medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Inferring the multiplicity of founder variants initiating HIV-1 infection: a systematic review and individual patient data meta-analysis

James Baxter, Sarah Langhorne, Ting Shi, Damien C. Tully, Ch. Julián Villabona-Arenas, Stéphane Hué, Jan Albert, Andrew Leigh Brown, Katherine E. Atkins

Usher Institute, The University of Edinburgh, Edinburgh, United Kingdom (J Baxter, T Shi PhD, K E Atkins PhD); Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom (S Langhorne MSc, Ch. J Villabona-Arenas ScD, D C Tully PhD, S Hué PhD, K E Atkins PhD); Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom (Ch. J Villabona-Arenas ScD, D C Tully PhD, S Hué PhD, K E Atkins PhD); Karolinska Institute, Stockholm, Sweden (Prof J Albert MD); Institute of Evolutionary Biology, The University of Edinburgh, Edinburgh, United Kingdom (Prof A Leigh Brown PhD)

Correspondence to:
James Baxter, Usher Institute, The University of Edinburgh, Edinburgh, EH9 3FL, United Kingdom,
James.Baxter@ed.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

1. Summary

Background

HIV-1 infections initiated by multiple founder variants are characterised by a higher viral load and a worse clinical prognosis, yet little is known about the routes of exposure through which multiple variant transmission is most likely, and whether methods of quantifying the number of founder variants differ in their accuracy.

Methods

We conducted a systematic review of studies that estimated founder variant multiplicity in HIV-1 infection, searching MEDLINE, EMBASE and Global Health databases for papers published between 1st January 1990 and 14th September 2020 (PROSPERO study CRD42020202672). Leveraging individual patient estimates from these studies, we performed a logistic meta-regression to estimate the probability that an HIV infection is initiated by multiple founder variants. We calculated a pooled estimate using a random effects model, subsequently stratifying this estimate across nine transmission routes in a univariable analysis. We then extended our model to adjust for different study methods in a multivariable analysis, recalculating estimates across the nine transmission routes.

Findings

We included 71 publications in our analysis, comprising 1664 individual patients. Our pooled estimate of the probability that an infection is initiated by multiple founder variants was 0·25 (95% CI: 0·21-0·30), with moderate heterogeneity ($Q = 137{\cdot}1,\ p < {\cdot}001, I^2 = 65{\cdot}3\%$). Our multivariable analysis uncovered differences in the probability of multiple variant infection by transmission route. Relative to a baseline of male-to-female transmission, the probability for female-to-male multiple variant transmission was significantly lower at 0·10 (95% CI: 0·05-0·21), while the probability for people-who-inject-drugs (PWID) transmission was significantly higher at 0·29 (0·13-0·52). There was no significant difference in the probability of multiple variant transmission between male-to-female transmission (0·16 (0·08-0·29)), post-partum mother-to-child (0·12 (0·02-0·51)), pre-partum mother-to-child (0·13 (0·05-0·32)), intrapartum mother-to-child (0.21 (0·08-0·44)) and men-who-have-sex-with-men (MSM) transmission (0·23 (0·03-0·7)).

Interpretation

We identified PWID transmissions are significantly more likely to result in an infection initiated by multiple founder variants, whilst female-to-male infections are significantly less likely. Quantifying how the routes of HIV infection impact the transmission of multiple variants allows us to better understand how the evolution and epidemiology of HIV-1 determine the clinical picture.

Funding

This study was supported by the MRC Precision Medicine Doctoral Training Programme (ref: 2259239) and a ERC Starting Grant awarded to KEA (award number 757688).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

2. Panel: Research in context

Evidence before this study

The majority of HIV-1 infections are initiated by a single, genetically homogeneous founder variant. Infections initiated by multiple founders, however, are associated with a significantly faster decline of CD4+ T Cells in untreated individuals, ultimately leading to an earlier onset of AIDS. Through our systematic search of MEDLINE, EMBASE and Global Health databases, we identified 82 studies that classify the founder variant multiplicity of acute HIV infections. As these studies vary in the methodology used to calculate the number of founder variants, it is difficult to evaluate the multiplicity of founder variants across routes of exposure.

Added value of this study

Using meta-regression, we estimated the probability of multiple founder infections across exposure routes by accounting for variability in methodology between studies. Our multivariable meta-regression adjusted for heterogeneity across study methodology and uncovered differences in the probability that an infection is initiated by multiple founder variants by transmission route, with the probability for female-to-male transmission significantly lower than for male-to-female transmission. By contrast, the probability for transmission among people-who-inject-drugs (PWID) was significantly higher. There was no difference in the probability of multiple founder variant transmission for mother-to-child transmission or men-who-have-sex-with-men (MSM) when compared with male-to-female.

Implications of all the available evidence

Because HIV-1 infections initiated by multiple founders are associated with a poorer prognosis, determining whether the route of infection affects the probability of transmission of multiple variants will facilitate an improved understanding of how the evolution and epidemiology of HIV-1 determine clinical progression. Our results identify that PWID transmissions are significantly more likely to result in an infection initiated by multiple founder variants compared to male-to-female. This reiterates the need for focussed public health programmes that reduce the burden of HIV-1 in this vulnerable risk group.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

3. Introduction

Transmission of HIV-1 results in a dramatic reduction in genetic diversity, with a large proportion of infections initiated by a single founder variant.[1,2] An appreciable minority of infections, however, appear to be the result of multiple founder variants simultaneously transmitted in a single exposure.[3] Importantly, these multiple founder infections are associated with both significant increases in set point viral load and the rate of CD4+ T lymphocyte decline.[4–7] HIV-1 infections initiated via different routes of exposure are subject to different virological, cellular and physiological environments, which likely influence the probability of acquiring infection.[8–10] For example, the probability of transmission upon exposure increases six-fold between heterosexual transmission and transmission between people who inject drugs (PWID), and up to eighteen-fold for men who have sex with men (MSM).[11]

Despite these differences in the probability of HIV-1 acquisition by route of exposure, there is currently no consensus about the effect of route of exposure on the transmission of multiple founder variants. Differences in selection pressure during transmission have been observed between sexual exposure routes, with reduced selection occurring during transmission from males to females than vice-versa, and less selection occurring between men who have sex with men (MSM) relative to those heterosexual exposure overall.[12,13] However, studies quantifying the number of founder variants are inconsistent with these findings, which may be due to differences in methodology and study population.[3,12,14,15] In sexual transmission, the probability of both transmission and founder variant multiplicity may also be influenced by inflammation, genital ulcerative disease and hormonal contraception, perhaps suggesting that the integrity of mucosal barrier underpins this process.[14,16] But, a significantly higher proportion of multiple founder infections in PWID transmissions, which bypass mucosal barriers altogether, has also not been consistently observed and so the role of exposure on the risk of acquiring a multiple founder infection remains unclear.[17,18] To estimate the role of exposure route on the acquisition of multiple HIV-1 founder variants, we conducted a meta-regression leveraging all available individual patient data, accounting for heterogeneity across methodology and study population.

4. Methods

4.1. Search Strategy and Eligibility Criteria

We searched MEDLINE, EMBASE and Global Health databases for papers published between 1 January 1990 to 14 September 2020 (S2: Supplementary Methods). To be included, studies must have reported original estimates of founder variant multiplicity in people acutely infected with HIV-1, be written in English and document ethical approval. Studies were excluded if they did not distinguish between single and multiple founder variants, if they did not detail the methods used, or if the study was conditional on having identified multiple founder variants. Additionally, studies were excluded if they solely reported data concerning people living with HIV-1 who had known or suspected superinfection, who were documented as having received pre-exposure prophylaxis, or if the transmitting partner was receiving antiretroviral treatment. No restrictions were placed on study design, geographic location, or age of participants. Publications were screened independently by SL and JB. Reviewers were blinded to the publication authors during the title and abstract screens and full text reviews were conducted independently, before a consensus was reached, with consultation with other co-authors when necessary.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

4.2. Data Extraction

Individual patient data (IPD) were collated from all studies, with authors contacted if these data were not available. Studies were excluded from further analysis if no IPD were obtained. Only individuals for whom a route of exposure was known were included. Additionally, we removed any entries for individuals with known or suspected superinfection, who were receiving pre-exposure prophylaxis or for whom the transmitting partner was receiving antiretroviral therapy. For this final individual patient dataset for analysis, we recorded whether an infection was initiated by one or multiple variants and nine predetermined covariates:

i. *Route of exposure*. Female-to-male (HSX-FTM), male-to-female (HSX-MTF), men-who-have-sex-with-men (MSM), pre-partum, intrapartum and post-partum mother to child (MTC), or people who inject drugs (PWID)).

ii. *Method of quantification.* Methodological groupings were defined by the properties of each approach, resulting in six levels: phylogenetic, haplotype, distance, model, or molecular (Table 1). Molecular methods interpret the formation of heteroduplexes during gel electrophoresis of viral RNA; haplotype methods identify linkage patterns of individual polymorphisms; distance and model-based methods assume a threshold or distribution of diversity that is reasonably expected to occur under a hypothesis of neutral exponential growth from a single founder and determine whether the observed diversity is consistent with the modelled values; and phylogenetic methods either use recipient sequences only, in which case a star-like topology is expected to be observed for single founder infections, or use source and recipient sequences from known transmission pairs, such that the number of distinct clades of recipient sequences nested within the source sequences corresponds to the number of founder variants.

iii. *HIV subtype*. Canonical geographically delimited subtypes (A-D, F-H, J and K) and circulating recombinant forms (e.g. CRF01_AE).[19,20] IPD where subtyping was unclear or not conducted were assigned 'unknown,' while putative recombinants not recognised as circulating recombinant forms were assigned 'recombinant.'

iv. *Delay between infection and sampling*. For sexual or injection drug use exposure, the delay was classified as either less than or equal to 21 days if the patient was seronegative at time of sampling (Feibig stages I-II) or more than 21 days if the patient was seropositive (Fiebig stages III-VI). For mother-to-child infections, if infection was confirmed at birth, or within 21 days of birth, the delay was classified as either less than or equal to 21 days. A positive mRNA or antibody test definitively reported after this period was classified as a delay of greater than 21 days.

v. *Number of genomes analysed per participant*.

vi. *Genomic region analysed*. Classified as envelope (Env), pol, gag or near full length genome (NFLG).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

vii. *Alignment length analysed*. Measured in base pairs, discretised at 250, 500, 1000, 2000, 4000, 8000, near full length genome (NFLG) intervals.

viii. *Use of single genome amplification (SGA) to generate viral sequences.* A binary classification (yes or no) as to whether the viral genomic data were generated using SGA. Regular bulk or near endpoint polymerase chain reaction (PCR) amplification can generate significant errors such as Taq-polymerase mediated template switching, nucleotide misincorporation or unequal amplicons resampling.[21,22] In SGA, serial dilutions of viral nucleic acids are made, which, assuming the proportion of positive PCR reaction at each dilution follows a null Poisson distribution, reduces the final reactions to contain a single variant that can be cloned, sequenced and then analysed.[22,23]

ix. *Study cohort*. The epidemiological cohort from which the patient was sampled.

If information from any of these nine covariates was missing or could not be inferred from the study, we classified its value as unknown. We excluded covariate levels for which there were fewer than 6 data points. For our base case analysis, we removed repeat measurements for the same individual, and used only those from the earliest study or, where the results of different methods were reported by the same study, the conclusive method used for each individual.

Molecular	**Haplotype**	**Distance**	**Model**	**Phylogenetic**	
				Recipient Only	*Source & Recipient*
Heteroduplex Assay	Highlighter plot Haplotype Frequency	Pairwise distance Diversity	Poissonfitter[24] • Goodness of fit • Starlike topology • tMRCA Other statistical or mathematical model	Starlike topology tMRCA (genealogy) Diversification	Paired topologies tMRCA (genealogy)

Table 1: Methods of quantification. Groupings of methods used to infer the founder variant multiplicity of HIV-1 infections. Model and phylogenetic methods may present as similar metrics such as the most recent common ancestor (tMRCA) and topology, but model-based approaches, unlike phylogenetic methods, do not use genealogical information in their calculation and instead are statistical models applied directly to the genomic data.

4.3. Pooled Meta-Analysis

We calculated a pooled estimates of the probability of multiple founder variant infection from our base case model: a 'one-step' generalised linear mixed model (GLMM) assuming an exact binomial distribution, with a normally distributed random effect on the intercept for within-study clustering and fitted by approximate maximum likelihood.[25] Heterogeneity was measured in terms of τ^2, the between-study variance; I^2, the percentage of variance

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

attributable to study heterogeneity; and Cochran's Q, an indicator of larger variation between studies than of subjects within studies.[26] Publication bias was assessed using funnel plots and Egger's regression test.[27]

Whilst pooled estimates obtained through a 'one-step' approach are usually congruent with the canonical 'two-step' meta-analysis model, discrepancies may arise due to differences in likelihood specification, weighting schemes, and specification of the intercept or estimation of residual variances.[28] We compared the results from our base case model with a two-step binomial-normal model to confirm our estimates were consistent. We performed additional sensitivity analyses to test the robustness of our pooled estimate to our exclusion criteria: iteratively excluding single studies, excluding studies that contained fewer than 10 participants, excluding studies that consisted solely of single founder infections, excluding IPD that did not use single genome amplification, and including only those data that matched our reference methodology of haplotype-based methods and whole genome analysis. In each of these sensitivity analyses, the base case model was refitted as previously described. To investigate the impact of our treatment of repeated measurements, we created 1000 datasets in which the included datapoint for each individual was sampled at random from a pool of their possible measurements. Each of these 1000 datasets thus contained a single datapoint per individual and we refitted the base case model to calculate a distribution of pooled estimates.

