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

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 transmission of multiple founder variants is most likely.

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 70 publications in our analysis, comprising 1657 individual patients. Our pooled estimate of the probability that an infection is initiated by multiple founder variants was 0·25 (95% CI: 0·21-0·29), with moderate heterogeneity ($Q = 132 \cdot 3, p < 0 \cdot 001, I^2 = 64 \cdot 2\%$). 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 predicted probability for female-to-male multiple variant transmission was significantly lower at 0·13 (95% CI: 0·08-0·20), while the probabilities for people-who-inject-drugs (PWID) and men-who-have-sex-with-men (MSM) transmissions were significantly higher at 0·37 (0·24-0·53) and 0·30 (0·33-0·40), respectively. There was no significant difference in the probability of multiple variant transmission between male-to-female transmission (0·21 (0·14-0·31)), post-partum mother-to-child (0·18 (0·03-0·57)), pre-partum mother-to-child (0·17 (0·08-0·33)), intrapartum mother-to-child (0·27 (0·14-0·40)).

Interpretation

We identified PWID and MSM 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 clinical outcomes.

Funding

This study was supported by the MRC Precision Medicine Doctoral Training Programme (ref: 2259239) and an ERC Starting Grant awarded to KEA (award number 757688). The funding sources played no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Panel: Research in context

Evidence before this study

Most 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 early 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

We estimated the probability that an HIV infection is initiated by multiple founder variants across exposure routes, leveraging individual patient data from 70 of the identified 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 exposure route. While overall, we estimated that 25% of infections are initiated by multiple founder variants, our analysis found that this probability for female-to-male transmission is significantly lower than for male-to-female transmission. By contrast, this probability was significantly higher among people-who-inject-drugs (PWID) and men-who-have-sex-with-men (MSM). There was no difference in the probability of multiple founder variant transmission for mother-to-child transmission when compared with male-to-female sexual transmission.

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 exposure affects the probability with which infections are initiated by multiple variants facilitates an improved understanding of how the evolution and epidemiology of HIV-1 determine clinical progression. Our results identify that PWID and MSM 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 these risk groups.

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 initiating infection after a single exposure.[3] Importantly, these infections caused by multiple founder variants are associated with elevated set point viral load and faster 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 per-act probability of transmission upon exposure is six times and eighteen times higher for transmission between people who inject drugs (PWID) and men who have sex with men (MSM) than for heterosexual transmission.[11]

Despite these differences in the probability of HIV-1 acquisition by route of exposure, there is currently no consensus about whether the route of exposure determines the probability that infection is initiated by multiple founder variants. Differences in selection pressure during transmission have been observed between sexual exposure routes, with less selection occurring during sexual transmission from males to females than vice-versa, and less selection during MSM transmission relative to heterosexual exposure overall.[12,13] Less selection should lead to more opportunities for infections initiated with more founder variants. Studies quantifying the number of founder variants are, however, inconsistent with these findings, which may be due to differences in methodology and study population.[3,12,14,15] Moreover, while acquisition risk during sexual transmission is known to be elevated during conditions that increase mucosal inflammation and compromise mucosal integrity, there is no consistent evidence that PWID transmissions, which bypass mucosal barriers altogether, are associated with a higher probability of founder variant initiation.[16,17] 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, and accounting for heterogeneity across methodology and study population.

Methods

Search Strategy and Eligibility Criteria

We searched MEDLINE, EMBASE and Global Health databases for papers published between 1 January 1990 to 14 September 2020 (Appendix S1, ppA2-A6). To be included, studies must have reported original estimates of founder variant multiplicity in people with acute or early HIV-1 infections, 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 founders. 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 known to be receiving antiretroviral treatment. No restrictions were placed on study design, geographic location, or age of participants. Studies were screened independently by SL and JB. Reviewers were blinded to study authorship during the title and abstract screens, and full text reviews were conducted independently before a consensus was reached; consulting other co-authors when necessary. This review conforms to PRISMA guidelines (Table S2, Appendix ppA7-A10).

Data Extraction

Individual patient data (IPD) were collated from all studies, with authors contacted if these data were not readily available. Studies were excluded from further analysis if IPD could not be 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 known to be receiving antiretroviral therapy. For the base-case dataset, we recorded whether an infection was initiated by one or multiple variants and eight predetermined covariates to be considered in the multivariable meta-regression:

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. *Quantification Method.* Methodological groupings were defined by the properties of each approach, resulting in six levels: phylogenetic, haplotype, distance, model, or molecular (Table S1).

iii. *HIV subtype*. Infecting subtypes were classed as either a canonical geographically delimited subtypes (A-D, F-H, J and K), a circulating recombinant form (CRF), or 'recombinant' (when a putative recombinant was identified but not designated a CRF).[18,19]

iv. *Delay between infection and sampling*. For sexual or PWID exposures, 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 reported after this period was classified as a delay of greater than 21 days.

v. *Number of genomes analysed per participant*. For studies that use single genome amplification, this was the number of consensus genomes obtained.

vi. *Genomic region analysed*. The region was classified as envelope (env), pol, gag or near full length genome (NFLG).

vii. *Alignment length analysed*. The length was measured in base pairs, discretised to the nearest 250, 500, 1000, 2000, 4000, 8000, and near full length genome (NFLG) intervals (~9000).

viii. *Use of single genome amplification (SGA) to generate viral sequences.* A binary classification was used to characterise whether the viral genomic data were generated using SGA. SGA mitigates the risk of Taq-polymerase mediated template switching, nucleotide misincorporation or unequal amplicons resampling encountered in regular bulk or near endpoint polymerase chain reaction (PCR) amplification.[20–22]

If information from any of the covariates iii-viii 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 main 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. Further details on covariate selection are in the supplementary methods (Appendix ppA2-A6).

Statistical Analysis

We calculated a pooled estimate of the probability of multiple founder variant infection using 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.[23] Heterogeneity was measured in terms of τ^2, the between-study variance; I^2, the percentage of variance attributable to study heterogeneity; and Cochran's Q, an indicator of larger variation between studies than of subjects within studies.[24] Publication bias was assessed using funnel plots and Egger's regression test.[25] All analyses were conducted in R 4.1.2.[26]

Pooled estimates obtained through a 'one-step' approach are usually congruent with the canonical 'two-step' meta-analysis model, however discrepancies may arise due to differences in weighting schemes, specification of the intercept or estimation of residual variances.[27] We compared the results from our 'one-step' model with a 'two-step' binomial-normal model to confirm our estimates were consistent. We also performed seven sensitivity analyses to test the robustness of our pooled estimate: i) iteratively excluding single studies, ii) excluding studies that contained fewer than ten participants, iii) setting variable thresholds of the number of genomes per patient, iv) excluding studies that consisted solely of single founder infections, v) excluding IPD that did not use single genome amplification, vi) including only those data that matched a 'gold-standard' methodology of haplotype-based methods and envelope gene analysis , and vii) an assessment of the effect of vaccine breakthrough, sequencing technologies, and molecular methods. To validate our down-sampling method that used only the most recent study for repeated individual data, we calculated a distribution of pooled estimates by refitting the pooling models to 1000 datasets, each containing one datapoint per individual sampled at random from an individual's possible measurements.

We extended our 'one-step' model by conducting a univariable meta-regression with each covariate contributing a fixed effect and assuming normally distributed random effects of publication. We extended this model to a multivariable analysis. Fixed effects were selected according to a 'keep it maximal' principle, in which covariates were only removed to facilitate a non-singular fit and to prevent multicolinearity.[28] We defined our reference case as heterosexual male-to-female transmission, and evaluated through a gold-standard methodology of haplotype-based methods, analysis of the envelope genomic region and a sampling delay of less than 21 days. We report stratified model estimates of the proportion of infections initiated by multiple founders and bootstrapped 95% confidence intervals across each covariate with all other covariates held at their reference case values. We performed four sensitivity analyses to test the robustness of the selected multivariable meta-regression model: i) iteratively excluding single studies, ii) excluding studies that contained fewer than ten participants, iii) excluding studies that consisted solely of single founder infections, and iv) excluding IPD that did not use single genome amplification. The re-sampling sensitivity analysis was repeated on our selected multivariable model as described above for the univariable model. Further details are in the supplementary methods (Appendix ppA2-A6).

Role of Funding Source

The funder of the study played no role in study design, data collection, data analysis, data interpretation, or writing of the report.

214

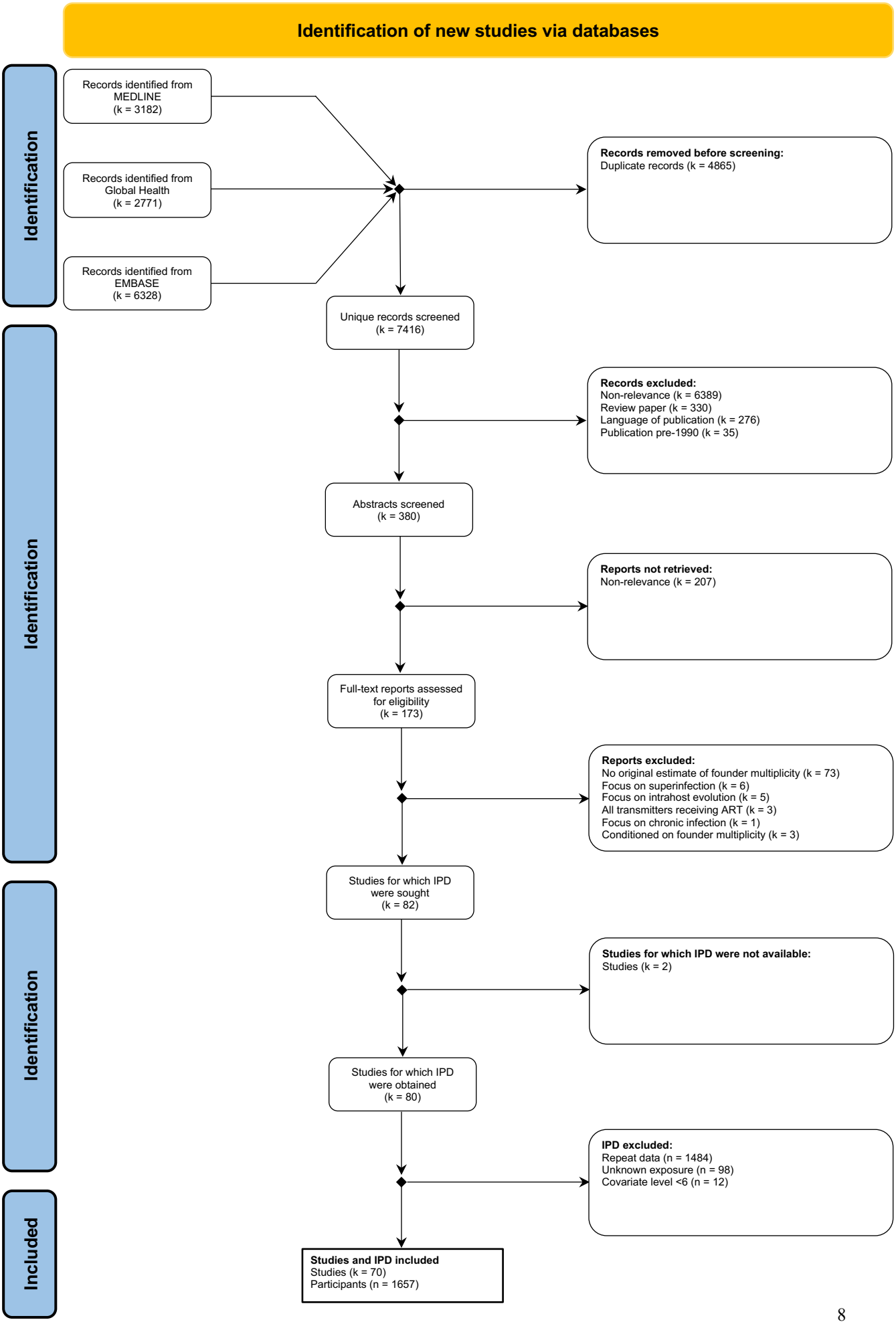

Identification of new studies via databases
Identification
Records identified from MEDLINE (k = 3182)
Records identified from Global Health (k = 2771)
Records identified from EMBASE (k = 6328)
Records removed before screening:
Duplicate records (k = 4865)
Unique records screened (k = 7416)
Identification
Records excluded:
Non-relevance (k = 6389)
Review paper (k = 330)
Language of publication (k = 276)
Publication pre-1990 (k = 35)
Abstracts screened (k = 380)
Reports not retrieved:
Non-relevance (k = 207)
Full-text reports assessed for eligibility (k = 173)
Reports excluded:
No original estimate of founder multiplicity (k = 73)
Focus on superinfection (k = 6)
Focus on intrahost evolution (k = 5)
All transmitters receiving ART (k = 3)
Focus on chronic infection (k = 1)
Conditioned on founder multiplicity (k = 3)
Identification
Studies for which IPD were sought (k = 82)
Studies for which IPD were not available:
Studies (k = 2)
Studies for which IPD were obtained (k = 80)
IPD excluded:
Repeat data (n = 1484)
Unknown exposure (n = 98)
Covariate level <6 (n = 12)
Included
Studies and IPD included
Studies (k = 70)
Participants (n = 1657)