4.4. Meta-regression

We extended our base case model by conducting a univariable meta-regression with each covariate contributing a fixed effect and, assuming normally distributed random effects of publication. Pooled heterogeneity measures were calculated for each covariate level. We extended the base case model in a multivariable analysis, where we defined publication and cohort as crossed random effects before sequentially adding fixed effects covariates and evaluating interactions; assessing convergence, singularity and multicollinearity between fixed effects. The fixed effects were selected according to a 'keep it maximal' principle, in which covariates were only removed to facilitate a non-singular fit.[29] We defined our reference case as heterosexual male-to-female transmission, evaluated through haplotype-based methods, analysis of the whole genome sequences and a sampling delay of less than 21 days. Stratified predictions of the proportion of infections initiated by multiple founders and bootstrapped 95% confidence intervals, conditioned on the reference case, were calculated. We performed sensitivity analyses to test the robustness of the selected multivariable meta-regression model: iteratively excluding single studies, excluding studies that contained fewer than 10 participants, excluding studies that consisted solely of single founder infections and excluding IPD that did not use single genome amplification. The re-sampling sensitivity analysis was repeated on our selected multivariable model as described above.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

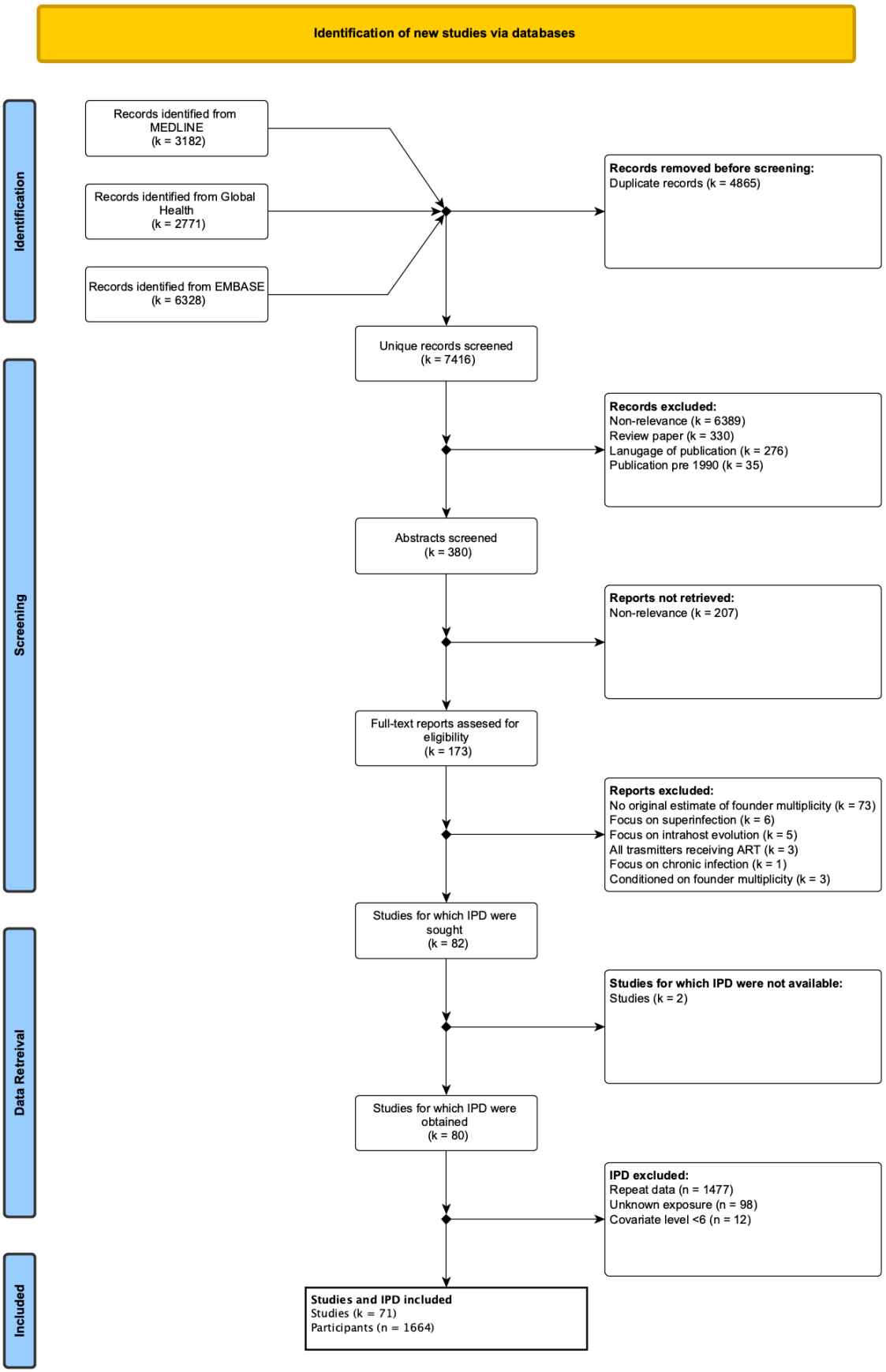

Figure 1: PRISMA flowchart outlining our systematic literature search and the application of exclusion criteria for the individual patient data meta-analysis.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

5. Results

5.1. Study and Patient Selection

Our search found 7416 unique papers, of which 7334 were excluded. Of the remaining 380 results, 207 were further excluded after abstract screening, leaving a total of 82 eligible studies for individual patient data (IPD) collation. We successfully extracted IPD from 80 of these studies, comprising 3251 data points. The 80 selected studies from which IPD were collated, were published between 1992 and 2020. Of the 3251 data points extracted, 1477 were excluded from our base case analysis to avoid repeated measurements; arising either between different studies that analysed the same individuals (resulting in the exclusion of five studies), or from repeat analysis of individuals within the same study. After excluding participants for whom the route of exposure was unknown or for whom one or more of their covariate values did not meet the minimum number of observations across the whole participants range of values, our final dataset for our base case analysis comprises estimates from 1664 unique patients across 71 studies.

5.2. Study and Patient Characteristics

Our base case dataset includes a median of 13 participants per study (range 2-124) and represents infections associated with heterosexual transmission (42·2%, (n = 703), MSM transmission (37·3%, n = 621), MTC mother-to-child transmission (14·1%, n = 234), and PWID transmission (6·4%, n = 106) (Fig. S3). Among heterosexual transmissions, 67·6% (n = 475) were male-to-female transmissions, 30% (n = 211) were female-to-male transmissions, with the remainder undisclosed (n = 17). Similarly, we subdivided MTC transmission according to the timing of infection with 44·4% (n = 104) pre-partum, 24·4% (n = 57) intrapartum, 4·7% (n = 11) post-partum, with the remainder undisclosed (n = 62). Our dataset spanned geographical regions and dominant subtypes, capturing the diversity of the HIV epidemic (Figs 2, S3). Across the base case dataset, phylogenetic methods constituted 37·1% (n = 618) of estimates, 26·7% (n = 445) were estimated using haplotype methods, 20·9% (n = 347) using molecular methods, and 12·9% (n = 215) and 2·34% (n = 39) of estimates were inferred using distance and model-based methods respectively (Table 2, Fig 2).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

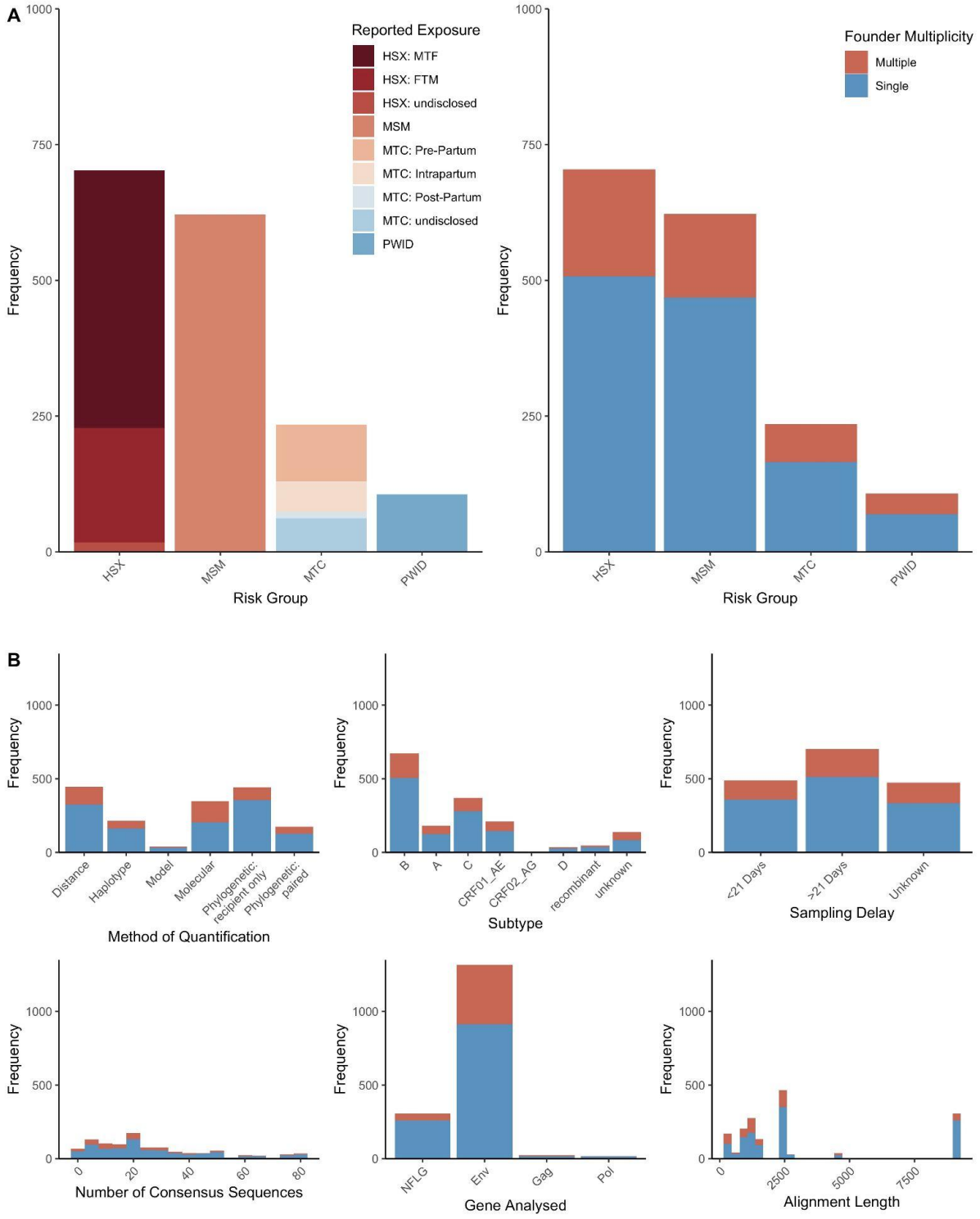

Figure 2: Individual patient data characteristics from the included studies that were tested for inclusion as fixed effects in the multivariable meta-regression model.

	Transmission Routes	Method	Genomic Region	Virus Subtype	Number of Participants	P(multiple founders)	Data Included	
							Participants	Multiple Founders
Wolinsky et al. (1992)[30]	MTC:undisclosed	Haplotype	Env; V3 & V4-V5	Unknown	3	0	3	0
Briant et al. (1995)[31]	MTC:undisclosed	Phylogenetic: source and recipient	Env; V3	B	4	0·75	4	3
Poss et al. (1995)[32]	HSX:MTF	Haplotype	Env; gp120	A, D	6	0·83	6	5
Wade et al. (1998)[33]	MTC:undisclosed MTC:PreP	Phylogenetic: source and recipient	Gag; p17	B	2	0·5	2	1
Long et al. (2000)[34]	HSX:MTF, HSX:FTM	Molecular	Env; gp120	A, D, C, Unknown	36	0·55	36	15
Dickover et al. (2001)[35]	MTC:IntraP MTC:PreP	Molecular	Env; gp120	B	23	0·26	23	6
Delwart et al. (2002)[36]	HSX:FTM HSX:MTF Unknown	Molecular	Env; V3	B	17	0·06	17	1
Learn et al. (2002)[37]	MSM	Molecular	Env; gp120	B	8	0·5	8	4
Long et al. (2002)[38]	HSX:MTF	Distance	Env; gp120	A, Unknown	5	0·5	2	0
Nowak et al. (2002)[39]	MTC:undisclosed	Phylogenetic: source and recipient	Env; V3	B	3	0·34	3	1
Renjifo et al. (2003)[40]	MTC:PreP	Molecular	Env; gp120	A, C, D	53	0·21	53	11
Sagar et al. (2003)[4]	HSX:MTF	Molecular	Env; gp120	Unknown	124	0·56	124	55
Verhofstede et al. (2003)[41]	MTC:IntraP MTC:PreP	Phylogenetic: source and recipient	Env; gp120	A	13	0·54	13	7

Derdeyn et al. (2004)[42]	HSX:MTF HSX:FTM	Phylogenetic: source and recipient	Env; gp120	C, G	7	0	7	0
Ritola et al. (2004)[43]	HSX:MTF HSX:FTM MSM	Molecular	Env; V1-V3	B	26	0·52	25	7
Sagar et al. (2004)[16]	HSX:MTF PWID MSM HSX:FTM	Molecular	Env; V1-V5	A, B, Unknown	17	0·24	17	4
Sagar et al. (2006)[44]	HSX:MTF	Distance	Env; V1-V3	A, D, Unknown, Recombinants	12	0·5	··	··
Gottlieb et al. (2008)[45]	MSM	Haplotype	Env; V1-V5	B	38	0·39	37	14
Keele et al. (2008)[3]	PWID MSM Unknown HSX:FTM HSX:MTF	Distance Haplotype Model Phylogenetic: recipient only	Env; gp160	B	102	0·24	44	15
Kwiek et al. (2008)[46]	MTC:IntraP MTC:PreP	Molecular	Env; V1-V2	C	48	0·42	48	28
Salazar-Gonzalez et al. (2008)[23]	HSX:MTF HSX:FTM	Distance Haplotype Phylogenetic: recipient only	Env; gp160	C, Unknown	12	0·34	12	4
Abrahams et al. (2009)[47]	HSX:FTM HSX:MTF	Distance Model Haplotype Phylogenetic: recipient only	Env; gp160	C, G	69	0·22	69	15
Haaland et al. (2009)[14]	HSX:MTF HSX:FTM	Haplotype Phylogenetic: source and	Env; gp160	A, C, Unknown	27	0·23	22	3

		recipient						
Kearney et al. (2009)[48]	MSM HSX:FTM HSX:MTF PWID	Phylogenetic: recipient only	pol	B	14	0·14	11	0
Novitsky et al. (2009)[49]	HSX:MTF HSX:FTM	Phylogenetic: recipient only	Env; gp120	C	8	0·25	8	2
Salazar-Gonzalez et al. (2009)[50]	MSM HSX:FTM	Distance Haplotype Model	NFLG	B, C	12	0·083	2	0
Bar et al. (2010)[17]	PWID	Phylogenetic: recipient only	Env; gp160	B	10	0·6	10	6
Fischer et al. (2010)[51]	MSM	Model	Env; gp120	B	3	0	··	··
Li et al. (2010)[15]	MSM	Distance Haplotype	Env; gp160	B	28	0·36	28	10
Masharsky et al. (2010)[18]	PWID	Haplotype	env	A, Recombinants	13	0·31	13	4
Zhang et al. (2010)[52]	MTC:IntraP	Phylogenetic: source and recipient	Env; V1-V5	C, Recombinants	6	0	6	0
Boeras et al. (2011)[53]	HSX:FTM HSX:MTF	Phylogenetic: source and recipient	Env; V1-V4	A, C	8	0	··	··
Collins-Fairclough et al. (2011)[54]	MSM HSX:FTM HSX:MTF HSX:undisclosed	Haplotype	Env; V1-C4	B	27	0·23	14	2
Herbeck et al. (2011)[55]	MSM	Distance	NFLG	B	9	0·11	9	1
Kishko et al. (2011)[56]	MTC:IntraP	Phylogenetic: source and recipient	Env; gp160	B	5	0·4	5	2

Nofemela et al. (2011)[57]	HSX:MTF	Haplotype	env	A, B, C, D, Recombinants	22	0·27	22	6
Novitsky et al. (2011)[58]	HSX:MTF HSX:FTM	Distance Haplotype Model Phylogenetic: recipient only	gag & Env; gp120	C	25	0·32	16	6
Rachinger et al. (2011)[59]	MSM	Phylogenetic: source and recipient	NFLG	B	1	0	··	··
Rieder et al. (2011)[60]	Unknown MSM HSX:MTF	Distance	Env; C2-V3-C3	A, B, C, G, CRF01AE, CRF02AG, CRF12BF, CRF14BG	143	0·11	··	··
Rolland et al. (2011)[61]	MSM HSX:MTF	Phylogenetic: recipient only	NFLG	B, CRF02AG	68	0·25	68	16
Cornelissen et al. (2012)[5]	MSM	Phylogenetic: recipient only	Env; V3-V4	B	31	0·13	31	4
Henn et al. (2012)[62]	unknown	Distance	NFLG	B	1	0	··	··
Kiwelu et al. (2012)[63]	HSX:MTF	Phylogenetic: recipient only	Env; gp120	A, C, D	50	0·27	43	10
Rossenkhan et al. (2012)[64]	HSX:MTF HSX:FTM	Phylogenetic: recipient only	gag & Env; gp120	C	20	0·15	5	0
Sturdevant et al. (2012)[65]	MTC:undisclosed	Haplotype Phylogenetic: recipient only	Env; gp160	C	43	0·12	43	5
Baalwa et al. (2013)[66]	HSX:MTF HSX:FTM	Haplotype	NFLG	A, D, Recombinants	12	0·17	12	2
Frange et al. (2013)[67]	MSM HSX:MTF	Phylogenetic: source and recipient	Env; C2-V5	B	8	0	8	0

	HSX:FTM							
Chaillon et al. (2014)[68]	MTC:PreP MTC:IntraP	Phylogenetic: source and recipient	Env; V1-V5	CRF01_AE	9	0·12	8	1
Sterrett et al. (2014)[69]	PWID	Distance Haplotype Model Phylogenetic: recipient only	Env; gp160	B, CRF01AE, CRF1501B, Recombinants	50	0·42	49	14
Wagner et al. (2014)[70]	MSM PWID	Phylogenetic: recipient only	NFLG	B	108	0·06	108	7
Chen et al. (2015)[71]	MSM	Haplotype	Env; gp160	B, CRF01AE, CRF07BC	30	0·2	18	3
Danaviah et al.(2015)[72]	MTC:PostP	Phylogenetic: source and recipient	Env; C2-V5	C	11	0·18	11	2
Deymier et al. (2015)[73]	HSX:FTM	Phylogenetic: recipient only	NFLG	C	6	0	5	0
Gounder et al. (2015)[74]	HSX:FTM HSX:MTF	Phylogenetic: recipient only	gag	C	22	0·27	22	6
Janes et al. (2015)[6]	MSM HSX:FTM HSX:MTF	Distance	Env; gp120	B, CRF01AE	163	0·29	100	32
Le et al. (2015)[75]	PWID	Phylogenetic: source and recipient	Env; gp120	B	2	0	2	0
Zanini et al. (2015)[76]	HSX:MTF MSM HSX:FTM	Distance	NFLG	B, C, CRF01AE	9	0·22	9	2
Chaillon et al. (2016)[77]	MSM PWID	Distance Phylogenetic: source and recipient	Env; C2-V3	B	30	53·3	30	16

Love et al. (2016)[78]	PWID MSM Unknown HSX:FTM HSX:MTF HSX:undisclosed	Model	Env; gp160	B, C	182	0·23	··	··
Novitsky et al. (2016)[79]	HSX:MTF HSX:FTM	Distance	Env; V1-C5	C	42	0·21	15	3
Oberle et al. (2016)[80]	MSM HSX:MTF	Phylogenetic: source and recipient	Env; gp160	B	9	0	2	0
Park et al. (2016)[81]	MSM	Model	Env; gp160	B, CRF02AG	59	0·17	··	··
Salazar-Gonzalez et al. (2016)[82]	unknown	Haplotype	Env; gp160	B	2	0	··	··
Smith et al. (2016)[83]	HSX:FTM HSX:MTF	Haplotype	Env; gp120	A, C, Recombinants	21	0	19	0
Tully et al. (2016)[12]	Unknown MSM PWID HSX:undisclosed NOSO	Distance Haplotype Model Phylogenetic: recipient only	Env; gp160, NFLG	B, C, CRF02AG	74	0·17	67	11
deCamp et al. (2017)[84]	MSM	Phylogenetic: recipient only	Env; gp120	B	46	0·28	43	12
Iyer et al. (2017)[85]	MSM HSX:FTM HSX:MTF	Haplotype	NFLG	B, C	8	0·13	7	1
Kijak et al. (2017)[86]	HSX:MTF HSX:FTM	Haplotype	NFLG	CRF01_AE, Recombinants	6	0·83	6	5
Ashokkumar et al. (2018)[87]	MTC:undisclosed	Haplotype	Env; gp120	C	8	0·25	8	2
Dukhovlinova et al. (2018)[88]	PWID	Model	Env; gp160	A	7	0	7	0
Leitner & Romero-Severson	MSM	Phylogenetic: source and	Various	A, B, C, D,	508	0.52	··	··

(2018)[89]	HSX:MTF HSX:FTM PWID HSX:undisclosed MTC:undisclosed Unknown NOSO	recipient		CRF01_AE, CRF14_BG				
Lewitus & Rolland (2019)[90]	Unknown MSM HSX:FTM HSX:MTF	Phylogenetic: recipient only	Env; gp160	B	72	0·29	··	··
Sivay et al. (2019)[91]	PWID	Model	Env; gp41	A, CRF01AE	7	0·43	7	3
Todesco et al. (2019)[92]	MSM	Phylogenetic: source and recipient	pol	B, CRF02AG, CRF07BC	8	0·25	7	2
Tovanabutra et al. (2019)[93]	MSM HSX:MTF	Haplotype	Env; gp160	CRF01_AE, recombinant	18	0·44	18	7
Brooks et al. (2020)[94]	HSX:FTM HSX:MTF	Phylogenetic: recipient only	NFLG	C	13	0·08	12	1
Leda et al. (2020)[95]	HSX:MTF MSM HSX:FTM	Model	Env; gp160	B, F, Recombinant	25	0·08	21	2
Liu et al. (2020)[96]	MSM	Haplotype	Env; gp120	B, CRF01_AE	8	0·25	8	2
Macharia et al. (2020)[7]	MSM	Phylogenetic: recipient only	NFLG	A	38	0·39	38	15
Martinez et al. (2020)[97]	MTC:IntraP MTC:PreP	Model	Env; gp160	B, C	4	0·25	4	1
Rolland et al. (2020)[98]	HSX:MTF MSM	Phylogenetic: recipient only	Env; gp160	A, B, C, CRF01AE	39	0·28	39	10
Villabona-Arenas et al. (2020)[99]	MSM HSX:undisclosed	Phylogenetic: source and recipient	Env; gp41, gp160,	A, B, C, D, G, Recombinants	112	0·23	49	12

	HSX:MTF HSX:FTM		gp120 & NFLG					

Table 2: Included studies selected for inclusion from our systematic literature search. We record the route of transmission: female-to-male (HSX:FTM), male-to-female (HSX:MTF), men-who-have-sex-with-men (MSM), mother-to-child pre-partum (MTC:PreP), intrapartum (MTC:IntP) and post-partum (MTC:PostP); people who inject drugs (PWID), or nosocomial (NOSO). Additionally, we tabulate the method grouping used to infer founder multiplicity, the genomic region analysed, the number of participants analysed and the proportion of infections initiated by multiple founders reported by each study. We note the number of single and multiple founder infections included within our base case dataset.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

5.3. Meta-analyses

5.3.1. Pooled Estimate

Our base case analysis using a GLMM estimated the probability that an infection is initiated by multiple founder variants to be 0.25 (95% CI: 0.21-0.29), identifying significant heterogeneity ($Q = 137.1$, $p < .001$, $I^2 = 65{\cdot}3\%$). Our sensitivity analyses revealed the pooled estimate is robust to the choice of model, the inclusion of estimates from repeat participants, and to the exclusion of studies that contained fewer than 10 participants (Fig. S4, S5). While analysing only data that matched our reference case study methodology did not change our estimate, it widened the confidence intervals of our estimate (0.25 (95% CI: 0.05-0.67)). We did not identify any studies that significantly influenced the pooled estimate (Fig. S6). Visual inspection of a funnel plot and a non-significant Egger's Test (t = -0·2663, df = 56, p = 0·7910), were consistent with an absence of publication bias in our dataset (Fig. S7).

5.3.2. Meta-Regression

We extended our base case binomial GLMM using uni- and multivariable fixed effects. Relative to a reference exposure route of male-to-female transmission, our univariable analysis found significantly lower odds of female-to-male transmission being initiated by multiple founder variants (Odds Ratio (OR): 0·56 (95% CI 0·33-0·87)), while other exposure routes were not significantly different. The univariable analyses also indicated significantly greater odds of multiple founder variants if the envelope genomic region was analysed (OR: 2·06 (95% CI:1·16-3·98)), relative to the whole genome. Other methodological covariates, however, such as method of quantification and sampling delay were not significantly associated with the odds that HIV-1 infection is initiated by multiple founder variants.

Our base case multivariable model calculated the probability of multiple founder variants across the seven routes of transmission controlling for method, genomic region and sampling delay (Fig. 3). Compared to a male-to-female transmission probability of 0·16 (95% CI: 0·08-0·29), there was no evidence that the probability of multiple founder variants differed across MSM (0·23 (0·03-0·7)) or MTC transmission. Stratifying MTC transmissions by the putative timing of infection, we calculated pre-partum were initiated by multiple founders with probability 0·13 (0·05-0·32), post-partum with probability 0·12 (0·02-0·51), and intrapartum transmissions with probability 0·21 (0·08-0·44).

By contrast, we found that female-to-male transmissions were less likely to be initiated by multiple founders than male-to-female transmissions, with probability 0·10 (95% CI: 0·05-0·21) (OR: 0·61 (95% CI 0·36-0·94)). Conversely, PWID transmission was more likely to be initiated by multiple founders (0·29 (0·13-0·52)), compared to male-to-female (OR: 2.19 (1·10-4·42)).