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

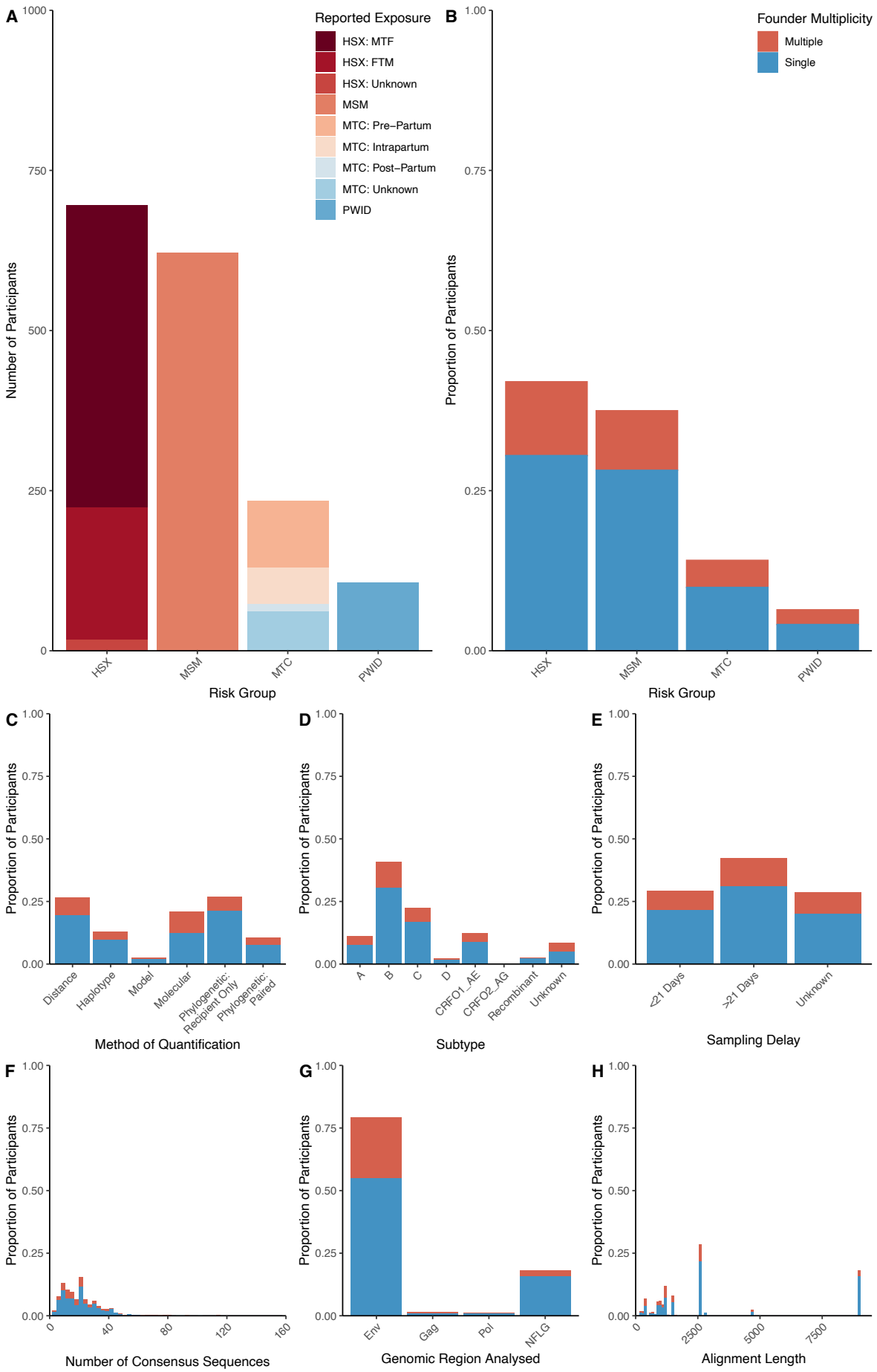
Results

Our search found 7416 unique papers, of which 7334 were excluded. Of the remaining 380, 207 were excluded after abstract screening, leaving a total of 82 eligible studies for IPD collation (Fig.1).[3,5–7,12,14–17,22,29–102] We 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, 1484 were excluded from our base case dataset 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 pertained to a covariate level that did not meet the minimum number (6) of observations across all participants, the base case dataset for our analysis comprised estimates from 1657 unique patients across 70 studies.

Our base case dataset includes a median of 13 participants per study (range 2-124) and represents infections associated with heterosexual transmission (42·0%, (n=696), MSM transmission (37·4%, n=621), MTC (14·1%, n=234), and PWID transmission (6·4%, n=106) (Fig.2; Table 1; Table S2, Appendix ppA11-17). Among heterosexual transmissions, 67·7% (n=471) were HSX:MTF transmissions, 29·9% (n=208) were HSX:FTM 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 over time (Fig. S1, Appendix ppA17). Across the base case dataset, 37·1% (n=618) estimates used phylogenetic methods, 26·4% (n=438) used haplotype methods, 20·9% (n=347) used molecular methods, and 13·0% (n=215) and 2·35% (n=39) of estimates were inferred using distance and model-based methods respectively (Fig.2).

Transmission Route	Number of studies	Number of participants	Number of participants where multiple founder variant estimated	Quantification methods		Genomic regions analysed		Number of sequences per participant
Heterosexual:								
MTF	32 (39)	471 (601)	147 (188)	Distance Haplotype Model Molecular Phylogenetic:R Phylogenetic:S&R	65 105 9 161 99 32	Env Gag Pol NFLG	437 14 1 19	21 [2-104]
FTM	25 (30)	208 (319)	39 (61)	Distance Haplotype Model Molecular Phylogenetic:R Phylogenetic:S&R	67 73 2 26 25 15	Env Gag Pol NFLG	179 8 3 18	22 [3-80]
Unknown	3 (4)	17 (22)	5 (7)	Distance Haplotype Phylogenetic:S&R	2 4 11	Env NFLG	15 2	13 [5-27]
MSM	28 (34)	621 (812)	154 (205)	Distance Haplotype Model Molecular Phylogenetic:R Phylogenetic:S&R	80 139 10 27 305 60	Env Pol NFLG	351 13 257	15 [2-149]
PWID	12 (13)	106 (116)	38 (45)	Distance Haplotype Model Molecular Phylogenetic:R Phylogenetic:S&R	1 63 14 9 14 5	Env Pol NFLG	101 1 4	24 [11-163]
MTC:								
Pre-partum	7 (7)	104 (104)	31 (31)	Model Molecular Phylogenetic:S&R	2 92 10	Env Gag	103 1	31 [6-49]
Intrapartum	7 (7)	57 (57)	25 (25)	Model Molecular Phylogenetic:S&R	2 32 23	Env	57	17 [6-31]
Post-partum	1 (1)	11 (11)	2 (2)	Phylogenetic:S&R	11	Env	11	··
Unknown	6 (6)	62 (62)	12 (12)	Haplotype Phylogenetic:S&R	54 8	Env Gag	61 1	37 [4-115]

Table 1: Summary of individual and study characteristics in our base case dataset. Transmission groups recorded as: female-to-male (HSX:FTM), male-to-female (HSX:MTF), men-who-have-sex-with-men (MSM), mother-to-child pre-partum, intrapartum), and post-partum; people who inject drugs (PWID). Numbers within parentheses refer to quantities before removal of repeat participants.



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

Our binomial GLMM pooled estimated the probability that an infection is initiated by multiple founder variants as 0·25 (95% CI: 0·21-0·29), identifying moderate heterogeneity ($Q = 132 \cdot 3, p < 0 \cdot 001$, $I^2 = 64 \cdot 2\%$). Visual inspection of a funnel plot and a non-significant Egger's Test (t = -0·7495, df = 55, p = 0·4568) were consistent with an absence of publication bias (Fig.S9 Appendix ppA26). Sensitivity analyses revealed the pooled estimate was 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.S2, Appendix ppA19). While restricting the analysis to participants for whom a large (>28) number of sequences were analysed did not change the pooled estimate (0·26 (0·20-0·34)), restricting the analysis to those individuals with fewer than 11 sequences reduced the estimate to 0·21 (0·17-0·25) (Fig.S3 Appendix ppA20). Analysing only data that matched our 'gold standard' study methodology slightly increases the pooled estimate (0.28 (95% CI: 0.22-0.35)) (Fig.S3, Appendix ppA20). We did not identify any studies or risk groups that individually influenced the pooled estimate significantly (Fig.S4, Fig.S5 Appendix ppA21-A22). A pooled estimate subgroup analysis of placebo and vaccine participants from studies for which vaccination status was available revealed no discernible influence of trial arm (Fig.S6, Appendix ppA23). Likewise, no discernible difference was identified between sequencing technologies on the pooled estimate (Fig.S7, Appendix ppA24).

We first extended our binomial GLMM with univariable fixed effects. Relative to a reference exposure route of HSX:MTF, we found significantly lower odds of HSX:FTM transmission being initiated by multiple founder variants (Odds Ratio (OR): 0·53 (95% CI 0·33-0·85)), while other exposure routes were not significantly different (Table 2). The univariable analyses also indicated significantly lower odds of identifying multiple founder variants when the near-full-length genome (NFLG) was analysed (OR: 0·38 (95% CI:0·19-0·68)), relative to the envelope genomic region, while molecular methods resulted in significantly greater odds (OR: 1·93 (1·02-3·45)), relative to haplotype methods. NFLG individuals continued to indicate significantly lower odds of identifying multiple founder variants in the absence of individuals analysed using molecular methods (Fig.S8, Appendix ppA25).

Next, we used a multivariable model to calculate the probability of multiple founder variants across the seven routes of exposure controlling for method, genomic region, and sampling delay (Fig.3, Table 2). A satisfactory fit was confirmed by inspection of binned residuals superimposed over 95% confidence intervals (Fig. S10, Appendix pp A27). Model estimated probabilities were calculated with respect to our 'gold standard' methodology. Compared to a HSX:MTF transmission probability of 0·21 (95% CI: 0·14-0·31), we found that HSX:FTM transmissions were less likely to be initiated by multiple founders than male-to-female transmissions, with probability 0·13 (95% CI: 0·08-0·21) (OR: 0·55 (95% CI 0·34-0·88)). Conversely, PWID and MSM transmissions were more likely to be initiated by multiple founders (0·37 (0·24-0·53) and 0·30 (0·22-0·40), respectively), compared to HSX:MTF (OR: 2·18 (1·11-3·89); 1·61 (1·00-2·34)) (Fig. 3A). Stratifying MTC transmissions by the putative timing of infection, we calculated pre-partum exposures were initiated by multiple founders with probability 0·17 (0·08-0·33), post-partum with probability 0·18 (0·03-0·57), and intrapartum transmissions with probability 0·27 (0·14-0·45).