We calculated the accuracy of estimating the probability of multiple founder variants compared to a gold-standard methodological reference scenario of using haplotype-based methods on whole genome sequences with individuals with less than 21 delays between infection and sampling. Our base case analysis indicates using model-based methods underestimates the chance of multiple founder variants (OR: 0·32 (95% CI: 0·05-0·82)), while using the gag or envelope genomic regions overestimates the chance of detecting multiple founder variants by (OR of 4·32; 95% CI:1·03-20·47 and 1·78 (0·99-3·86) respectively). Our sensitivity analyses revealed the odds ratios calculated using

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

the uni- and multivariable models are robust to inclusion of data from repeated participants, and to the exclusion of studies that contained fewer than 10 participants, of studies that consisted solely of single founder infections, and of individual data that did not use single genome amplification (Fig. S7).

		Univariable		Multivariable	
		Odds Ratio [95% CI]	**p-value**	**Odds Ratio [95% CI]**	**p-value**
Reported Exposure					
	Heterosexual: male-to-female	1 (reference)	-	1 (reference)	-
	Heterosexual: female-to-male	**0·56 [0·33-0·87]**	**0·011**	**0·61 [0·36-0·94]**	**0·026**
	Heterosexual: undisclosed	1·77 [0·33-4·82]	0·359	1·61 [0·33-5·66]	0·455
	MSM	1·28 [0·83-2·06]	0·.299	1·46 [0·92-2·19]	0·086
	Mother-to-child: pre-partum	1·18 [0·48-2·45]	0·699	0·82 [0·33-1·97]	0·653
	Mother-to-child: intrapartum	1·76 [0·75-3·78]	0·199	1·43 [0·60-3·39]	0·397
	Mother-to-child: post-partum	0·70 [0·01-2·98]	0·740	0·76 [0·00-3·62]	0·783
	Mother-to-child: undisclosed	1·17 [0·40-4·22]	0·758	0·85 [0·28-2·32]	0·755
	PWID	2·01 [0·87-4·41]	0·062	**2·19 [1·10-4·42]**	**0·025**
Grouped Method					
	Haplotype	1 (reference)	-	1 (reference)	
	Distance	0·69 [0·32-1·42]	0·309	0·93 [0·45-2·14]	0·843
	Model	0·48 [0·11-1·41]	0·210	**0·32 [0·05-0·82]**	**0·047**
	Molecular	1·76 [0·95-3·03]	0·068	1·62 [0·76-3·09]	0·149
	Phylogenetic: recipient only	0·66 [0·39-1·18]	0·133	0·59 [0·37-1·02]	0·074
	Phylogenetic: source & recipient	0·81 [0·48-1·49]	0·507	0·81 [0·45-1·36]	0·487
Genomic Region					
	NFLG	1 (reference)	-	1 (reference)	-
	Envelope	**2·06 [1·16-3·98]**	**0·019**	1·78 [0·99-3·86]	0·085
	Gag	2·37 [0·32-8·91]	0·255	**4·32 [1·03-20·47]**	**0·043**
	Pol	0·63 [0·00-3·.06]	0·610	0·61 [0·00-2·45]	0·576
Sampling Delay					
	<21 Days	1 (reference)	-	1 (reference)	-
	>21 Days	1·05 [0·69-1·46]	0·809	1·09 [0·71-1·52]	0·642
	Unknown	1·34 [0·73-2·29]	0·291	1·17 [0·66-2·03]	0·552

Table 2: Odds ratios that an HIV-1 infection is initiated by multiple founder variants, inferred from fixed effects coefficients from the univariable and selected multivariable meta-regression models. Significant effects in bold. MSM - men who have sex with men; PWID - people who inject drugs; NFLG - near full length genome.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

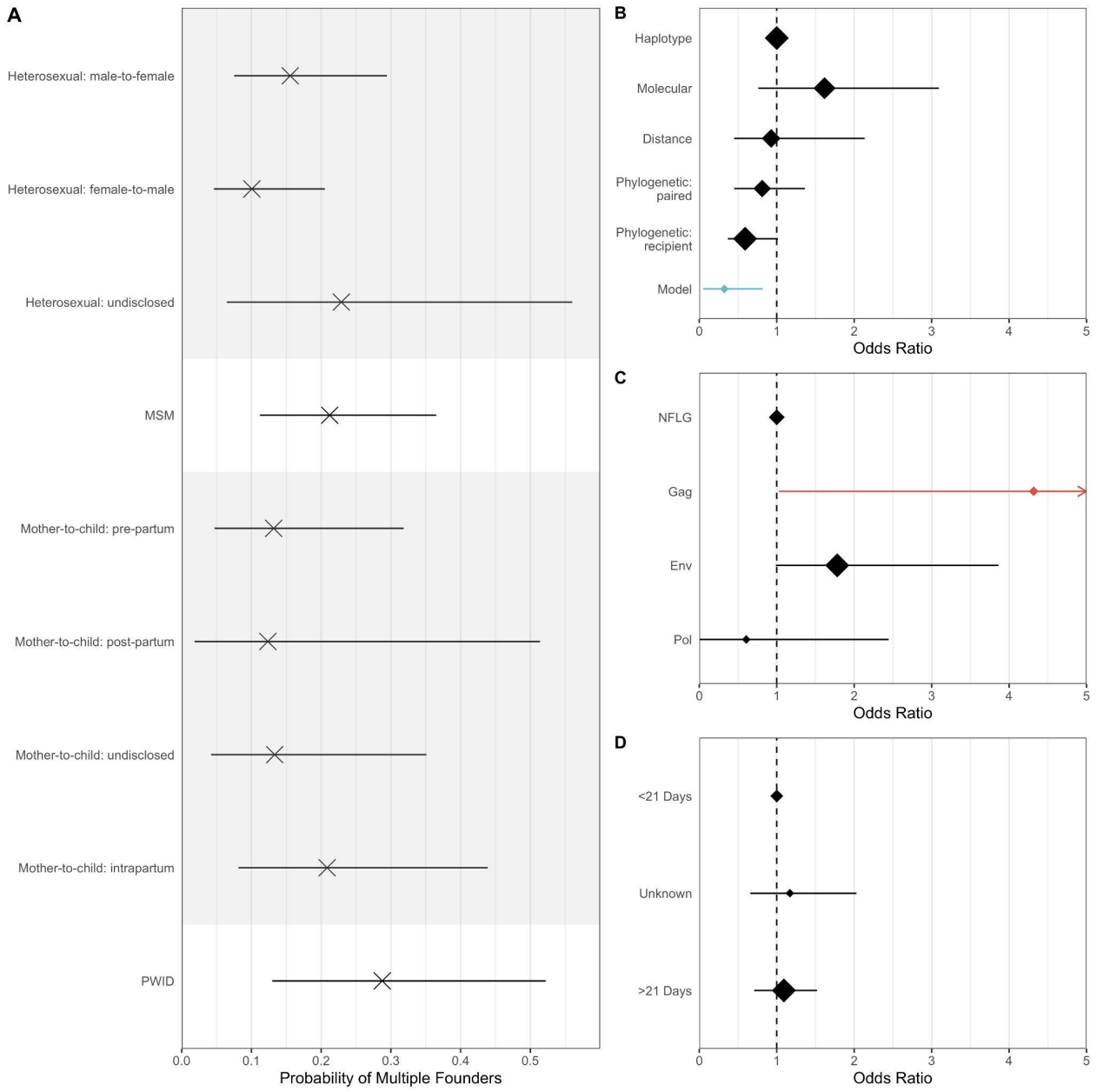

Figure 3: Predictions and coefficients obtained from the multivariable model. A) predicted probabilities of an infection being initiated by multiple founder variants, stratified by the route of exposure. B-D) Inferred odds ratios of fixed effects variables. Blue denotes that a covariate level significantly decreases the odds of an infection being initiated by multiple founders, whilst red indicates covariate levels for which the odds are significantly greater. For each plot, the reference case is marked at the top of the y axis, with the dotted line at x=1 demarcating the reference plane.

6. Discussion

Using data from 71 previous studies, we estimated that a quarter of HIV-1 infections are initiated by multiple founder variants. When controlling for different methodologies across studies, the probability that an infection is initiated by multiple founders decreased relatively by 37.5% for female-to-male infections with respect to a baseline of

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

male-to-female infections, but increased by 81.25% for infections transmitted between people who inject drugs. Further, we found that model-based methods, representing a group of approaches that determine founder multiplicity by comparing the observed distribution of diversity with that expected under neutral exponential outgrowth from single variant transmission, were less likely to identify multiple founder infections. Together these results suggest that while the exposure route probably influences the number of founder variants, previous comparison has been difficult due to different study methodologies.

Our pooled estimate is consistent with the seminal study of Keele et al., who found 23·5% (24/102) of their participants had infections initiated by multiple founders.[3] Our stratified predicted probabilities are also in line with those of previous smaller studies. A nine-study meta-analysis of 354 subjects found 0·34 of PWID infections were initiated by multiple founders compared with 0·29 (95% CI: 0·13-0·52) in our study; 0·2 for heterosexual infections compared to 0·23 (0·06-0·56) and 0·25 for MSM infections for which we calculated 0·23 (0·03-0·7).[12] Likewise, an earlier meta-analysis of five studies and 235 subjects found PWID infections were at significantly greater odds than heterosexual infections of being initiated by a single founder, with the frequency of founder variant multiplicity increasing 3-fold, while a smaller, non-signficant 1·5-fold increase was observed with respect to MSM transmissions.[17] In both instances, these studies restricted the number of participants so that the methodology in estimating founder variant multiplicity was consistent across all subjects. In this study, in contrast, we were able to extend our meta-analysis by leveraging individual level data to control for methodological sources of heterogeneity.

We did not identify any significant effect of sampling delay on the probability that an infection is identified to be initiated by multiple founders. While previous work has shown a negative association between detection of multiple founder variants and the delay from infection to sampling, this discrepancy is likely due to the range of the delay analysed. Specifically, Leitner and Romero-Severson found a reduced chance of multiple founder variants over a period of 8 years, while our study analyses over a shorter time span of less or greater than a 3 week delay[89]

Certain routes of transmission that our analysis found to be associated with a higher or lower probability of multiple founder variants, have previously been identified as having higher or lower probabilities of transmission, respectively. For example, we estimated that female-to-male multiple variant transmission is 39% less likely than that of male-to-female, while the per exposure transmission probability has been estimated at half as likely.[11] Similarly, MSM infections are 46% more likely to be initiated by multiple founders, but here the probability of infection following a given exposure can be up to 33-fold greater than that in male-to-female infections. By contrast, although PWID infections were found to be the most likely to be initiated by multiple founders, PWID are less likely to be infected upon exposure than MSM, and 14-fold greater than male-to-female infections. Further, mother-to-child infections are not significantly more likely than MSM or heterosexual infections to be initiated by multiple founders, but the probability of infection for mother-to-child exposures is 16-times and 565-times greater, respectively. Our results suggest a complicated relationship between the probability of transmission and the probability of multiple founder transmission.

Our analysis has some limitations. First, our definition of single and multiple founder variants is determined by the individual studies, however questions remain concerning the definition of a founder variant. Recent studies have

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

suggested a continuum of genotypic diversity exists, rather than discrete variants that give rise to distinct phylogenetic diversification trajectories and may not be reflected by this binary classification.[90,98] Indeed, although a threshold is specified for distance-based methods, above which the observed diversity is defined to be to great to be explained by neutral exponential growth, this threshold often varies between publications.[100,101] For example, both Keele et al and Li et al analysed the diversity of the envelope protein, but whilst the former classifies populations with less than 0.47% diversity as homogenous, Li et al included samples up to 0.75%.[3,15] The distinction between single and multiple founder variants may further be blurred by non-coalescent sources of variation such as recombination and APOBEC mediated hypermutation, which would erroneously inflate diversity measures unless accounted for.[102,103] Ultimately, the classification of multiple/single founders is subjective and may also be informed by cognitive biases of the authors. This is pertinent to studies which recruit participants from specific, often marginalised risk groups (e.g. MSM, PWID), where authors may have been more likely to classify multiple founder infections based on their prior assumptions.Second, we acknowledge that under the hypothesis that the proportion of infections initiated by multiple founders varies by transmission route, our point estimate will be influenced by the relative proportion of transmission routes in our dataset. Globally, it is estimated that 70% of infections are transmitted heterosexually, compared to 42.2% in our dataset, which reflects the longstanding geographical bias of research towards patients in the global north.[104] Therefore, our point estimate should be considered a summary of the published data over the course of the HIV-1 epidemic, and not a global estimate at any fixed point in time. Third, we were unable to account for the stage of infection in the transmitter, despite recent findings that transmitters with acute infections are more likely to initiate multiple variant infections, because we had insufficient data regarding the transmitting partner within our dataset.[99] Finally, we acknowledge the bootstrapped confidence intervals are wide and may lead to uncertainty in our estimates. These arise as a product of small sample sizes for certain observations, and the crossed random effects of publication and cohort used in the meta-regression. In particular, our finding that infections analysed using gag are significantly more likely to be initiated by multiple founders demonstrates substantial uncertainty, and is arguably unlikely considering the mutation rate of envelope is significantly higher than gag during primary infection.[105] We note that in this case, the results of our univariable analysis of genomic region analysed are more consistent with our prior expectations.

This systematic review and meta analysis has demonstrated that infections initiated by multiple founders account for a quarter of HIV-1 infections across all known routes of transmission. We find that transmissions involving people who inject drugs are significantly more likely to be initiated by multiple founder variants, whilst female-to-male infections are significantly less likely, relative to male-to-female infections. Quantifying how the routes of HIV infection impact the transmission of multiple variants allows us to better understand the evolution, epidemiology and clinical picture of HIV transmission.

7. Contributors

KEA conceived the study. JB, SL, DT, KEA designed the study. JB and SL extracted the data. JB performed the experiments and analysed the data. All authors interpreted the data. JB and KEA drafted the manuscript, with critical revisions from all authors. All authors approved the final version of the manuscript

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

8. Declaration of Interests

The authors declare no competing interests

9. Data Sharing

Code use in this study is available at https://github.com/J-Baxter/foundervariantsHIV_sysreview

10. Acknowledgements

JB was supported by the MRC Precision Medicine Doctoral Training Programme (ref: 2259239); CJV-A and KEA were funded by an ERC Starting Grant (award number 757688) awarded to KEA. We are grateful to Kamini Gounder, Mary Kearney, Vladimir Novitsky, Morgane Rolland and Sodsai Tovanabutra for agreeing to share additional individual patient data with the authors in order to complete this study.

11. References

1 Zhu T, Mo H, Wang N, *et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993; **261**: 1179 LP – 1181.
2 Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, Simmonds P. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J Virol* 1993; **67**: 3345 LP – 3356.
3 Keele BF, Giorgi EE, Salazar-Gonzalez JF, *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci* 2008; **105**: 7552–7.
4 Sagar M, Lavreys L, Baeten JM, *et al.* Infection with multiple human immunodeficiency virus type 1 variants is associated with faster disease progression. *J Virol* 2003; **77**: 12921–6.
5 Cornelissen M, Pasternak AO, Grijsen ML, *et al.* HIV-1 Dual Infection Is Associated With Faster CD4+ T-Cell Decline in a Cohort of Men With Primary HIV Infection. *Clin Infect Dis* 2012; **54**: 539–47.
6 Janes H, Herbeck JT, Tovanabutra S, *et al.* HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. *Nat Med* 2015; **21**: 1139.
7 Macharia GN, Yue L, Staller E, *et al.* Infection with multiple HIV-1 founder variants is associated with lower viral replicative capacity, faster CD4+ T cell decline and increased immune activation during acute infection. *PLoS Pathog* 2020; **16**: e1008853–e1008853.
8 Kariuki SM, Selhorst P, Ariën KK, Dorfman JR. The HIV-1 transmission bottleneck. *Retrovirology* 2017; **14**: 22.
9 Joseph SB, Swanstrom R, Kashuba ADM, Cohen MS. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nat Rev Microbiol* 2015; **13**: 414–25.
10 Talbert-Slagle K, Atkins KE, Yan K-K, *et al.* Cellular Superspreaders: An Epidemiological Perspective on HIV Infection inside the Body. *PLOS Pathog* 2014; **10**: e1004092.
11 Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. *AIDS* 2014; **28**.
12 Tully DC, Ogilvie CB, Batorsky RE, *et al.* Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. *PLoS Pathog* 2016; **12**: e1005619–e1005619.
13 Carlson JM, Schaefer M, Monaco DC, *et al.* HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* 2014; **345**: 1254031–1254031.
14 Haaland RE, Hawkins PA, Salazar-Gonzalez J, *et al.* Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1. *PLOS Pathog* 2009; **5**: e1000274.
15 Li H, Bar KJ, Wang S, *et al.* High multiplicity infection by HIV-1 in men who have sex with men. *PLoS Pathog* 2010; **6**: e1000890.
16 Sagar M, Kirkegaard E, Long EM, *et al.* Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes. *J Virol* 2004; **78**: 7279–83.
17 Bar KJ, Li H, Chamberland A, *et al.* Wide variation in the multiplicity of HIV-1 infection among injection drug

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

users. *J Virol* 2010; **84**: 6241–7.
18 Masharsky AE, Dukhovlinova EN, Verevochkin SV, *et al.* A Substantial Transmission Bottleneck among Newly and Recently HIV-1-Infected Injection Drug Users in St Petersburg, Russia. *J Infect Dis* 2010; **201**: 1697–702.
19 Robertson DL, Anderson JP, Bradac JA, *et al.* HIV-1 nomenclature proposal. *Science* 2000; **288**: 55.
20 Archer J, Robertson DL. Understanding the diversification of HIV-1 groups M and O. *Aids* 2007; **21**: 1693–700.
21 Meyerhans A, Vartanian J-P, Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Res* 1990; **18**: 1687–91.
22 Simmonds P, Balfe P, Peutherer JF, Ludlam CA, Bishop JO, Brown AJ. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. *J Virol* 1990; **64**: 864–72.
23 Salazar-Gonzalez JF, Bailes E, Pham KT, *et al.* Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 2008; **82**: 3952–70.
24 Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, Bhattacharya T. Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. *BMC Bioinformatics* 2010; **11**: 532.
25 Riley RD, Legha A, Jackson D, *et al.* One‐stage individual participant data meta‐analysis models for continuous and binary outcomes: Comparison of treatment coding options and estimation methods. *Stat Med* 2020; **39**: 2536–55.
26 Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. John Wiley & Sons, 2011.
27 Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Bmj* 1997; **315**: 629–34.
28 Burke DL, Ensor J, Riley RD. Meta‐analysis using individual participant data: one‐stage and two‐stage approaches, and why they may differ. *Stat Med* 2017; **36**: 855–75.
29 Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* 2013; **68**: 10.1016/j.jml.2012.11.001.
30 Wolinsky SM, Wike CM, Korber BT, *et al.* Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science* 1992; **255**: 1134 LP – 1137.
31 Briant L, Wade CM, Puel J, Brown AJ, Guyader M. Analysis of envelope sequence variants suggests multiple mechanisms of mother-to-child transmission of human immunodeficiency virus type 1. *J Virol* 1995; **69**: 3778–88.
32 Poss M, Martin HL, Kreiss JK, *et al.* Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. *J Virol* 1995; **69**: 8118–22.
33 Wade CM, Lobidel D, Brown AJ. Analysis of human immunodeficiency virus type 1 env and gag sequence variants derived from a mother and two vertically infected children provides evidence for the transmission of multiple sequence variants. *J Gen Virol* 1998; **79**: 1055–68.
34 Long EM, Martin HL, Kreiss JK, *et al.* Gender differences in HIV-1 diversity at time of infection. *Nat Med* 2000; **6**: 71–5.
35 Dickover RE, Garratty EM, Plaeger S, Bryson YJ. Perinatal transmission of major, minor, and multiple maternal human immunodeficiency virus type 1 variants in utero and intrapartum. *J Virol* 2001; **75**: 2194–203.
36 Delwart E, Magierowska M, Royz M, *et al.* Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS* 2002; **16**. https://journals.lww.com/aidsonline/Fulltext/2002/01250/Homogeneous_quasispecies_in_16_out_of_17.7.aspx.
37 Learn GH, Muthui D, Brodie SJ, *et al.* Virus population homogenization following acute human immunodeficiency virus type 1 infection. *J Virol* 2002; **76**: 11953–9.
38 Long EM, Rainwater SMJ, Lavreys L, Mandaliya K, Overbaugh J. HIV type 1 variants transmitted to women in Kenya require the CCR5 coreceptor for entry, regardless of the genetic complexity of the infecting virus. *AIDS Res Hum Retroviruses* 2002; **18**: 567–76.
39 Nowak P, Karlsson AC, Naver L, Bohlin AB, Piasek A, Sönnerborg A. The selection and evolution of viral quasispecies in HIV-1 infected children. *HIV Med* 2002; **3**: 1–11.
40 Renjifo B, Chung M, Gilbert P, *et al.* In-utero transmission of quasispecies among human immunodeficiency virus type 1 genotypes. *Virology* 2003; **307**: 278–82.
41 Verhofstede C, Demecheleer E, De Cabooter N, *et al.* Diversity of the human immunodeficiency virus type 1 (HIV-1) env sequence after vertical transmission in mother-child pairs infected with HIV-1 subtype A. *J Virol* 2003; **77**: 3050–7.
42 Derdeyn CA, Decker JM, Bibollet-Ruche F, *et al.* Envelope-Constrained Neutralization-Sensitive HIV-1 After Heterosexual Transmission. *Science* 2004; **303**: 2019 LP – 2022.
43 Ritola K, Pilcher CD, Fiscus SA, *et al.* Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. *J Virol* 2004; **78**: 11208–18.
44 Sagar M, Wu X, Lee S, Overbaugh J. HIV-1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection and these modifications affect antibody neutralization sensitivity. *J Virol* 2006; **80**: 9586–98.
45 Gottlieb GS, Heath L, Nickle DC, *et al.* HIV-1 variation before seroconversion in men who have sex with men:

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

analysis of acute/early HIV infection in the multicenter AIDS cohort study. *J Infect Dis* 2008; **197**: 1011–5.
46 Kwiek JJ, Russell ES, Dang KK, *et al.* The molecular epidemiology of HIV-1 envelope diversity during HIV-1 subtype C vertical transmission in Malawian mother-infant pairs. *AIDS Lond Engl* 2008; **22**: 863–71.
47 Abrahams M-R, Anderson JA, Giorgi EE, *et al.* Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol* 2009; **83**: 3556–67.
48 Kearney M, Maldarelli F, Shao W, *et al.* Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J Virol* 2009; **83**: 2715–27.
49 Novitsky V, Lagakos S, Herzig M, *et al.* Evolution of proviral gp120 over the first year of HIV-1 subtype C infection. *Virology* 2009; **383**: 47–59.
50 Salazar-Gonzalez JF, Salazar MG, Keele BF, *et al.* Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 2009; **206**: 1273–89.
51 Fischer W, Ganusov VV, Giorgi EE, *et al.* Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PloS One* 2010; **5**: e12303–e12303.
52 Zhang H, Tully DC, Hoffmann FG, He J, Kankasa C, Wood C. Restricted genetic diversity of HIV-1 subtype C envelope glycoprotein from perinatally infected Zambian infants. *PloS One* 2010; **5**: e9294–e9294.
53 Boeras DI, Hraber PT, Hurlston M, *et al.* Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proc Natl Acad Sci U S A* 2011; **108**: E1156–63.
54 Collins-Fairclough AM, Charurat M, Nadai Y, *et al.* Significantly longer envelope V2 loops are characteristic of heterosexually transmitted subtype B HIV-1 in Trinidad. *PloS One* 2011; **6**.
55 Herbeck JT, Rolland M, Liu Y, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 2011; **85**: 7523–34.
56 Kishko M, Somasundaran M, Brewster F, Sullivan JL, Clapham PR, Luzuriaga K. Genotypic and functional properties of early infant HIV-1 envelopes. *Retrovirology* 2011; **8**: 67.
57 Nofemela A, Bandawe G, Thebus R, *et al.* Defining the human immunodeficiency virus type 1 transmission genetic bottleneck in a region with multiple circulating subtypes and recombinant forms. *Virology* 2011; **415**: 107–13.
58 Novitsky V, Wang R, Margolin L, *et al.* Transmission of single and multiple viral variants in primary HIV-1 subtype C infection. *PLoS One* 2011; **6**.
59 Rachinger A, Groeneveld PHP, van Assen S, Lemey P, Schuitemaker H. Time-measured phylogenies of gag, pol and env sequence data reveal the direction and time interval of HIV-1 transmission. *AIDS* 2011; **25**. https://journals.lww.com/aidsonline/Fulltext/2011/05150/Time_measured_phylogenies_of_gag,_pol_and_env.3.aspx.
60 Rieder P, Joos B, Scherrer AU, *et al.* Characterization of Human Immunodeficiency Virus Type 1 (HIV-1) Diversity and Tropism in 145 Patients With Primary HIV-1 Infection. *Clin Infect Dis* 2011; **53**: 1271–9.
61 Rolland M, Tovanabutra S, DeCamp AC, *et al.* Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat Med* 2011; **17**: 366–71.
62 Henn MR, Boutwell CL, Charlebois P, *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012; **8**: e1002529–e1002529.
63 Kiwelu IE, Novitsky V, Margolin L, *et al.* HIV-1 subtypes and recombinants in Northern Tanzania: distribution of viral quasispecies. *PLoS One* 2012; **7**.
64 Rossenkhan R, Novitsky V, Sebunya TK, Musonda R, Gashe BA, Essex M. Viral diversity and diversification of major non-structural genes vif, vpr, vpu, tat exon 1 and rev exon 1 during primary HIV-1 subtype C infection. *PloS One* 2012; **7**: e35491–e35491.
65 Sturdevant CB, Dow A, Jabara CB, *et al.* Central nervous system compartmentalization of HIV-1 subtype C variants early and late in infection in young children. *PLoS Pathog* 2012; **8**: e1003094–e1003094.
66 Baalwa J, Wang S, Parrish NF, *et al.* Molecular identification, cloning and characterization of transmitted/founder HIV-1 subtype A, D and A/D infectious molecular clones. *Virology* 2013; **436**: 33–48.
67 Frange P, Meyer L, Jung M, *et al.* Sexually-transmitted/founder HIV-1 cannot be directly predicted from plasma or PBMC-derived viral quasispecies in the transmitting partner. *PloS One* 2013; **8**: e69144–e69144.
68 Chaillon A, Gianella S, Wertheim JO, Richman DD, Mehta SR, Smith DM. HIV migration between blood and cerebrospinal fluid or semen over time. *J Infect Dis* 2014; **209**: 1642–52.
69 Sterrett S, Learn GH, Edlefsen PT, *et al.* Low multiplicity of HIV-1 infection and no vaccine enhancement in VAX003 injection drug users. In: Open forum infectious diseases. Oxford University Press, 2014.
70 Wagner GA, Pacold ME, Kosakovsky Pond SL, *et al.* Incidence and prevalence of intrasubtype HIV-1 dual infection in at-risk men in the United States. *J Infect Dis* 2014; **209**: 1032–8.
71 Chen Y, Li N, Zhang T, *et al.* Comprehensive Characterization of the Transmitted/Founder env Genes From a Single MSM Cohort in China. *J Acquir Immune Defic Syndr 1999* 2015; **69**: 403–12.
72 Danaviah S, de Oliveira T, Bland R, *et al.* Evidence of long-lived founder virus in mother-to-child HIV transmission. *PloS One* 2015; **10**: e0120389–e0120389.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

73 Deymier MJ, Ende Z, Fenton-May AE, *et al.* Heterosexual Transmission of Subtype C HIV-1 Selects Consensus-Like Variants without Increased Replicative Capacity or Interferon-α Resistance. *PLoS Pathog* 2015; **11**: e1005154–e1005154.