We calculated the accuracy of different methods by comparing their estimated probability of multiple founder variants to a gold-standard methodological reference scenario of haplotype-based methods on whole genome sequences with individuals with less than 21 delays between infection and sampling. Our analysis indicates using model-based methods underestimates the chance of multiple founder variants (OR: 0·36 (95% CI: 0·09-0·87)), while using molecular methods results in an overestimation (OR: 2·05 (1·09-3·53)). Compared to the envelope genomic

region, analysis of near-full-length genome fragments likely underestimates the proportion of multiple founder infections (OR: 0·31 (95% CI: 0·13-0·62)). Our sensitivity analyses revealed the odds ratios calculated using 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.S11, Appendix ppA28).

		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	**0·53 [0·33-0·85]**	**0·006**	**0·55 [0·34-0·90]**	**0·011**
	Heterosexual: undisclosed	1·81 [0·40-5·91]	0·340	1·72 [0·25-5·24]	0·364
	MSM	1·33 [0·83-2·03]	0·238	**1·61 [1·00-2·34]**	**0·023**
	Mother-to-child: pre-partum	1·26 [0·54-2·65]	0·589	0·76 [0·38-1·59]	0·479
	Mother-to-child: intrapartum	1·87 [0·81-4·02]	0·148	1·34 [0·58-2·86]	0·461
	Mother-to-child: post-partum	0·73 [0·01-3·56]	0·772	0·79 [0·00-3·58]	0·794
	Mother-to-child: undisclosed	1·23 [0·45-3·50]	0·701	0·79 [0·29-2·23]	0·637
	PWID	2·08 [0·91-4·15]	0·05	**2·18 [1·17-3·89]**	**0·018**
Quantification Method					
	Haplotype	1 (reference)	-	1 (reference)	
	Distance	0·76 [0·35-1·58]	0·443	1·46 [0·76-2·86]	0·251
	Model	0·53 [0·09-1·39]	0·265	**0·36 [0·09-0·87]**	**0·057**
	Molecular	**1·93 [1·02-3·45]**	**0·026**	**2·05 [1·09-3·53]**	**0·018**
	Phylogenetic: recipient only	0·72 [0·42-1·24]	0·234	0·83 [0·48-1·54]	0·473
	Phylogenetic: source & recipient	0·90 [0·49-1·57]	0·730	0·95 [0·52-1·92]	0·852
Genomic Region					
	Envelope	1 (reference)	-	1 (reference)	-
	NFLG	**0·38 [0·19-0·63]**	**0·002**	**0·31[0·13-0·62]**	**0·000**
	Gag	1·13 [0·22-4·51]	0·857	2·14[0·45-6·80]	0·220
	Pol	0·31 [0·00-1·36]	0·171	0·31[0·00-1·13]	0·155
Sampling Delay					
	<21 Days	1 (reference)	-	1 (reference)	-
	>21 Days	1·09 [0·74-1·61]	0·629	1·16 [0·78-1·65]	0·434
	Unknown	1·42 [0·84-2·72]	0·201	1·39 [0·81-2·43]	0·220

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

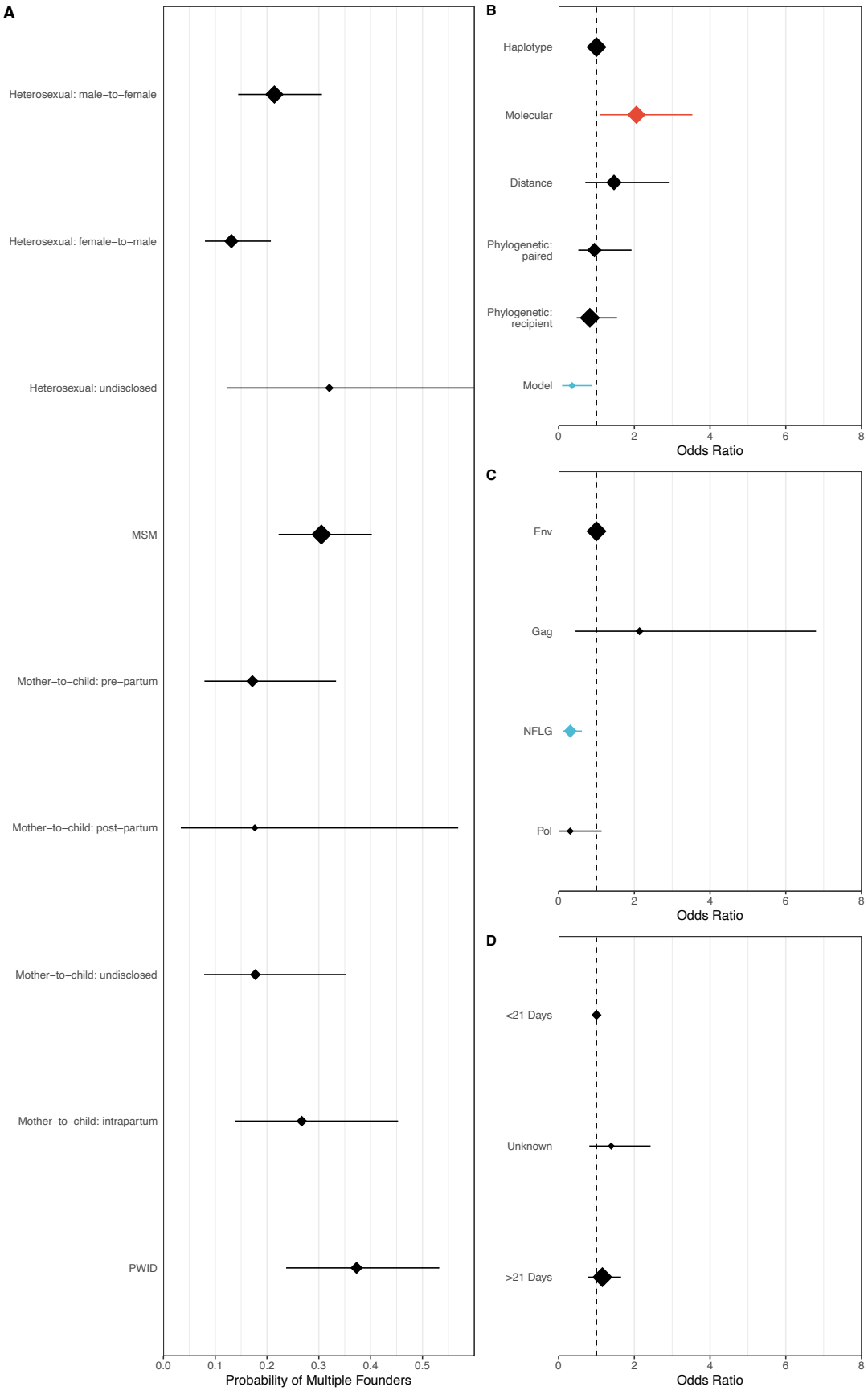


Figure 3: Model estimated probabilities and coefficients obtained from the multivariable model. A) Model estimated probabilities of an infection being initiated by multiple founder variants, stratified by the route of exposure. B-D) Inferred odds ratios of fixed effects variables. Blue denotes that a 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 (dotted line) and the marker size scales with sample size.

Discussion

Using data from 70 published 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 from 0·21 (95% CI: 0·14-0·31) for male-to-female infections, to 0·13 (95% CI: 0·08-0·21) for female-to-male infections, but increased for MSM and PWID infections (0·30 (0·22-0·40) and 0·37 (0·24-0·53), respectively). 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 whereas molecular methods overestimated. 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, however, were marginally higher than those of previous studies. A nine-study meta-analysis of 354 subjects found 0·34 of PWID infections were initiated by multiple founders compared with 0·37 (0·24-0·53) in our study and 0·25 for MSM infections for which we calculated (0·30 (0·22-0·40) [12] An earlier meta-analysis of five studies and 235 subjects also 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-significant 1·5-fold increase was observed with respect to MSM transmissions.[16] In both instances, these studies restricted participants so that the methodology in estimating founder variant multiplicity was consistent. In this study, we were able to leverage individual level data to control for methodological sources of heterogeneity across publications.

Across sexual transmission routes, the probability of multiple founder variants is positively, albeit weakly, associated with an increase in the risk of transmission given exposure. Nonetheless, the probability that infection is initiated by multiple founders remain remarkable consistent. For example, while male-to-male exposures may be up to eighteen times more likely to result in transmission than male-to-female exposures, we calculated a 1·6 fold increase in the risk of multiple founders.[11] Previously, Thompson et al. reconciled the low probability of acquisition with the relatively high probability of multiple founders by assuming only a fraction of exposures occur in environments conducive for transmission.[103] In sexual transmission, this could be induced through epithelial damage arising from ulceration or microtrauma; enhancing translocation of viral particles or driving inflammation that propagates recruitment of permissive target cells.[8] Despite a higher constitutive abundance of permissive cells in the adult human foreskin, the endocervical epithelium and its junction with the ectocervical epithelium are much more susceptible to inflammation and micro-abrasions, reflecting the transmission bias observed in heterosexual transmission.[104,105]

Our analysis has some limitations. First, our classification of founder variant multiplicity is determined by the individual studies, but explicitly defining a founder variant remains challenging. Recent studies have suggested a continuum of genotypic diversity exists, rather than discrete variants, that gives rise to distinct phylogenetic

diversification trajectories and may not be reflected by a binary classification.[32,68] Although a threshold is specified for distance-based methods, this often varies between publications.[106,107] 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 recombination and hypermutation.[108,109] Our finding that the analysis of near-full-length genomes were associated with a significant decrease in the odds of multiple founders, suggests earlier studies that rely on smaller, highly variable, fragments of envelop, may have overestimated the frequency of infections initiated by multiple founder variants. Similarly, our sensitivity analyses revealed a subtle correlation between the number of genomes analysed and the probability of observing multiple founder variants, pointing to the possibility that using too few genomes could limit the chance of observing multiple founders.

Second, we acknowledge that some heterogeneity associated with our estimates is encapsulated within the classification of route of exposure. Relying on self-reported route of exposure may bias our results if misclassification occurs systematically across studies. Similarly, insufficient data were available to properly consider risk factors such as genital ulceration, early stage of disease in the transmitter or receptive anal intercourse. These risk factors may confound or mediate any association between the exposure type and the probability of multiple founder variants, potentially hindering a deeper mechanistic understanding as to the risk factors underpinning founder variant multiplicity.[11] Also, under the hypothesis that the proportion of infections initiated by multiple founders varies by transmission route, our point estimate will be influenced by their relative proportions in our dataset. Globally, it is estimated that 70% of infections are transmitted heterosexually, compared to 42.2% in our dataset.[110] 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.

Finally, for several covariates the bootstrapped confidence intervals are wide and may lead to some uncertainty. These are a product of small sample sizes for certain observations, combined with the random effect of publication used in the meta-regression.

This systematic review and meta-analysis demonstrate that infections initiated by multiple founders account for a quarter of HIV-1 infections across major routes of transmission. We find that transmissions involving PWID and MSM are significantly more likely to be initiated by multiple founder variants, whilst HSX:FTM infections are significantly less likely, relative to HSX:MTF 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.

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. JB and KEA verified the data. All authors interpreted the data. JB and KEA drafted the manuscript, with critical revisions from all authors. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of Interests

The authors declare no competing interests.

Data Sharing

Code and individual patient data used in this study is publicly available at https://github.com/J-Baxter/foundervariantsHIV_sysreview.

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 Morgane Rolland for agreeing to share additional individual patient data with the authors to complete this study. We thank the four anonymous reviewers for their helpful feedback.

404 # References

1 Zhu T, Mo H, Wang N, *et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary
406 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
408 external envelope protein of human immunodeficiency virus type 1 upon primary infection. *Journal of Virology* 1993; **67**: 3345 LP – 3356.

410 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. *Proceedings of the National Academy of Sciences* 2008; **105**:
412 7552–7.

4 Sagar M, Lavreys L, Baeten JM, *et al.* Infection with multiple human immunodeficiency virus type 1 variants is
414 associated with faster disease progression. *Journal of virology* 2003; **77**: 12921–6.

5 Cornelissen M, Pasternak AO, Grijsen ML, *et al.* HIV-1 Dual Infection Is Associated With Faster CD4+ T-Cell
416 Decline in a Cohort of Men With Primary HIV Infection. *Clinical Infectious Diseases* 2012; **54**: 539–47.