74 Gounder K, Padayachi N, Mann JK, *et al.* High frequency of transmitted HIV-1 Gag HLA class I-driven immune escape variants but minimal immune selection over the first year of clade C infection. *PloS One* 2015; **10**: e0119886–e0119886.

75 Le AQ, Taylor J, Dong W, *et al.* Differential evolution of a CXCR4-using HIV-1 strain in CCR5wt/wt and CCR5Δ32/Δ32 hosts revealed by longitudinal deep sequencing and phylogenetic reconstruction. *Sci Rep* 2015; **5**: 17607.

76 Zanini F, Brodin J, Thebo L, *et al.* Population genomics of intrapatient HIV-1 evolution. *Elife* 2015; **4**: e11282.

77 Chaillon A, Gianella S, Little SJ, *et al.* Characterizing the multiplicity of HIV founder variants during sexual transmission among MSM. *Virus Evol* 2016; **2**. DOI:10.1093/ve/vew012.

78 Love TMT, Park SY, Giorgi EE, Mack WJ, Perelson AS, Lee HY. SPMM: estimating infection duration of multivariant HIV-1 infections. *Bioinforma Oxf Engl* 2016; **32**: 1308–15.

79 Novitsky V, Moyo S, Wang R, Gaseitsiwe S, Essex M. Deciphering multiplicity of HIV-1C infection: transmission of closely related multiple viral lineages. *PloS One* 2016; **11**.

80 Oberle CS, Joos B, Rusert P, *et al.* Tracing HIV-1 transmission: envelope traits of HIV-1 transmitter and recipient pairs. *Retrovirology* 2016; **13**: 62.

81 Park SY, Mack WJ, Lee HY. Enhancement of viral escape in HIV-1 Nef by STEP vaccination. *AIDS Lond Engl* 2016; **30**: 2449–58.

82 Salazar-Gonzalez JF, Salazar MG, Tully DC, *et al.* Use of Dried Blood Spots to Elucidate Full-Length Transmitted/Founder HIV-1 Genomes. *Pathog Immun* 2016; **1**: 129–53.

83 Smith SA, Burton SL, Kilembe W, *et al.* Diversification in the HIV-1 Envelope Hyper-variable Domains V2, V4, and V5 and Higher Probability of Transmitted/Founder Envelope Glycosylation Favor the Development of Heterologous Neutralization Breadth. *PLoS Pathog* 2016; **12**: e1005989–e1005989.

84 DeCamp AC, Rolland M, Edlefsen PT, *et al.* Sieve analysis of breakthrough HIV-1 sequences in HVTN 505 identifies vaccine pressure targeting the CD4 binding site of Env-gp120. *PloS One* 2017; **12**: e0185959–e0185959.

85 Iyer SS, Bibollet-Ruche F, Sherrill-Mix S, *et al.* Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. *Proc Natl Acad Sci U S A* 2017; **114**: E590–9.

86 Kijak GH, Sanders-Buell E, Chenine A-L, *et al.* Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection. *PLoS Pathog* 2017; **13**: e1006510–e1006510.

87 Ashokkumar M, Aralaguppe SG, Tripathy SP, Hanna LE, Neogi U. Unique phenotypic characteristics of recently transmitted HIV-1 subtype C envelope glycoprotein gp120: use of CXCR6 coreceptor by transmitted founder viruses. *J Virol* 2018; **92**: e00063-18.

88 Dukhovlinova E, Masharsky A, Vasileva A, *et al.* Characterization of the Transmitted Virus in an Ongoing HIV-1 Epidemic Driven by Injecting Drug Use. *AIDS Res Hum Retroviruses* 2018; **34**: 867–78.

89 Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nat Microbiol* 2018; **3**: 983–8.

90 Lewitus E, Rolland M. A non-parametric analytic framework for within-host viral phylogenies and a test for HIV-1 founder multiplicity. *Virus Evol* 2019; **5**: vez044.

91 Sivay MV, Grabowski MK, Zhang Y, *et al.* Phylogenetic Analysis of Human Immunodeficiency Virus from People Who Inject Drugs in Indonesia, Ukraine, and Vietnam: HPTN 074. *Clin Infect Dis* 2019; published online Dec. DOI:10.1093/cid/ciz1081.

92 Todesco E, Wirden M, Calin R, *et al.* Caution is needed in interpreting HIV transmission chains by ultradeep sequencing. *Aids* 2019; **33**: 691–9.

93 Tovanabutra S, Sirijatuphat R, Pham PT, *et al.* Deep Sequencing Reveals Central Nervous System Compartmentalization in Multiple Transmitted/Founder Virus Acute HIV-1 Infection. *Cells* 2019; **8**: 902.

94 Brooks K, Jones BR, Dilernia DA, *et al.* HIV-1 variants are archived throughout infection and persist in the reservoir. *PLOS Pathog* 2020; **16**: e1008378.

95 Leda AR, Hunter J, Castro de Oliveira U, *et al.* HIV-1 genetic diversity and divergence and its correlation with disease progression among antiretroviral naïve recently infected individuals. *Virology* 2020; **541**: 13–24.

96 Liu Y, Jia L, Su B, *et al.* The genetic diversity of HIV-1 quasispecies within primary infected individuals. *AIDS Res Hum Retroviruses* 2020.

97 Martinez DR, Tu JJ, Kumar A, *et al.* Maternal Broadly Neutralizing Antibodies Can Select for Neutralization-Resistant, Infant-Transmitted/Founder HIV Variants. *mBio* 2020; **11**: e00176-20.

98 Rolland M, Tovanabutra S, Dearlove B, *et al.* Molecular dating and viral load growth rates suggested that the eclipse phase lasted about a week in HIV-1 infected adults in East Africa and Thailand. *PLoS Pathog* 2020; **16**: e1008179–e1008179.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

99 Villabona-Arenas ChJ, Hall M, Lythgoe KA, *et al.* Number of HIV-1 founder variants is determined by the recency of the source partner infection. *Science* 2020; **369**: 103 LP – 108.

100 Lee HY, Giorgi EE, Keele BF, *et al.* Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol* 2009; **261**: 341–60.

101 Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 1991; **129**: 555–62.

102 Simon V, Zennou V, Murray D, Huang Y, Ho DD, Bieniasz PD. Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. *PLoS Pathog* 2005; **1**: e6.

103 Bourara K, Liegler TJ, Grant RM. Target cell APOBEC3C can induce limited G-to-A mutation in HIV-1. *PLoS Pathog* 2007; **3**: e153.

104 Shaw GM, Hunter E. HIV transmission. *Cold Spring Harb Perspect Med* 2012; **2**: a006965.

105 Novitsky V, Wang R, Rossenkhan R, Moyo S, Essex M. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. *Infect Genet Evol* 2013; **19**: 361–8.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12. Supplementary

Supplementary **29**

S1: PRISMA Checklist 30

S2: Supplementary Methods 34

Full search query submitted to MEDLINE, EMBASE and Global Health databases 34

Software and Computational Methods 34

S3: Time Structure of Route of Exposure and Method 35

S4: Sensitivity Analyses for Pooling 36

S5: Leave-One-Out Cross Validation 37

S6: Evaluation of Publication Bias 38

S7: Sensitivity Analyses for Meta-regression (i) 39

S9: Supplementary References 40

12.1. S1: PRISMA Checklist

PRISMA-IPD Section/topic	Item No	Checklist item	Reported on page
Title			
Title	1	Identify the report as a systematic review and meta-analysis of individual participant data.	1
Abstract			
Structured summary	2	Provide a structured summary including as applicable:	2
		Background: state research question and main objectives, with information on participants, interventions, comparators and outcomes.	
		Methods: report eligibility criteria; data sources including dates of last bibliographic search or elicitation, noting that IPD were sought; methods of assessing risk of bias.	
		Results: provide number and type of studies and participants identified and number (%) obtained; summary effect estimates for main outcomes (benefits and harms) with confidence intervals and measures of statistical heterogeneity. Describe the direction and size of summary effects in terms meaningful to those who would put findings into practice.	
		Discussion: state main strengths and limitations of the evidence, general interpretation of the results and any important implications.	
		Other: report primary funding source, registration number and registry name for the systematic review and IPD meta-analysis.	
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of the questions being addressed with reference, as applicable, to participants, interventions, comparisons, outcomes and study design (PICOS). Include any hypotheses that relate to particular types of participant-level subgroups.	4
Methods			
Protocol and registration	5	Indicate if a protocol exists and where it can be accessed. If available, provide registration information including registration number and registry name. Provide publication details, if applicable.	2

Eligibility criteria	6	Specify inclusion and exclusion criteria including those relating to participants, interventions, comparisons, outcomes, study design and characteristics (e.g. years when conducted, required minimum follow-up). Note whether these were applied at the study or individual level i.e. whether eligible participants were included (and ineligible participants excluded) from a study that included a wider population than specified by the review inclusion criteria. The rationale for criteria should be stated.	5
Identifying studies - information sources	7	Describe all methods of identifying published and unpublished studies including, as applicable: which bibliographic databases were searched with dates of coverage; details of any hand searching including of conference proceedings; use of study registers and agency or company databases; contact with the original research team and experts in the field; open adverts and surveys. Give the date of last search or elicitation.	5
Identifying studies - search	8	Present the full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	33
Study selection processes	9	State the process for determining which studies were eligible for inclusion.	5
Data collection processes	10	Describe how IPD were requested, collected and managed, including any processes for querying and confirming data with investigators. If IPD were not sought from any eligible study, the reason for this should be stated (for each such study).	5-6
		If applicable, describe how any studies for which IPD were not available were dealt with. This should include whether, how and what aggregate data were sought or extracted from study reports and publications (such as extracting data independently in duplicate) and any processes for obtaining and confirming these data with investigators.	
Data items	11	Describe how the information and variables to be collected were chosen. List and define all study level and participant level data that were sought, including baseline and follow-up information. If applicable, describe methods of standardising or translating variables within the IPD datasets to ensure common scales or measurements across studies.	6
IPD integrity	A1	Describe what aspects of IPD were subject to data checking (such as sequence generation, data consistency and completeness, baseline imbalance) and how this was done.	6
Risk of bias assessment in individual studies.	12	Describe methods used to assess risk of bias in the individual studies and whether this was applied separately for each outcome. If applicable, describe how findings of IPD checking were used to inform the assessment. Report if and how risk of bias assessment was used in any data synthesis.	6-7
Specification of outcomes and effect measures	13	State all treatment comparisons of interests. State all outcomes addressed and define them in detail. State whether they were pre-specified for the review and, if applicable, whether they were primary/main or secondary/additional outcomes. Give the principal measures of effect (such as risk ratio, hazard ratio, difference in means) used for each outcome.	6

Synthesis methods	14	Describe the meta-analysis methods used to synthesise IPD. Specify any statistical methods and models used. Issues should include (but are not restricted to): · Use of a one-stage or two-stage approach. · How effect estimates were generated separately within each study and combined across studies (where applicable). · Specification of one-stage models (where applicable) including how clustering of patients within studies was accounted for. · Use of fixed or random effects models and any other model assumptions, such as proportional hazards. · How (summary) survival curves were generated (where applicable). · Methods for quantifying statistical heterogeneity (such as I^2 and t^2). · How studies providing IPD and not providing IPD were analysed together (where applicable). · How missing data within the IPD were dealt with (where applicable).	6-7
Exploration of variation in effects	A2	If applicable, describe any methods used to explore variation in effects by study or participant level characteristics (such as estimation of interactions between effect and covariates). State all participant-level characteristics that were analysed as potential effect modifiers, and whether these were pre-specified.	7
Risk of bias across studies	15	Specify any assessment of risk of bias relating to the accumulated body of evidence, including any pertaining to not obtaining IPD for particular studies, outcomes or other variables.	NA
Additional analyses	16	Describe methods of any additional analyses, including sensitivity analyses. State which of these were pre-specified.	7
Results			
Study selection and IPD obtained	17	Give numbers of studies screened, assessed for eligibility, and included in the systematic review with reasons for exclusions at each stage. Indicate the number of studies and participants for which IPD were sought and for which IPD were obtained. For those studies where IPD were not available, give the numbers of studies and participants for which aggregate data were available. Report reasons for non-availability of IPD. Include a flow diagram.	9
Study characteristics	18	For each study, present information on key study and participant characteristics (such as description of interventions, numbers of participants, demographic data, unavailability of outcomes, funding source, and if applicable duration of follow-up). Provide (main) citations for each study. Where applicable, also report similar study characteristics for any studies not providing IPD.	9-10
IPD integrity	A3	Report any important issues identified in checking IPD or state that there were none.	NA
Risk of bias within studies	19	Present data on risk of bias assessments. If applicable, describe whether data checking led to the up-weighting or down-weighting of these assessments. Consider how any potential bias impacts on the robustness of meta-analysis conclusions.	35