6 Janes H, Herbeck JT, Tovanabutra S, *et al.* HIV-1 infections with multiple founders are associated with higher
418 viral loads than infections with single founders. *Nature Medicine* 2015; **21**: 1139.

7 Macharia GN, Yue L, Staller E, *et al.* Infection with multiple HIV-1 founder variants is associated with lower
420 viral replicative capacity, faster CD4+ T cell decline and increased immune activation during acute infection. *PLoS pathogens* 2020; **16**: e1008853–e1008853.

422 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
424 of founder viruses. *Nature reviews Microbiology* 2015; **13**: 414–25.

10 Talbert-Slagle K, Atkins KE, Yan K-K, *et al.* Cellular Superspreaders: An Epidemiological Perspective on HIV
426 Infection inside the Body. *PLOS Pathogens* 2014; **10**: e1004092.

11 Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a
428 systematic review. *AIDS* 2014; **28**.

12 Tully DC, Ogilvie CB, Batorsky RE, *et al.* Differences in the Selection Bottleneck between Modes of Sexual
430 Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. *PLoS pathogens* 2016; **12**: e1005619–e1005619.

432 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.

434 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 Pathogens* 2009; **5**: e1000274.

436 15 Li H, Bar KJ, Wang S, *et al.* High multiplicity infection by HIV-1 in men who have sex with men. *PLoS pathogens* 2010; **6**: e1000890.

438 16 Bar KJ, Li H, Chamberland A, *et al.* Wide variation in the multiplicity of HIV-1 infection among injection drug users. *Journal of virology* 2010; **84**: 6241–7.

440 17 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. *The Journal of Infectious Diseases*
442 2010; **201**: 1697–702.

18 Robertson DL, Anderson JP, Bradac JA, *et al.* HIV-1 nomenclature proposal. *Science* 2000; **288**: 55.

444 19 Archer J, Robertson DL. Understanding the diversification of HIV-1 groups M and O. *Aids* 2007; **21**: 1693–700.

20 Meyerhans A, Vartanian J-P, Wain-Hobson S. DNA recombination during PCR. *Nucleic acids research* 1990; **18**: 1687–91.

21 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. *Journal of virology* 1990; **64**: 864–72.

22 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. *Journal of virology* 2008; **82**: 3952–70.

23 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. *Statistics in medicine* 2020; **39**: 2536–55.

24 Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. John Wiley & Sons, 2011.

25 Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Bmj* 1997; **315**: 629–34.

26 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2021 https://www.R-project.org/.

27 Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in medicine* 2017; **36**: 855–75.

28 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.

29 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. *Journal of virology* 2004; **78**: 7279–83.

30 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. *Journal of virology* 2009; **83**: 3556–67.

31 Sagar M, Kirkegaard E, Lavreys L, Overbaugh J. Diversity in HIV-1 envelope V1–V3 sequences early in infection reflects sequence diversity throughout the HIV-1 genome but does not predict the extent of sequence diversity during chronic infection. *AIDS Research & Human Retroviruses* 2006; **22**: 430–7.

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

33 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. *Journal of virology* 1995; **69**: 3778–88.

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

35 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. *Journal of general virology* 1998; **79**: 1055–68.

36 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 pathogens* 2012; **8**: e1003094–e1003094.

37 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. *Clinical Infectious Diseases* 2011; **53**: 1271–9.

38 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 research and human retroviruses* 2018; **34**: 867–78.

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

40 Rossenkhan R, Rolland M, Labuschagne JPL, *et al.* Combining Viral Genetics and Statistical Modeling to Improve HIV-1 Time-of-Infection Estimation towards Enhanced Vaccine Efficacy Assessment. *Viruses* 2019; **11**: 607.

41 Chen Y, Li N, Zhang T, *et al.* Comprehensive Characterization of the Transmitted/Founder env Genes From a Single MSM Cohort in China. *Journal of acquired immune deficiency syndromes (1999)* 2015; **69**: 403–12.

42 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**.

43 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.

44 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.

45 Herbeck JT, Rolland M, Liu Y, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *Journal of virology* 2011; **85**: 7523–34.

46 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 pathogens* 2016; **12**: e1005989–e1005989.

47 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. *Journal of virology* 1995; **69**: 8118–22.

48 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. *Journal of virology* 2003; **77**: 3050–7.

49 Park SY, Mack WJ, Lee HY. Enhancement of viral escape in HIV-1 Nef by STEP vaccination. *AIDS (London, England)* 2016; **30**: 2449–58.

50 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.

51 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.

52 Long EM, Martin HL, Kreiss JK, *et al.* Gender differences in HIV-1 diversity at time of infection. *Nature medicine* 2000; **6**: 71–5.

53 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. *The Journal of experimental medicine* 2009; **206**: 1273–89.

54 Rolland M, Tovanabutra S, DeCamp AC, *et al.* Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature medicine* 2011; **17**: 366–71.

55 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.

56 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 pathogens* 2015; **11**: e1005154–e1005154.

57 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.

58 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.

59 Kiwelu IE, Novitsky V, Margolin L, *et al.* HIV-1 subtypes and recombinants in Northern Tanzania: distribution of viral quasispecies. *PLoS One* 2012; **7**.

60 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.

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

62 Gottlieb GS, Heath L, Nickle DC, *et al.* HIV-1 variation before seroconversion in men who have sex with men: analysis of acute/early HIV infection in the multicenter AIDS cohort study. *The Journal of infectious diseases* 2008; **197**: 1011–5.

63 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**.

64 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.

65 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. *The Journal of infectious diseases* 2014; **209**: 1032–8.

66 Sagar M, Lavreys L, Baeten JM, *et al.* Infection with multiple human immunodeficiency virus type 1 variants is associated with faster disease progression. *Journal of virology* 2003; **77**: 12921–6.

67 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.

68 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 pathogens* 2020; **16**: e1008179–e1008179.

69 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.

70 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. *Journal of virology* 2004; **78**: 11208–18.

71 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.

72 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. *Journal of virology* 2001; **75**: 2194–203.

73 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. *Clinical Infectious Diseases* 2019; published online Dec. DOI:10.1093/cid/ciz1081.

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

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

578 76 Iyer SS, Bibollet-Ruche F, Sherrill-Mix S, *et al.* Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. *Proceedings of the National Academy of Sciences of the United States of America* 2017; **114**:
580 E590–9.

77 Zhang H, Tully DC, Hoffmann FG, He J, Kankasa C, Wood C. Restricted genetic diversity of HIV-1 subtype C
582 envelope glycoprotein from perinatally infected Zambian infants. *PloS one* 2010; **5**: e9294–e9294.

78 Boeras DI, Hraber PT, Hurlston M, *et al.* Role of donor genital tract HIV-1 diversity in the transmission
584 bottleneck. *Proceedings of the National Academy of Sciences of the United States of America* 2011; **108**: E1156–63.

586 79 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.

588 80 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.

590 81 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.

592 82 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**.

594 83 Love TMT, Park SY, Giorgi EE, Mack WJ, Perelson AS, Lee HY. SPMM: estimating infection duration of multivariant HIV-1 infections. *Bioinformatics (Oxford, England)* 2016; **32**: 1308–15.

596 84 Liu Y, Jia L, Su B, *et al.* The genetic diversity of HIV-1 quasispecies within primary infected individuals. *AIDS research and human retroviruses* 2020.

598 85 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 (London, England)* 2008; **22**: 863–71.

600 86 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 Medicine* 2002; **3**: 1–11.

602 87 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**.

604 88 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.

606 89 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**.

608 90 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.

610 91 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
612 viruses. *Journal of virology* 2018; **92**: e00063-18.

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

93 Rossenkhan R, Novitsky V, Sebunya TK, Musonda R, Gashe BA, Essex M. Viral diversity and diversification of
616 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.

618 94 Learn GH, Muthui D, Brodie SJ, *et al.* Virus population homogenization following acute human immunodeficiency virus type 1 infection. *Journal of virology* 2002; **76**: 11953–9.

620 95 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 pathogens* 2012; **8**: e1002529–e1002529.

622 96 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*
624 *research and human retroviruses* 2002; **18**: 567–76.

97 Derdeyn CA, Decker JM, Bibollet-Ruche F, *et al.* Envelope-Constrained Neutralization-Sensitive HIV-1 After
626 Heterosexual Transmission. *Science* 2004; **303**: 2019 LP – 2022.

98 Kearney M, Maldarelli F, Shao W, *et al.* Human immunodeficiency virus type 1 population genetics and
628 adaptation in newly infected individuals. *Journal of virology* 2009; **83**: 2715–27.

99 Chaillon A, Gianella S, Wertheim JO, Richman DD, Mehta SR, Smith DM. HIV migration between blood and
630 cerebrospinal fluid or semen over time. *The Journal of infectious diseases* 2014; **209**: 1642–52.

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

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

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

638 103 Thompson RN, Wymant C, Spriggs RA, Raghwani J, Fraser C, Lythgoe KA. Link between the numbers of particles and variants founding new HIV-1 infections depends on the timing of transmission. *Virus Evolution*
640 2019; **5**. DOI:10.1093/ve/vey038.

104 Miller CJ, Shattock RJ. Target cells in vaginal HIV transmission. *Microbes and infection* 2003; **5**: 59–67.

642 105 Anderson D, Politch JA, Pudney J. HIV infection and immune defense of the penis. *Am J Reprod Immunol* 2011; **65**: 220–9.

644 106 Lee HY, Giorgi EE, Keele BF, *et al.* Modeling sequence evolution in acute HIV-1 infection. *Journal of Theoretical Biology* 2009; **261**: 341–60.

646 107 Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 1991; **129**: 555–62.

648 108 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.

650 109 Bourara K, Liegler TJ, Grant RM. Target cell APOBEC3C can induce limited G-to-A mutation in HIV-1. *PLoS Pathog* 2007; **3**: e153.

652 110 Shaw GM, Hunter E. HIV transmission. *Cold Spring Harbor perspectives in medicine* 2012; **2**: a006965.

s

654 # Appendices

Supplementary Methods **A2**

656 **Prisma Checklist** **A7**

Table of Selected Studies **A11**

658 **Temporal Structure of Exposure and Method** **A18**

Sensitivity Analyses of Pooling **A19**

660 **Influence of Methodology and Number of Genomes Analysed** **A20**

Leave-One-Out Cross Validation: Studies **A21**

662 **Leave-One-Out Cross Validation: Transmission Routes** **A22**

Comparison of Vaccine Escape and Placebo Participants **A23**

664 **Comparison of Sequencing Technologies** **A24**

Evaluating the Impact of Molecular Methods **A25**

666 **Evaluation of Publication Bias** **A26**

Binned Residuals Plot **A27**

668 **Sensitivity Analyses for Meta-regression** **A28**

Supplementary References **A29**

670

672

Supplementary Methods

Protocol Registration

This systematic review and meta-analysis was registered with PROSPERO following the initial literature search (PROSPERO study CRD42020202672).

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/)))

Databases Queried:

- 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

Data Extraction

- Route of exposure

We used the route of exposure of horizontally transmitted infections as reported by the original studies. These data are typically ascertained from risk behaviour questionnaires or enrolment criteria for a study cohort. In the majority of cases, a single route of exposure was reported. We stratified the route of exposure as much as possible, given the data available. This resulted in the following levels being included in the models: 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), and people who inject drugs (PWID)).

We refer to these as routes of exposure, rather than transmission route as an element of uncertainty is always present; both because there can be multiple concurrent routes of exposure and more generally because self-reported exposure does not necessarily match with transmission. For the same reason we chose not to stratify sexual exposure route into receptive vs. insertive anal sex for male-to-male and vaginal vs. anal sex for male-to-female.

We categorised mother-to-child (MTC) infections into pre-partum, intrapartum and post-partum was using the criteria below. Where these data were nor reported or the results ambiguous, we used the category 'unknown timing' (Bertolli et al., 1996):

Timing	HIV RNA+/PCR+
Pre-partum	Infant tests positive at birth
Intrapartum	Infant tests negative at birth, but later tests positive after no more than 3 months
Post-partum	Infant tests negative at birth, but later tests positive after more than 3 months

- Sampling Delay

For horizontally transmitted infections, the delay between infection and sampling (not diagnosis) was determined according to seropositivity. A delay of less than or equal to 21 days was recorded 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 vertical transmissions, 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.

- Methodologies

A binary classification (yes or no) was inferred 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. (Meyerhans et al. 1990, Simmonds et al. 1990). 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. (Simmonds et al. 1990, Salazar-Gonzalez et al. 2008).

Quantification method groupings were defined by the properties of each approach, resulting in six levels: phylogenetic, haplotype, distance, model, or molecular (Table S1). Prior the widespread application of sequencing, molecular methods such as heteroduplex mobility assays could provide a qualitative measure of diversity (Novitsky et al. 1996). Heterogeneous genomic segments would form heteroduplexes during gel electrophoresis of viral RNA, allowing one to distinguish genetically similar and dissimilar segments. Although estimates derived from these assays were regarded as close approximations of viral diversity, they only consider a tiny fraction of the whole genome and cannot provide further information regarding phylogenetics, or functional attributes of any substitutions. As a result they may lead to overestimation of the number of founder variants initiating infection.

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 (Slatkin and Hudson, 1991; Lee et al., 2009). Within the model category, we include any mathematical or statistical model which tests whether the observed patterns of diversity can be explained by the transmission of a single variant. For example, this includes Poissonfitter, where frequency distributions of Hamming Distances that significantly diverge from the expected Poisson distribution, after

controlling for APOBEC mediated hypermutation, represent an over-dispersed population (Giorgi et al., 2010); simple probabilistic models expressing the expected number of substitutions, and estimates of the time to most recent common ancestor that do not involve the reconstruction of genealogies.

Molecular	**Haplotype**	**Distance**	**Model**	**Phylogenetic**	
				Recipient Only	*Source & Recipient*
Heteroduplex Assay	Highlighter plot Haplotype Frequency	Pairwise distance Diversity	Poissonfitter ● Goodness of fit ● Starlike topology ● tMRCA	Starlike topology tMRCA (genealogy) Diversification	Paired topologies tMRCA (genealogy)

Table S1: 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.

Haplotype methods identify linkage patterns of individual polymorphisms across samples from a patient. In the study of HIV founder infection multiplicity, this category mostly concerns the use of highlighter plots, that visually map nucleotide mismatches along an aligned gene segment (Keele et al. 2008). Inspection of these graphs facilitates an approximate enumeration of the number of variants initiating an infection and allow for inference of putative recombinants and APOBEC mediated hypermutation, which would erroneously inflate diversity measures. Haplotype methods may also refer to modelling the distribution of haplotypes obtained through longitudinal deep-sequence samples.

Phylogenetic methods are here defined as approaches that explicitly reconstruct ancestral genealogical relationships directly from sequence data. These 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.

Statistical Models

- Pooled estimates models

We assumed a binary outcome y_{ik} of whether the infection of individual k of study i was initiated by multiple founder variants (1) or not (0) with probability p_{ik}. For the two-step model, we first fit a logit model to these binary outcome data for each study, i, where θ_i is the effect size of study i, x_{ik} is whether the infection of individual k of study i was initiated by multiple founder variants (1) or not (0) and a_i is the intercept :

$$y_{ik} \sim Bernoulli(p_{ik})$$

$$logit(p_{ik}) = a_i + \theta_i x_{ik}$$

We then accounted for between study variation in the effect sizes by assuming a random effects model such that each of the estimated study effect sizes, $\hat{\theta}_i$, is a sample from a different normal distribution with a mean equal to an underlying study-specific effect size. This study-specific effect size is itself drawn from a normal distribution with constant mean and variance, τ^2 (the between-study variance) :

$$\hat{\theta}_i \sim N\left(\theta_i, var(\hat{\theta}_i)\right)$$

$$\theta_i \sim N(\theta, \tau^2)$$

For the one-step model, individual-level and study-level variation are considered simultaneously:

$$y_{ik} \sim Bernoulli(p_{ik})$$

$$logit(p_{ik}) = \alpha_0 + \theta_i x_{ik}$$

$$\theta_i = \theta + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \tau^2)$$

Here, α_0 represents a fixed intercept and θ_i is the random effect of study acting on observation x_{ik}. θ_i is the sum of θ, the mean study effect size, and ε_i, the study-specific effect drawn from a normal distribution with variance, τ^2 (the between-study variance). We compared the results from our one-step model with a two-step model to confirm our estimates were consistent.

- Univariable and multivariable models

We extended our one-step model by conducting a univariable meta-regression with each covariate contributing a fixed effect and assuming normally distributed random effects of publication. We report results for univariable models that analysed the role of route of exposure, quantification method, genome region analysed and sampling delay as fixed effects (β_{ik}).

$$logit(p_{ik}) = \alpha_0 + \beta_{ik} x_{ik} + \theta_i x_{ik} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \tau^2)$$

A multivariable model was built from the fixed effects used in the univariable analysis. The fixed effects ($\beta_{n_{ik}}$, $n \in [1, m]$) were selected according to a 'keep it maximal' principle, in which covariates were only removed to facilitate a non-singular fit. The selected model is outlined here:

$$logit(p_{ik}) = \alpha_0 + \sum_{n=1}^{m} \beta_{n_{ik}} x_{n_{ik}} + \theta_i x_{ik} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \tau^2)$$

The selected model was assessed for convergence, singularity, multicollinearity using the R package ggeffects. We calculated the proportion of binned residuals within 95% confidence limits. Model estimated probabilities per transmission route were calculated controlling baseline covariates as our 'gold standard' methodology (envelope genomic region, a short delay, haplotype analysis).

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. Further details on how to run the analysis are included in the README.md.

The analyses were conducted in R 4.1.2, principally using the following packages:
lme4, 1.1-27.1, (Bates et al. 2007)
metafor, 3.0-2, (Viechtbauer 2010)
tidyverse, 1.3.1, (including ggplot2 3.3.5, stringr 1.4.0, forcats 0.5.1 & dplyr 1.0.7) (Wickham et al., 2019)
reshape2 1.4.4 (Wickham, 2012)
ggeffects 1.1.1 (Lüdecke, 2018)
mltools 0.5.2
parallel 3.6.2

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, A2

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.	4
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.	4
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.	A2
Study selection processes	9	State the process for determining which studies were eligible for inclusion.	4
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).	4
		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.	5, A2-A4
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.	5
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.	5
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.	5

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, A4-A6
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.	6
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.	6
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.	9-11
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.	A18-A27
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.	10, A11-A17

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.	12-14
		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.	A18-A27
Discussion			
Summary of evidence	24	Summarise the main findings, including the strength of evidence for each main outcome.	15
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.	16
Conclusions	26	Provide a general interpretation of the findings in the context of other evidence.	15-16
Implications	A4	Consider relevance to key groups (such as policy makers, service providers and service users). Consider implications for future research.	16
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, 7

Table S2: PRISMA checklist referencing the necessary steps taken to pages in this manuscript.

846 **Table of Selected Studies**

	Transmission Routes	Method	Genomic Region	Range of Genomes Analysed Per Participant	Virus Subtype	Number of Participants	P(multiple founders)	Data Included	
								Participants	Multiple Founders
Wolinsky et al. (1992)[79]	MTC:undisclosed	Haplotype	Env; V3 & V4-V5	9-18	Unknown	3	0	3	0
Briant et al. (1995)[33]	MTC:undisclosed	Phylogenetic: source and recipient	Env; V3	··	B	4	0·75	4	3
Poss et al. (1995)[47]	HSX:MTF	Haplotype	Env; gp120	10-17	A, D	6	0·83	6	5
Wade et al. (1998)[35]	MTC:undisclosed MTC:PreP	Phylogenetic: source and recipient	Gag; p17	49-115	B	2	0·5	2	1
Long et al. (2000)[52]	HSX:MTF, HSX:FTM	Molecular	Env; gp120	··	A, D, C, Unknown	36	0·55	36	15
Dickover et al. (2001)[72]	MTC:IntraP MTC:PreP	Molecular	Env; gp120	··	B	23	0·26	23	6
Delwart et al. (2002)[63]	HSX:FTM HSX:MTF Unknown	Molecular	Env; V3	2-141	B	17	0·06	17	1
Learn et al. (2002)[94]	MSM	Molecular	Env; gp120	··	B	8	0·5	8	4
Long et al. (2002)[96]	HSX:MTF	Distance	Env; gp120	··	A, Unknown	5	0·5	2	0
Nowak et al. (2002)[86]	MTC:undisclosed	Phylogenetic: source and recipient	Env; V3	44-71	B	3	0·34	3	1
Renjifo et al. (2003)[64]	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)[48]	MTC:IntraP MTC:PreP	Phylogenetic: source and recipient	Env; gp120	11-36	A	13	0·54	13	7
Derdeyn et al. (2004)[97]	HSX:MTF HSX:FTM	Phylogenetic: source and recipient	Env; gp120	13-20	C, G	7	0	7	0
Ritola et al. (2004)[70]	HSX:MTF HSX:FTM MSM	Molecular	Env; V1-V3	··	B	26	0·52	25	7
Sagar et al. (2004)[29]	HSX:MTF PWID MSM HSX:FTM	Molecular	Env; V1-V5	··	A, B, Unknown	17	0·24	17	4
Sagar et al. (2006)[60]	HSX:MTF	Distance	Env; V1-V3	10-25	A, D, Unknown, Recombinants	12	0·5	··	··
Gottlieb et al. (2008)[62]	MSM	Haplotype	Env; V1-V5	11-19	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	10-67	B	102	0·24	44	15
Kwiek et al. (2008)[85]	MTC:IntraP MTC:PreP	Molecular	Env; V1-V2	··	C	48	0·42	48	28
Salazar-Gonzalez et al. (2008)[22]	HSX:MTF HSX:FTM	Distance Haplotype Phylogenetic: recipient only	Env; gp160	24-48	C, Unknown	12	0·34	12	4
Abrahams et al. (2009)[30]	HSX:FTM HSX:MTF	Distance Model Haplotype Phylogenetic: recipient only	Env; gp160	15-42	C, G	69	0·22	69	15
Haaland et al. (2009)[14]	HSX:MTF	Haplotype	Env; gp160	22-73	A, C,	27	0·23	22	3

	HSX:FTM	Phylogenetic: source and recipient			Unknown				
Kearney et al. (2009)[98]	MSM HSX:FTM HSX:MTF PWID	Phylogenetic: recipient only	pol	··	B	14	0·14	11	0
Novitsky et al. (2009)[51]	HSX:MTF HSX:FTM	Phylogenetic: recipient only	Env; gp120	11-33	C	8	0·25	8	2
Salazar-Gonzalez et al. (2009)[53]	MSM HSX:FTM	Distance Haplotype Model	NFLG	4-26	B, C	12	0·083	2	0
Bar et al. (2010)[16]	PWID	Phylogenetic: recipient only	Env; gp160	19-163	B	10	0·6	10	6
Fischer et al. (2010)[90]	MSM	Model	Env; gp120	··	B	3	0	··	··
Li et al. (2010)[15]	MSM	Distance Haplotype	Env; gp160	23-89	B	28	0·36	28	10
Masharsky et al. (2010)[17]	PWID	Haplotype	env	18-29	A, Recombinants	13	0·31	13	4
Zhang et al. (2010)[77]	MTC:IntraP	Phylogenetic: source and recipient	Env; V1-V5	25-30	C, Recombinants	6	0	6	0
Boeras et al. (2011)[78]	HSX:FTM HSX:MTF	Phylogenetic: source and recipient	Env; V1-V4	31-73	A, C	8	0	··	··
Collins-Fairclough et al. (2011)[82]	MSM HSX:FTM HSX:MTF HSX:undisclosed	Haplotype	Env; V1-C4	5-20	B	27	0·23	14	2
Herbeck et al. (2011)[45]	MSM	Distance	NFLG	10-113	B	9	0·11	9	1
Kishko et al. (2011)[55]	MTC:IntraP	Phylogenetic: source and recipient	Env; gp160	10-22	B	5	0·4	5	2
Nofemela et al. (2011)[44]	HSX:MTF	Haplotype	env	5-18	A, B, C, D,	22	0·27	22	6