Results of individual studies	20	For each comparison and for each main outcome (benefit or harm), for each individual study report the number of eligible participants for which data were obtained and show simple summary data for each intervention group (including, where applicable, the number of events), effect estimates and confidence intervals. These may be tabulated or included on a forest plot.	11-18
Results of syntheses	21	Present summary effects for each meta-analysis undertaken, including confidence intervals and measures of statistical heterogeneity. State whether the analysis was pre-specified, and report the numbers of studies and participants and, where applicable, the number of events on which it is based.	19
		When exploring variation in effects due to patient or study characteristics, present summary interaction estimates for each characteristic examined, including confidence intervals and measures of statistical heterogeneity. State whether the analysis was pre-specified. State whether any interaction is consistent across trials.	
		Provide a description of the direction and size of effect in terms meaningful to those who would put findings into practice.	
Risk of bias across studies	22	Present results of any assessment of risk of bias relating to the accumulated body of evidence, including any pertaining to the availability and representativeness of available studies, outcomes or other variables.	NA
Additional analyses	23	Give results of any additional analyses (e.g. sensitivity analyses). If applicable, this should also include any analyses that incorporate aggregate data for studies that do not have IPD. If applicable, summarise the main meta-analysis results following the inclusion or exclusion of studies for which IPD were not available.	19+35
Discussion			
Summary of evidence	24	Summarise the main findings, including the strength of evidence for each main outcome.	21
Strengths and limitations	25	Discuss any important strengths and limitations of the evidence including the benefits of access to IPD and any limitations arising from IPD that were not available.	21-22
Conclusions	26	Provide a general interpretation of the findings in the context of other evidence.	23
Implications	A4	Consider relevance to key groups (such as policy makers, service providers and service users). Consider implications for future research.	22
Funding			
Funding	27	Describe sources of funding and other support (such as supply of IPD), and the role in the systematic review of those providing such support.	2 & 23

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12.2. S2: Supplementary Methods

12.2.1. Full search query submitted to MEDLINE, EMBASE and Global Health databases

(((((transmi*.af. or found*.af. or bottleneck.af. or single.af. or multiple.af. or multiplicity.af. or breakthrough.ti. or TF.af.) and (virus*.af. or variant*.af. or strain.af. or lineage.af. or phenotyp*.af.)) and (HIV.ti. or HIV-1.ti. or human immunodeficiency virus.ti. or env.ti. or envelope.ti or gag.ti. or pol.ti.)) and ((single genome amplification.af. or sga.af. or sgs.af. or ((sequencing.af. or characterized.af.) and (single genome.af. or deep.af. or whole genome.af. or full length.af. or full-length.af.))) or divers*.af. or distance.af. or poisson-fitter.af. or fitness.af. or (monophyletic.af. or paraphyletic.af. or polyphyletic.af.) or (phylogenetic*.af. and (clade.af. or topology.af. or tree.af. or linked.af. or diver*.af. or distance.af. or sieve.af. or molecular dating.af.)))) not ((SIV.ti,ab. or simian immunodeficiency.ti,ab. or fiv.ti,ab. or feline immunodeficiency virus.ti,ab. or exp Hepacivirus/ or Hepatitis.ti,ab. or exp Flaviviridae/ or Tuberculosis.ti,ab. or Enterovirus.ti,ab. or exp Spumavirus/ or diarrhoea.ti,ab. or diarrhea.ti,ab. or superinfection.ti. or exp Malaria/ or CMV.ti,ab. or HPV.ti,ab. or SHIV.ti,ab. OR exp HIV-2/ or phylogeo*.af. or network.ti. or exp HIV Protease Inhibitors/ or exp HIV Integrase Inhibitors/)))

Set to these databases:

- Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R)
- Global Health 1910 to 2020 Week 36
- EMBASE & EMBASE Classic 1947 – Sep 11

12.2.2. Software and Computational Methods

- All code associated with this study is available under GNU General Public License v3.0 at the following GitHub repository: foundervariantsHIV_sysreview.

- The analyses were conducted in R 3.6.1, using the following packages:
  lme4, 1.1-23, (Bates et al. 2007); metafor, 2.4-0, (Viechtbauer 2010); performance, 0.6.1, ; cowplot, 1.0.0, ; ggplot2, 3.3.2,; dplyr, 1.0.3, (Wickham et al. 2015); forcats, 0.5.0, ; mltools, 0.3.5, ; parallel, 3.6.1,; reshape2, 1.4.3, ; stringr, 1.4.0, ; tidyr, 1.0.

12.3. S3: Time Structure of Route of Exposure and Method

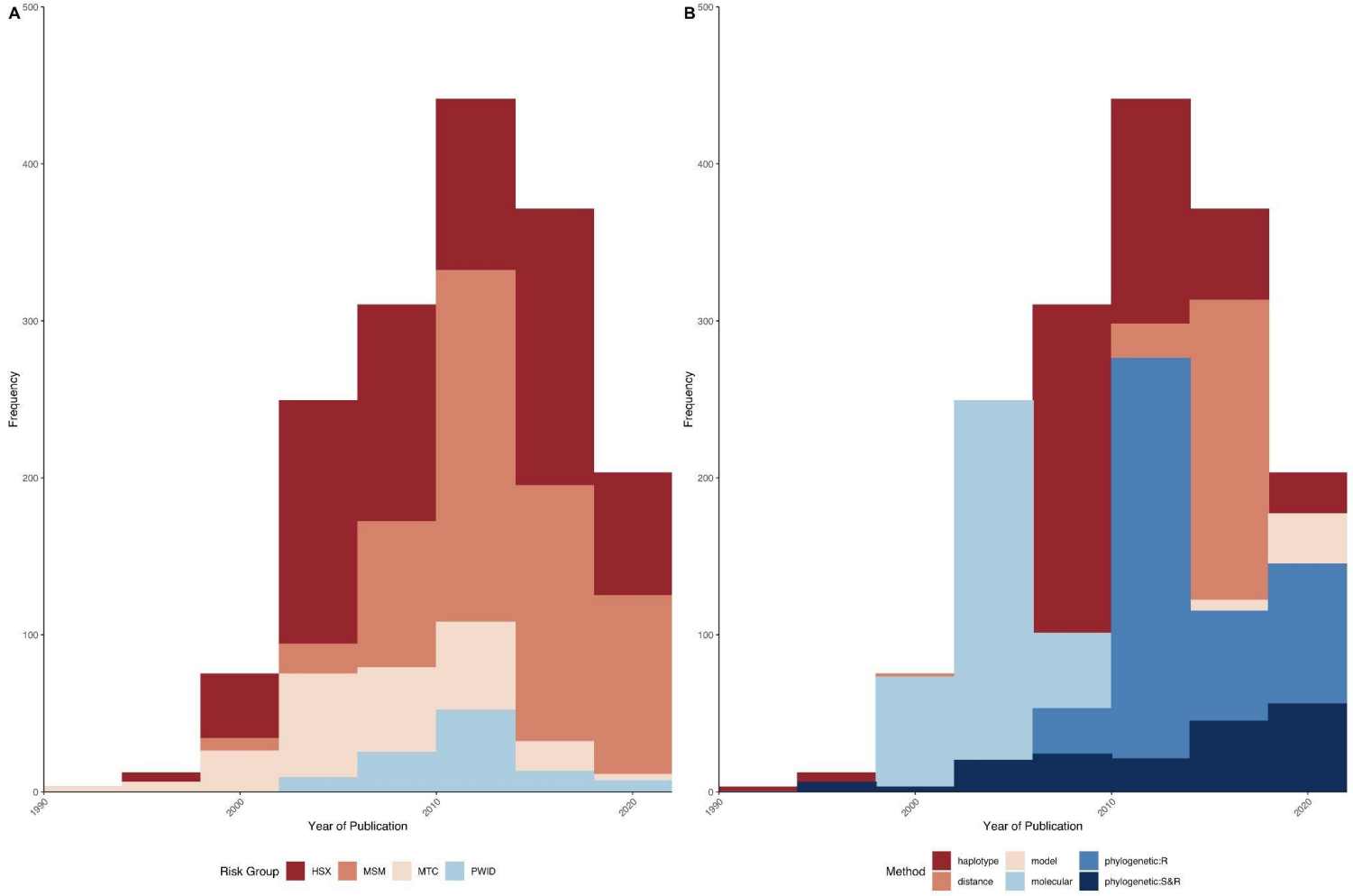

Figure S3: Distributions of transmission route (A) and grouped method (B) over time, highlighting the epidemiologic and methodological step-changes that occurred over the three decades in which the selected studies were published. Importantly, this means that earlier methods may be biased to those transmission routes that were more common in earlier studies.

12.4. S4: Sensitivity Analyses for Pooling

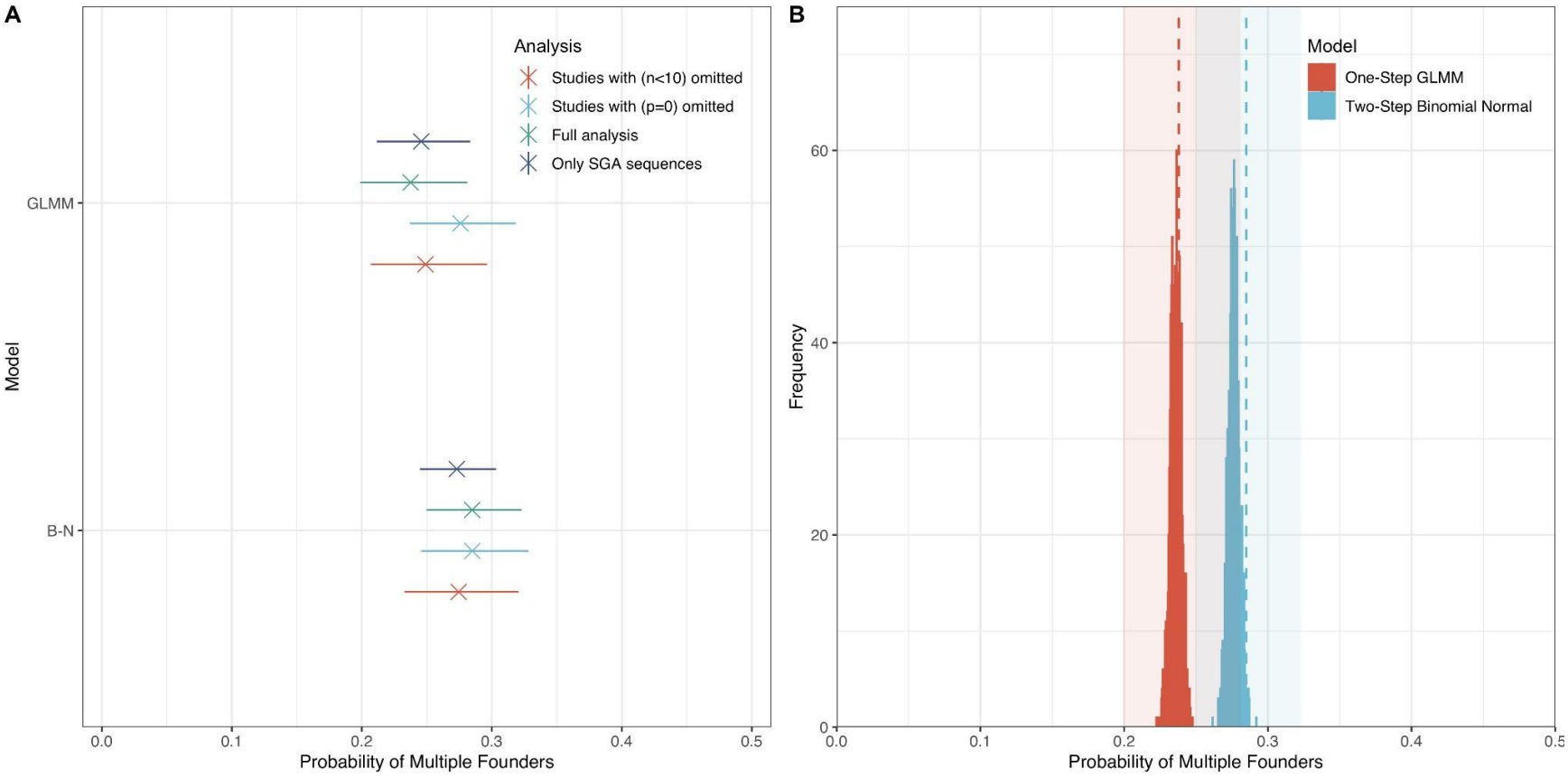

Figure S4: A visual comparison of the pooled estimates of the probability that an infection is initiated by multiple founders by the one-step (GLMM) and two-step (Binomial-Normal (B-N)) models and respective sensitivity analyses. Plot (A) shows both models calculate concordant estimates and are robust to sensitivity analyses designed to test our inclusion/exclusion criteria, and biases introduced by small or minimal-effect studies. B) reports the distribution of estimates, recalculated from 1000 datasets in which the representative datapoint for each individual was sampled at random from a pool of their possible measurements. The dashed lines and shaded areas denote the original point estimate and confidence intervals, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12.5. S5: Leave-One-Out Cross Validation

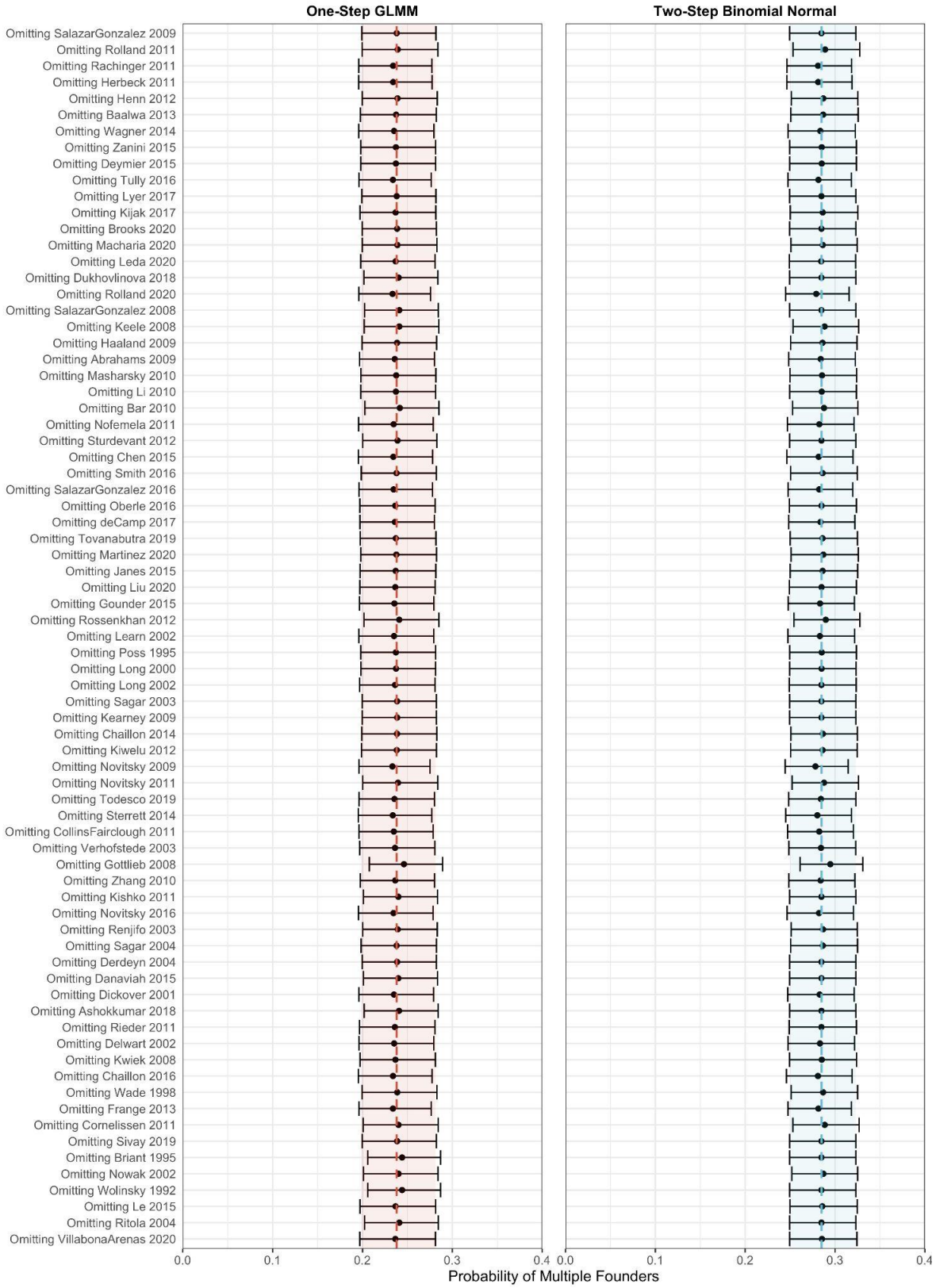

Figure S5: For both one-step and two-step models, we visually inspect the influence of each study included in our analysis on the pooled estimate that an infection is initiated by multiple founders. We find that in iteratively excluding individual studies, no discernible impact on the overall pooled estimate is made.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12.6. S6: Evaluation of Publication Bias

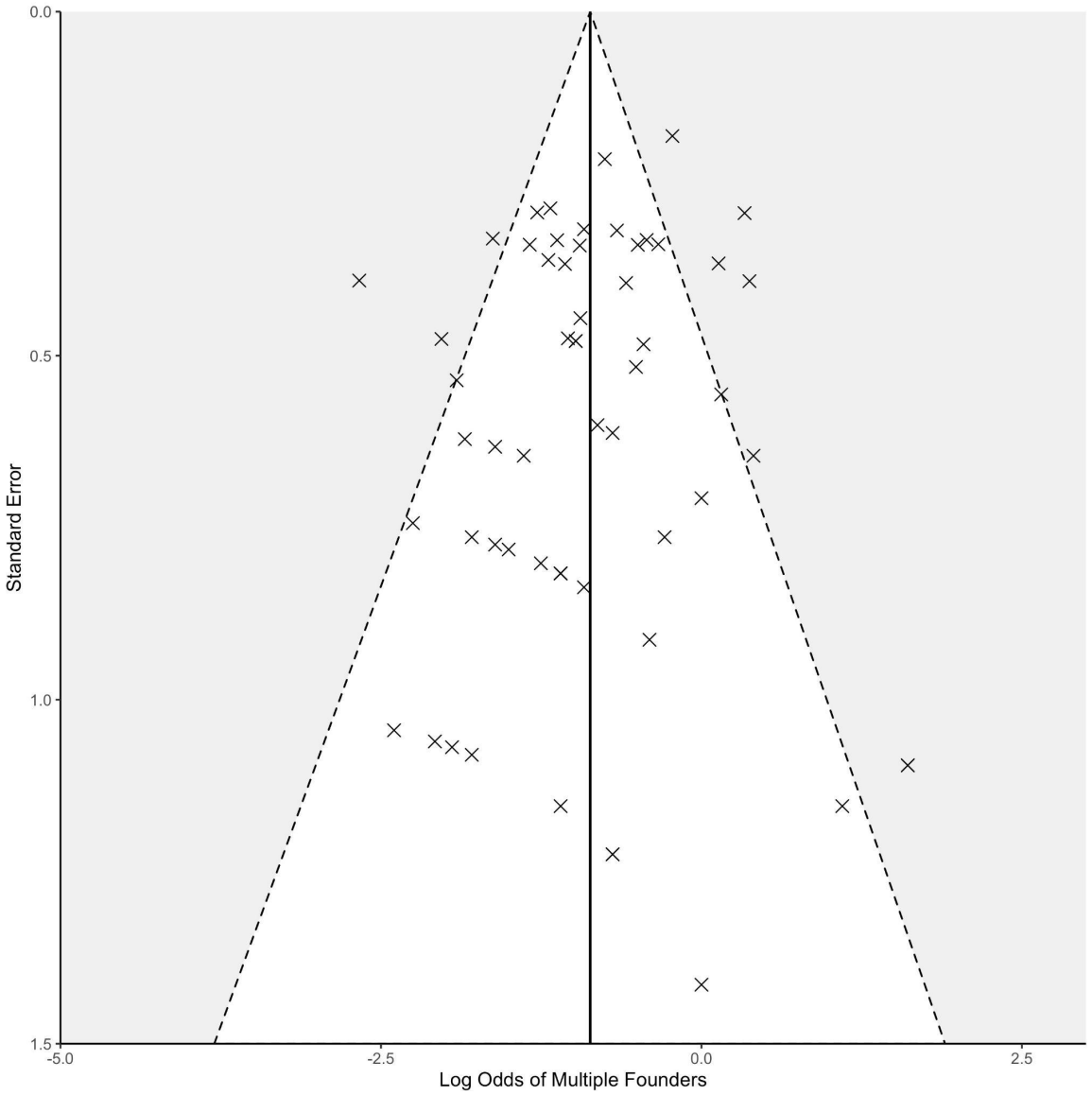

Figure S6: Funnel plot to visually evaluate the presence of publication bias. In the absence of publication bias, study estimates are distributed symmetrically with respect to the pooled estimate (vertical solid black line). Here, the log odds of an infection being initiated by multiple founders for each study, plotted against the standard error for each study indicate an absence of publication bias. This conclusion was supported by an (Egger's Regression Test: t = -0.2663, df = 56, p = 0.7910).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12.7. S7: Sensitivity Analyses for Meta-regression (i)

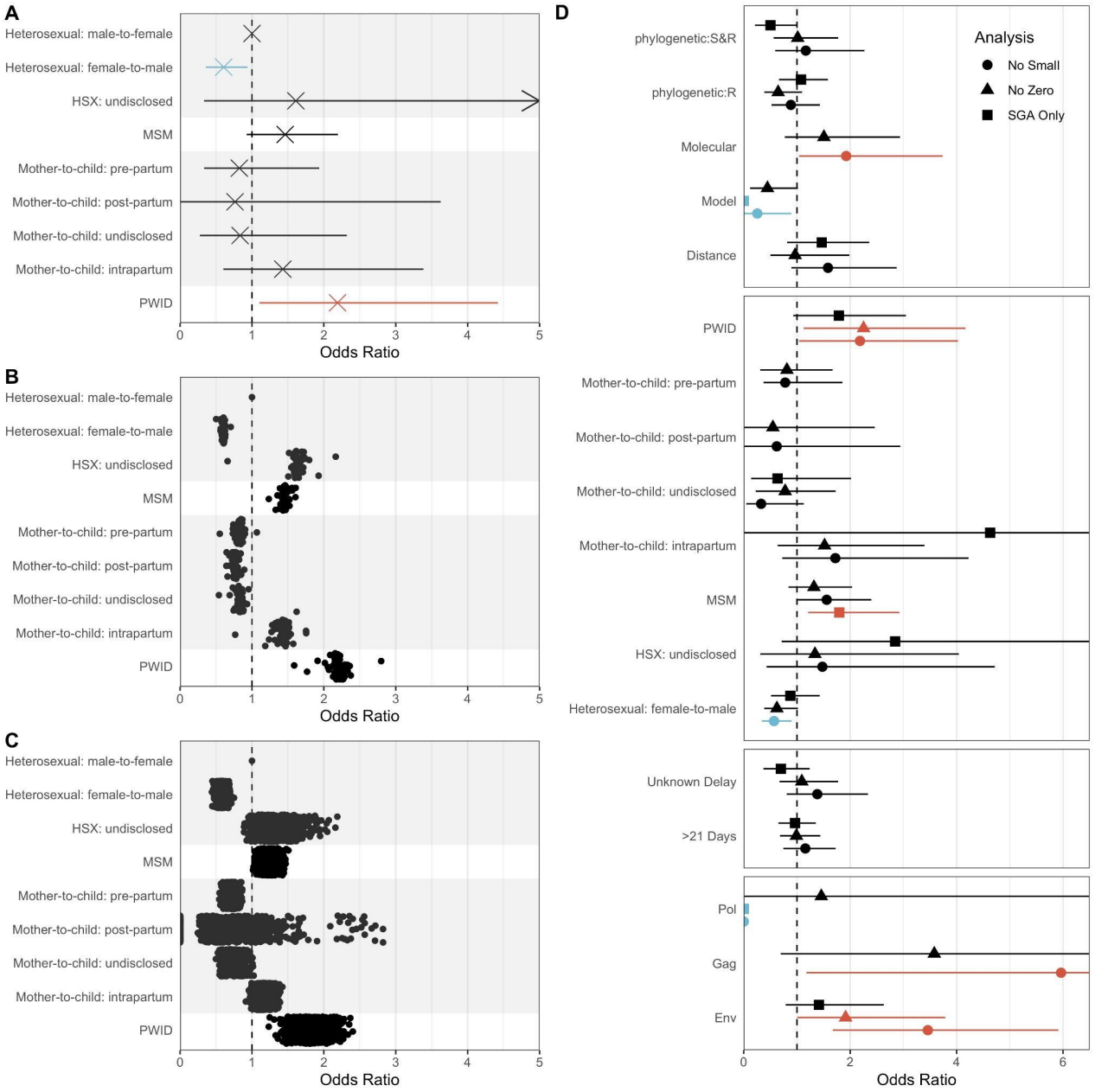

Figure S7: Odds ratios that an infection is initiated by multiple founders, stratified by route of transmission, as calculated in the main analysis (A), following the iterative exclusion of individual studies (B) and bootstrapped estimates recalculated from 1000 datasets in which the representative datapoint for each individual was sampled at random from a pool of their possible measurements (C). Panel (D) plots the odds ratios of all covariate levels included in the meta-regression, stratifying by previously defined sensitivity analyses. Overly generous confidence intervals in (D), particularly under the condition of single genome analysis (SGA) only data, is likely due to small sample sizes in at those levels (n<10).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.14.21259809; this version posted July 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12.8. S8: Supplementary References

Bates D, Sarkar D, Bates MD, Matrix L. 2007. The lme4 package. R Package Version 2:74.

Lüdecke D. 2018. ggeffects: Tidy data frames of marginal effects from regression models. J. Open Source Softw. 3:772.

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36:1–48.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer

Wickham H, Francois R, Henry L, Müller K. 2015. dplyr: A grammar of data manipulation. R Package Version 04 3.