					Recombinants				
Novitsky et al. (2011)[89]	HSX:MTF HSX:FTM	Distance Haplotype Model Phylogenetic: recipient only	gag & Env; gp120	6-33	C	25	0·32	16	6
Rachinger et al. (2011)[87]	MSM	Phylogenetic: source and recipient	NFLG	··	B	1	0	··	··
Rieder et al. (2011)[37]	Unknown MSM HSX:MTF	Distance	Env; C2-V3-C3	14-16	A, B, C, G, CRF01AE, CRF02AG, CRF12BF, CRF14BG	143	0·11	··	··
Rolland et al. (2011)[54]	MSM HSX:MTF	Phylogenetic: recipient only	NFLG	2-14	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)[95]	unknown	Distance	NFLG	··	B	1	0	··	··
Kiwelu et al. (2012)[59]	HSX:MTF	Phylogenetic: recipient only	Env; gp120	5-62	A, C, D	50	0·27	43	10
Rossenkhan et al. (2012)[93]	HSX:MTF HSX:FTM	Phylogenetic: recipient only	gag & Env; gp120	3-92	C	20	0·15	5	0
Sturdevant et al. (2012)[36]	MTC:undisclosed	Haplotype Phylogenetic: recipient only	Env; gp160	16-46	C	43	0·12	43	5
Baalwa et al. (2013)[69]	HSX:MTF HSX:FTM	Haplotype	NFLG	20-82	A, D, Recombinants	12	0·17	12	2
Frange et al. (2013)[80]	MSM HSX:MTF HSX:FTM	Phylogenetic: source and recipient	Env; C2-V5	19-43	B	8	0	8	0
Chaillon et al. (2014)[99]	MTC:PreP	Phylogenetic: source and	Env; V1-V5	6-32	CRF01_AE	9	0·12	8	1

	MTC:IntraP	recipient							
Sterrett et al. (2014)[67]	PWID	Distance Haplotype Model Phylogenetic: recipient only	Env; gp160	12-41	B, CRF01AE, CRF1501B, Recombinants	50	0·42	49	14
Wagner et al. (2014)[65]	MSM PWID	Phylogenetic: recipient only	NFLG	··	B	108	0·06	108	7
Chen et al. (2015)[41]	MSM	Haplotype	Env; gp160	6-26	B, CRF01AE, CRF07BC	30	0·2	18	3
Danaviah et al.(2015)[50]	MTC:PostP	Phylogenetic: source and recipient	Env; C2-V5	··	C	11	0·18	11	2
Deymier et al. (2015)[56]	HSX:FTM	Phylogenetic: recipient only	NFLG	6-9	C	6	0	5	0
Gounder et al. (2015)[57]	HSX:FTM HSX:MTF	Phylogenetic: recipient only	gag	12-16	C	22	0·27	22	6
Janes et al. (2015)[6]	MSM HSX:FTM HSX:MTF	Distance	Env; gp120	2-28	B, CRF01AE	163	0·29	100	32
Le et al. (2015)[102]	PWID	Phylogenetic: source and recipient	Env; gp120	··	B	2	0	2	0
Zanini et al. (2015)[100]	HSX:MTF MSM HSX:FTM	Distance	NFLG	··	B, C, CRF01AE	9	0·22	9	2
Chaillon et al. (2016)[39]	MSM PWID	Distance Phylogenetic: source and recipient	Env; C2-V3	··	B	30	53·3	30	16
Love et al. (2016)[83]	PWID MSM Unknown HSX:FTM HSX:MTF	Model	Env; gp160	10-163	B, C	182	0·23	··	··

	HSX:undisclosed								
Novitsky et al. (2016)[42]	HSX:MTF HSX:FTM	Distance	Env; V1-C5	12-54	C	42	0·21	15	3
Oberle et al. (2016)[88]	MSM HSX:MTF	Phylogenetic: source and recipient	Env; gp160	8-27	B	9	0	2	0
Park et al. (2016)[49]	MSM	Model	Env; gp160	3-13	B, CRF02AG	59	0·17	··	··
Salazar-Gonzalez et al. (2016)[92]	unknown	Haplotype	Env; gp160	12-19	B	2	0	··	··
Smith et al. (2016)[46]	HSX:FTM HSX:MTF	Haplotype	Env; gp120	5-104	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)[81]	MSM	Phylogenetic: recipient only	Env; gp120	4-30	B	46	0·28	43	12
Iyer et al. (2017)[76]	MSM HSX:FTM HSX:MTF	Haplotype	NFLG	7-24	B, C	8	0·13	7	1
Kijak et al. (2017)[75]	HSX:MTF HSX:FTM	Haplotype	NFLG	··	CRF01_AE, Recombinants	6	0·83	··	··
Ashokkumar et al. (2018)[91]	MTC:undisclosed	Haplotype	Env; gp120	4-22	C	8	0·25	8	2
Dukhovlinova et al. (2018)[38]	PWID	Model	Env; gp160	8-46	A	7	0	7	0
Leitner & Romero-Severson (2018)[74]	MSM HSX:MTF HSX:FTM PWID HSX:undisclosed MTC:undisclosed Unknown NOSO	Phylogenetic: source and recipient	Various	··	A, B, C, D, CRF01_AE, CRF14_BG	508	0.52	··	··

Lewitus & Rolland (2019)[32]	Unknown MSM HSX:FTM HSX:MTF	Phylogenetic: recipient only	Env; gp160	11-47	B	72	0·29	··	··
Sivay et al. (2019)[73]	PWID	Model	Env; gp41	··	A, CRF01AE	7	0·43	7	3
Todesco et al. (2019)[34]	MSM	Phylogenetic: source and recipient	pol	··	B, CRF02AG, CRF07BC	8	0·25	7	2
Tovanabutra et al. (2019)[43]	MSM HSX:MTF	Haplotype	Env; gp160	5-70	CRF01_AE, recombinant	18	0·44	18	7
Brooks et al. (2020)[61]	HSX:FTM HSX:MTF	Phylogenetic: recipient only	NFLG	5-22	C	13	0·08	12	1
Leda et al. (2020)[58]	HSX:MTF MSM HSX:FTM	Model	Env; gp160	··	B, F, Recombinant	25	0·08	21	2
Liu et al. (2020)[84]	MSM	Haplotype	Env; gp120	4-31	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)[101]	MTC:IntraP MTC:PreP	Model	Env; gp160	20-47	B, C	4	0·25	4	1
Rolland et al. (2020)[68]	HSX:MTF MSM	Phylogenetic: recipient only	Env; gp160	2-42	A, B, C, CRF01AE	39	0·28	39	10
Villabona-Arenas et al. (2020)[71]	MSM HSX:undisclosed HSX:MTF HSX:FTM	Phylogenetic: source and recipient	Env; gp41, gp160, gp120 & NFLG	5-149	A, B, C, D, G, Recombinants	112	0·23	49	12

Table S3: 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

Temporal Structure of Exposure and Method

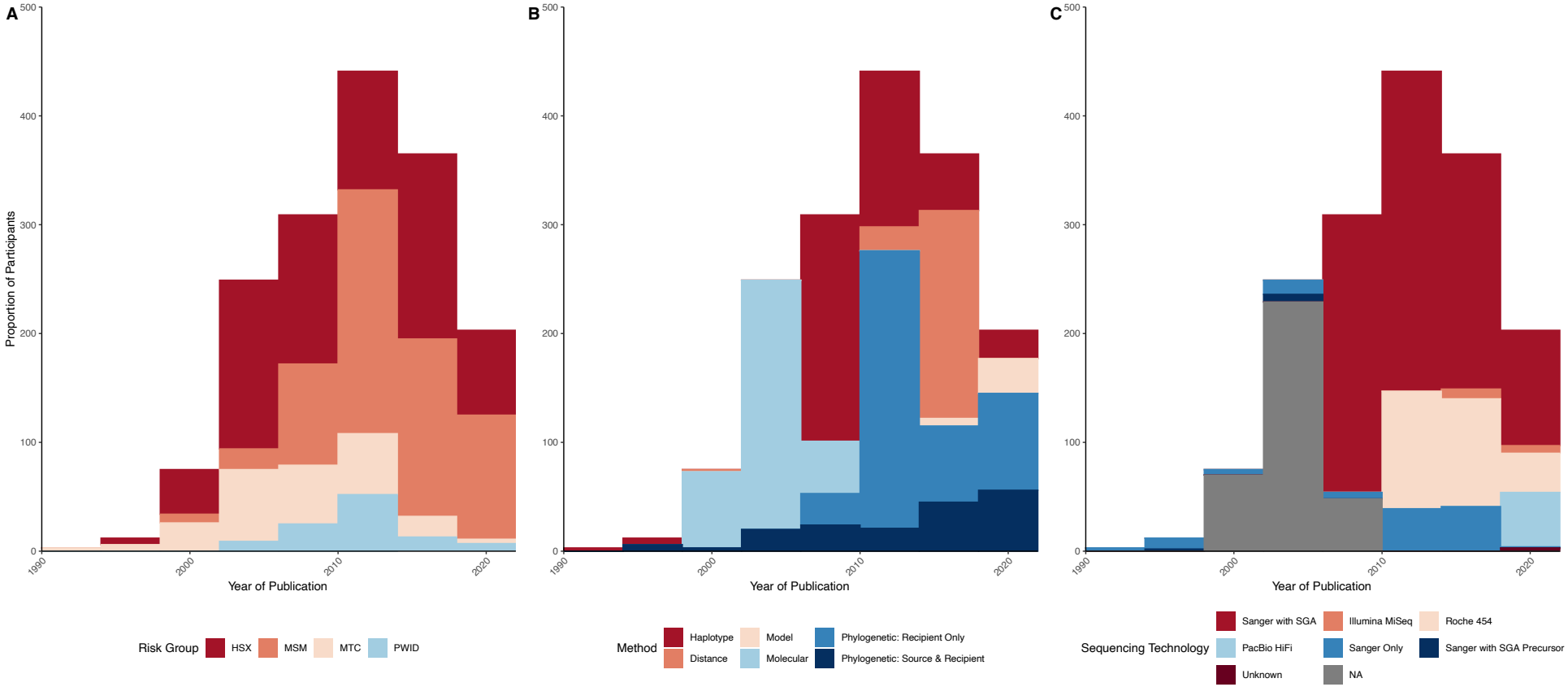


Figure S1: Distributions of transmission route (A), grouped method (B) and sequencing technology (C) over time, highlighting the epidemiologic and methodological step-changes that occurred over the three decades in which the selected studies were published. This means that earlier methods may be biased to those transmission routes that were more common in earlier studies.

Sensitivity Analyses of Pooling

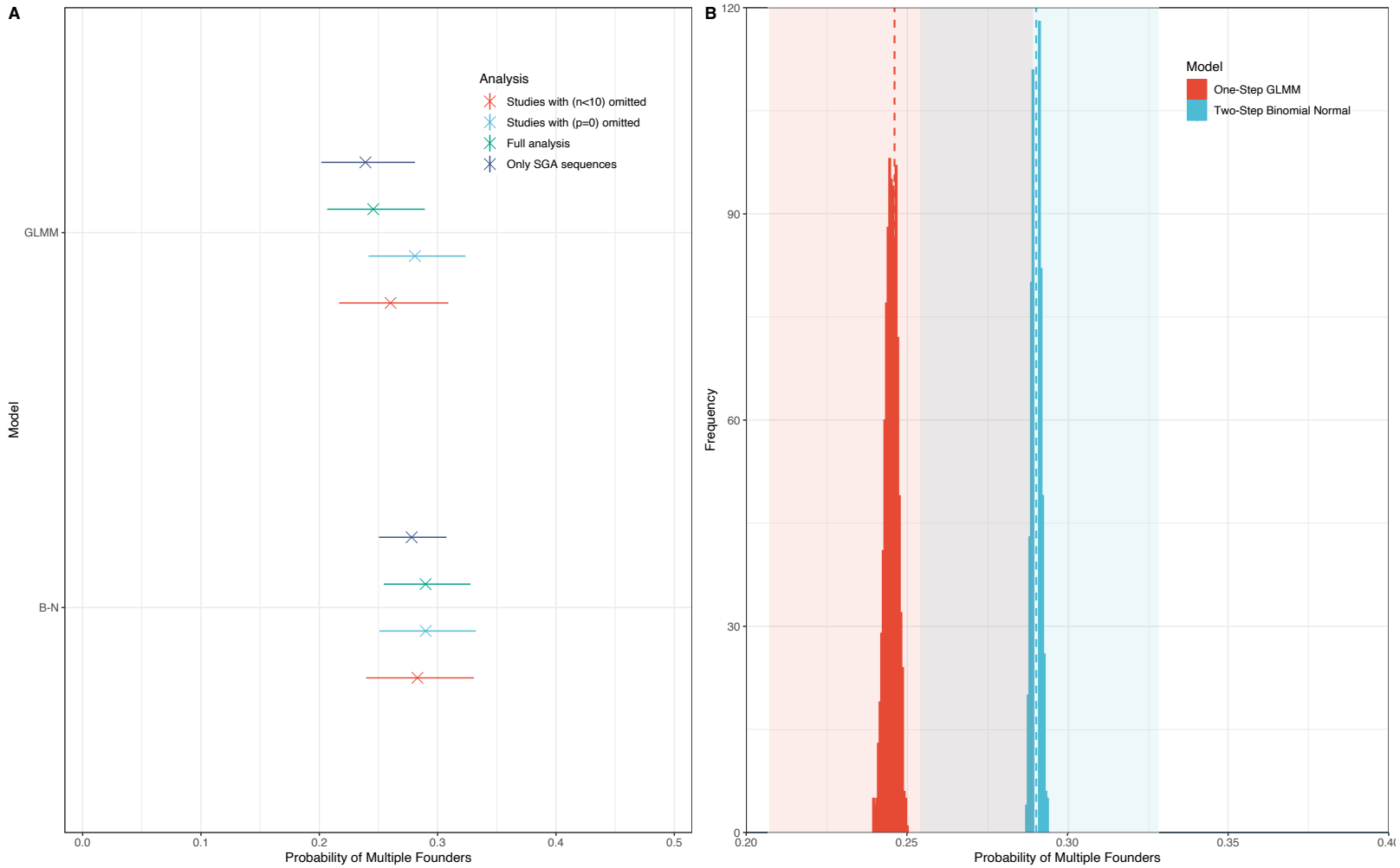


Figure S2: A comparison of the pooled estimates of the probability that an infection is initiated by multiple founders by the one-step (GLMM) and two-step (Binomial-Normal (B-N)) models and respective sensitivity analyses. Plot (A) shows both models calculate concordant estimates and are robust to sensitivity 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.

Influence of Methodology and Number of Genomes Analysed

We analysed data from participants spanning the interquartile range (11 - 28 genomes), and then restricted the analysis to participants with higher than the upper quartile value (numbers of genomes>28) or lower than the lower quartile value (number of genomes <11). Restricting the analysis to participants for whom a large (>28) or small (<11) number of sequences were analysed adjusted the pooled estimate to 0·26 (0·20-0·34) and 0·21 (0·17-0·25), respectively (Fig.S3). The model fitted to participants spanning the interquartile range also revealed a slight increase in the probability of observing multiple founder variants when compared to the original estimates (0·27 (0·24-0·31)). These findings suggest the presence of a subtle correlation between the number of genomes analysed and the probability of observing multiple founder variants.

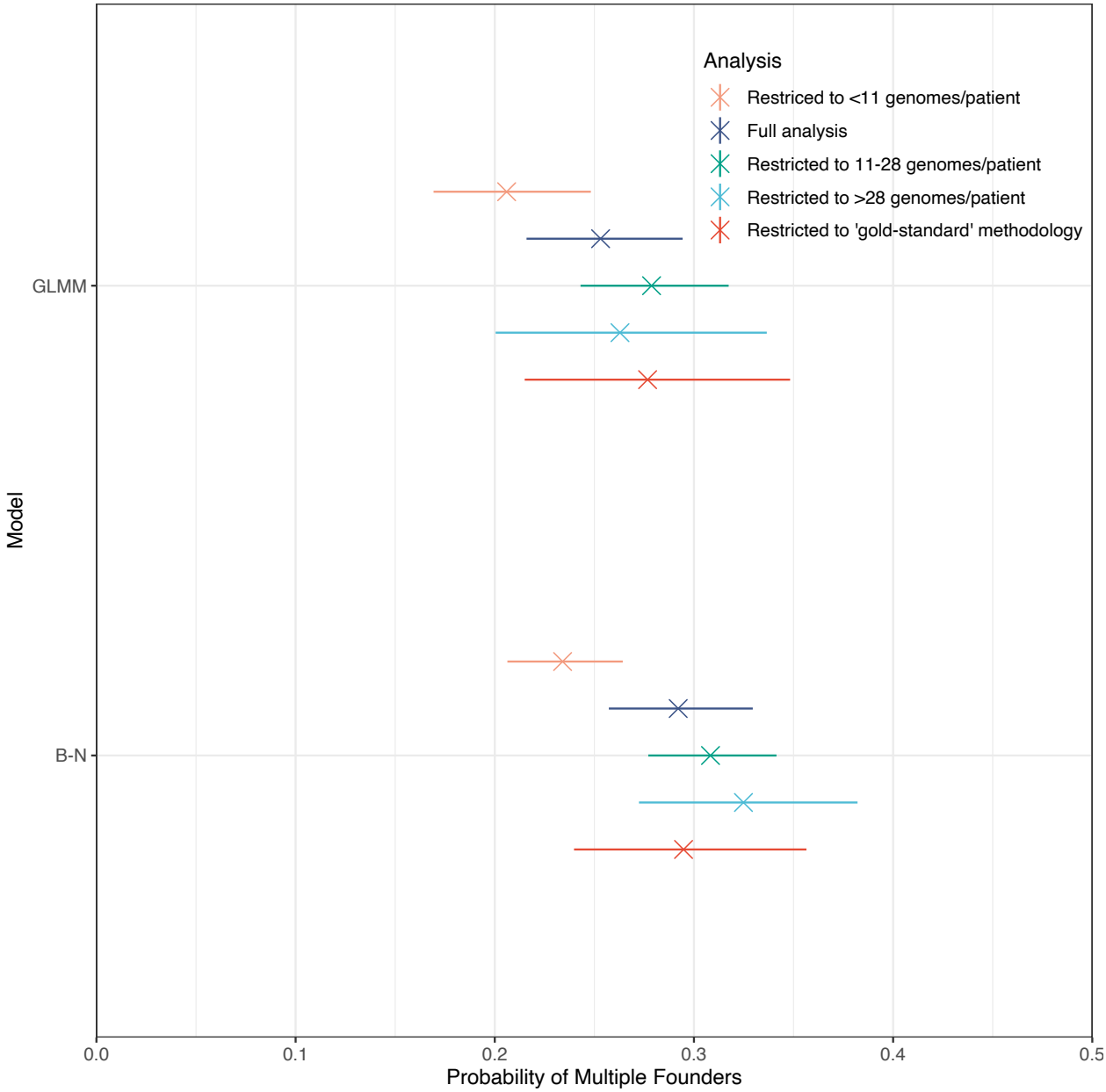


Figure S3: Comparing the pooled estimates of the probability that an infection is initiated by multiple founders by the one-step (GLMM) and two-step (Binomial-Normal (B-N)) models under our 'gold-standard' methodology, and when varying the threshold of the number of genomes analysed per patient.

Leave-One-Out Cross Validation: Studies

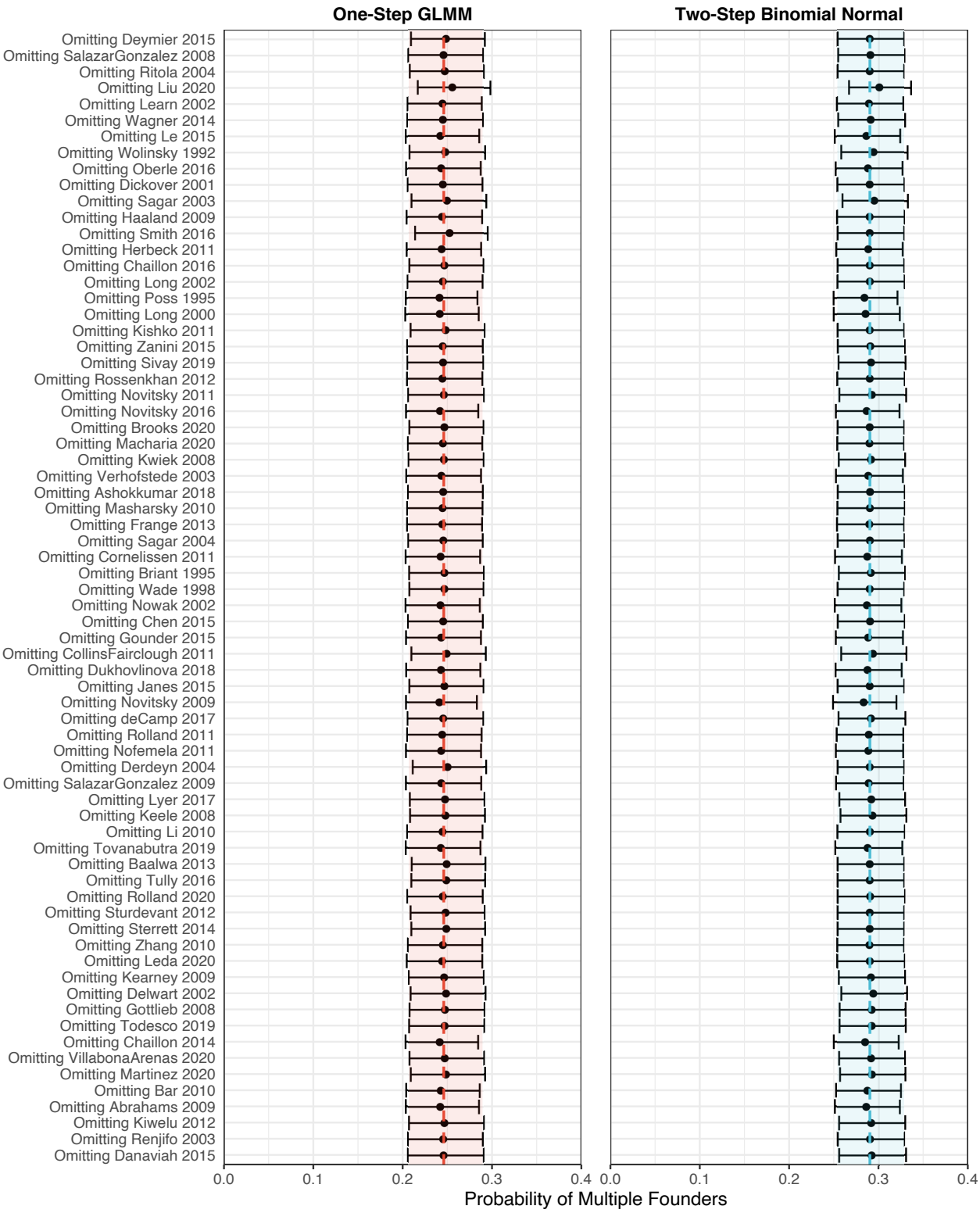


Figure S4: For both one-step and two-step models, we visually inspect the influence of each study included in our analysis on the pooled estimate that an infection is initiated by multiple founders. We find that in iteratively excluding individual studies, no discernible impact on the overall pooled estimate is made. The dashed lines and shaded areas denote the original point estimate and confidence intervals, respectively.

Leave-One-Out Cross Validation: Transmission Routes

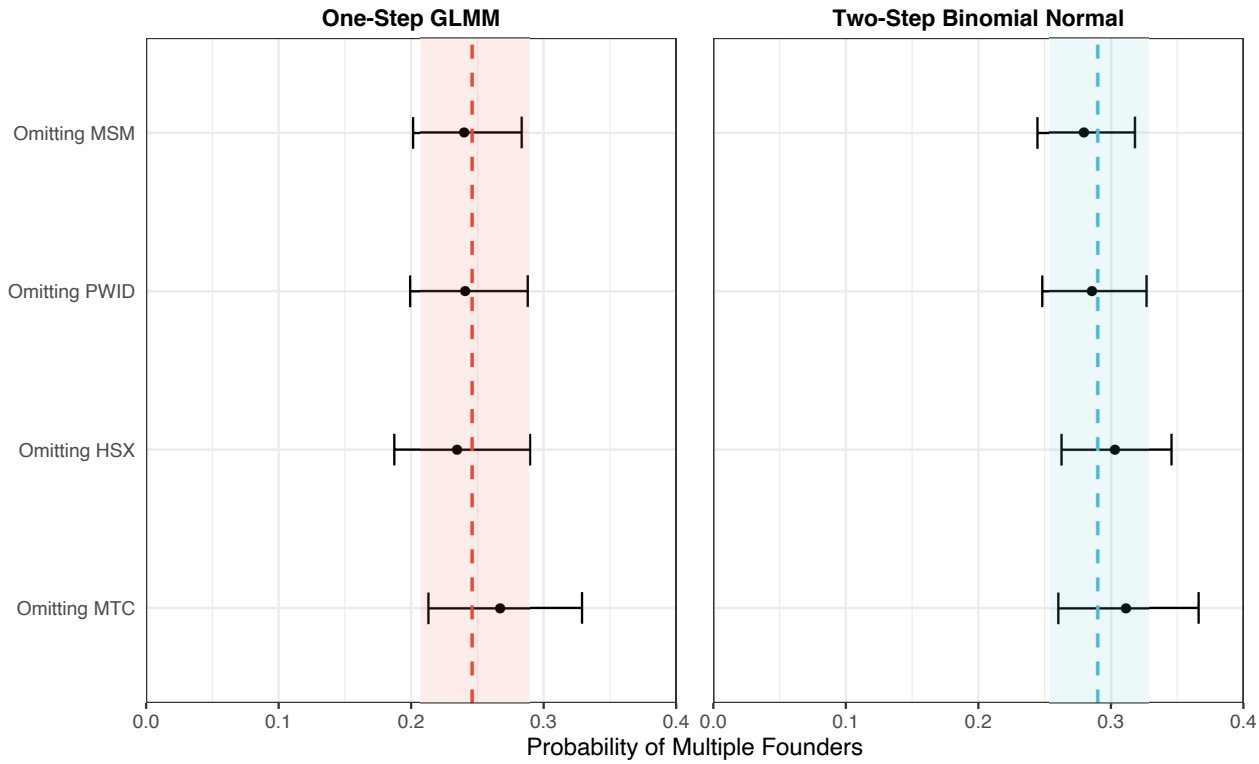


Figure S5: For both one-step and two-step models, we visually inspect the influence of each risk group included in our analysis on the pooled estimate that an infection is initiated by multiple founders. We find no discernible impact on the overall pooled estimate is made. The dashed lines and shaded areas denote the original point estimate and confidence intervals, respectively.

Comparison of Vaccine Escape and Placebo Participants

Some of the selected studies included participants enrolled on vaccine trials. As breakthrough infections of vaccine-recipients may not reflect natural infection, we compare vaccine and placebo arms of trials for which these data were available. This analysis included participants from HTVN502 and RV144 (A third vaccine trial (HTVN505) is not included as participant vaccine status was not available). Estimates of founder multiplicity were extracted from Rolland et al 2011 (HTVN502), and Janes et al 2015 (RV144), following our inclusion criteria of selecting the first instance for which data are available (HTVN502 participants were also subsequently analysed by Janes et al.). We did not find any significant difference between vaccine-breakthrough and placebo infections.

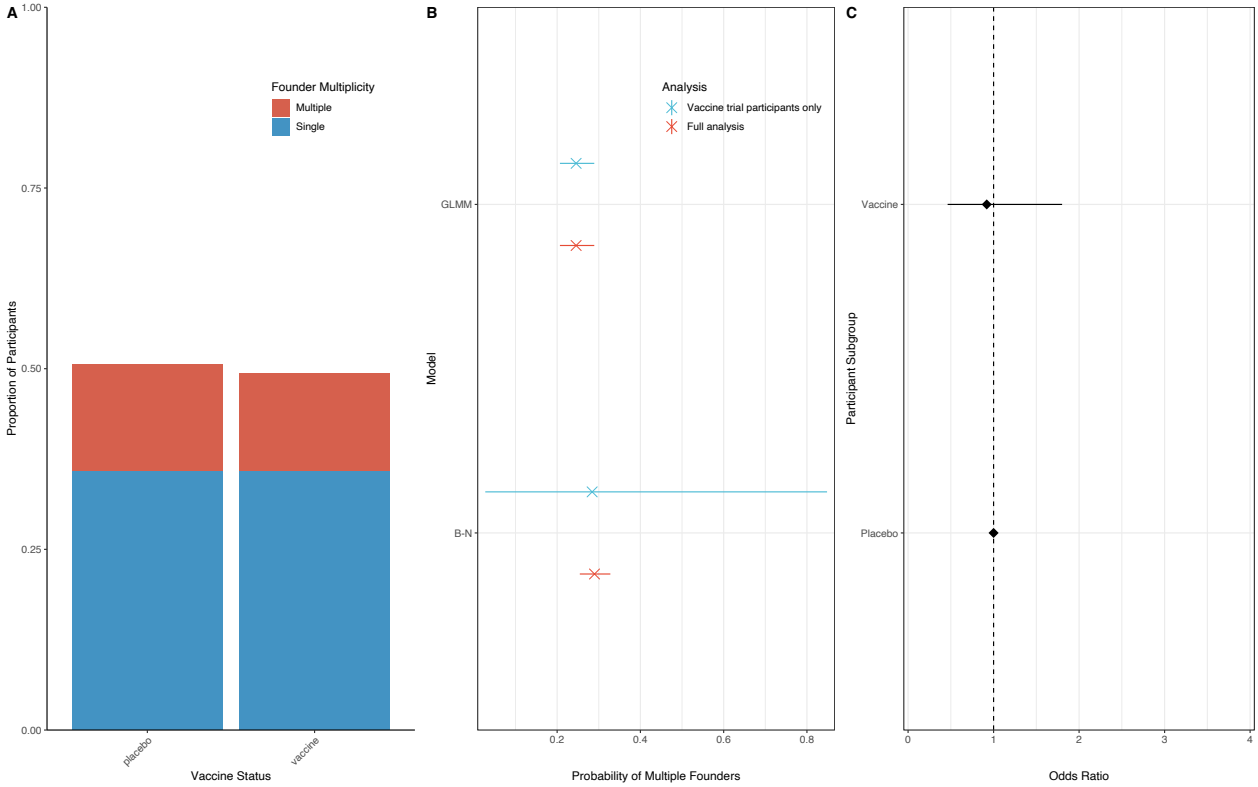


Figure S6: A) The proportion of infections identified as being initiated by multiple founders, segregated by vaccine status. B) Pooling estimates calculated using one and two-step models for vaccine trial only datapoints, compared to the base case dataset. C) Univariable analysis finding no significant difference in the odds of observing multiple founder variants between vaccinated and placebo arms of vaccine trial participants.

Comparison of Sequencing Technologies

Of 70 selected studies in our base case dataset, eleven studies used deep sequencing (Roche 454 – 9 studies, Ilumina – 1 study, PacBio HiFi- 1 study). To investigate whether the higher resolution of deep sequencing approaches influenced the observation of multiple founder variants initiating HIV infection, we conducted a univariable regression across those studies that used sequence-based methods. (0.22 (95% CI: 0.19-0.27)) was slightly lower than our original pooled estimate (0·25 (95% CI: 0·21-0·29)). In our univariable analysis, we did not find the odds of observing multiple founder variants differed significantly across sequencing methodologies.

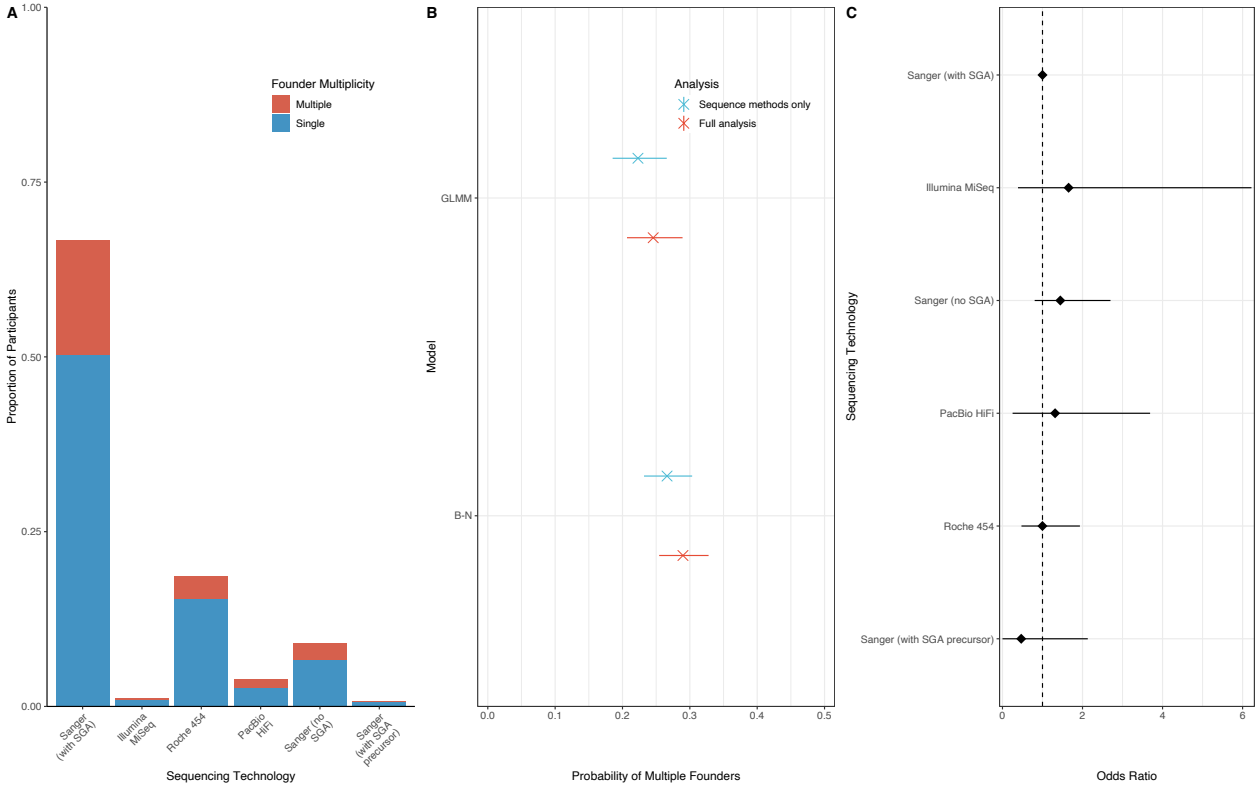


Figure S7: A) The proportion of infections identified as being initiated by multiple founders, segregated by sequencing technology. B) Pooling estimates calculated using one and two-step models for sequence methods only datapoints, compared to the base case dataset. C) Univariable analysis finding no significant difference in the odds of observing multiple founder variants across sequencing technologies.

Evaluating the Impact of Molecular Methods

Both our multivariable and univariable analyses identified differences in the probability of multiple founders according to the genomic region analysed. Molecular methods, which we defined as approaches that rely on the formation of heteroduplexes during gel electrophoresis of viral RNA, are very sensitive. This allows one to distinguish genetically similar and dissimilar segments, however the use of these methods on short fragments of envelope may produce false positive results. Indeed, in both our univariable and multivariable analyses, there were significantly greater odds of recording multiple founder infections if molecular methods were used. To evaluate the impact of molecular methods as a confounder on the genomic region, we recalculated our pooled estimate under different scenarios and re-fitted a univariable model of genomic region in the absence of molecular methods. Of 1657 individuals, 1315 in our base case dataset were analysed using the envelope genomic region. Pooled estimates for envelope only individuals and envelope only individuals without molecular methods under the GLMM were 0·28 (0·23-0·32) and (0·25 0·21-0·29) respectively. A univariable analysis of genomic region fitted to the main dataset excluding molecular methods reported findings consistent with the main univariable analysis.

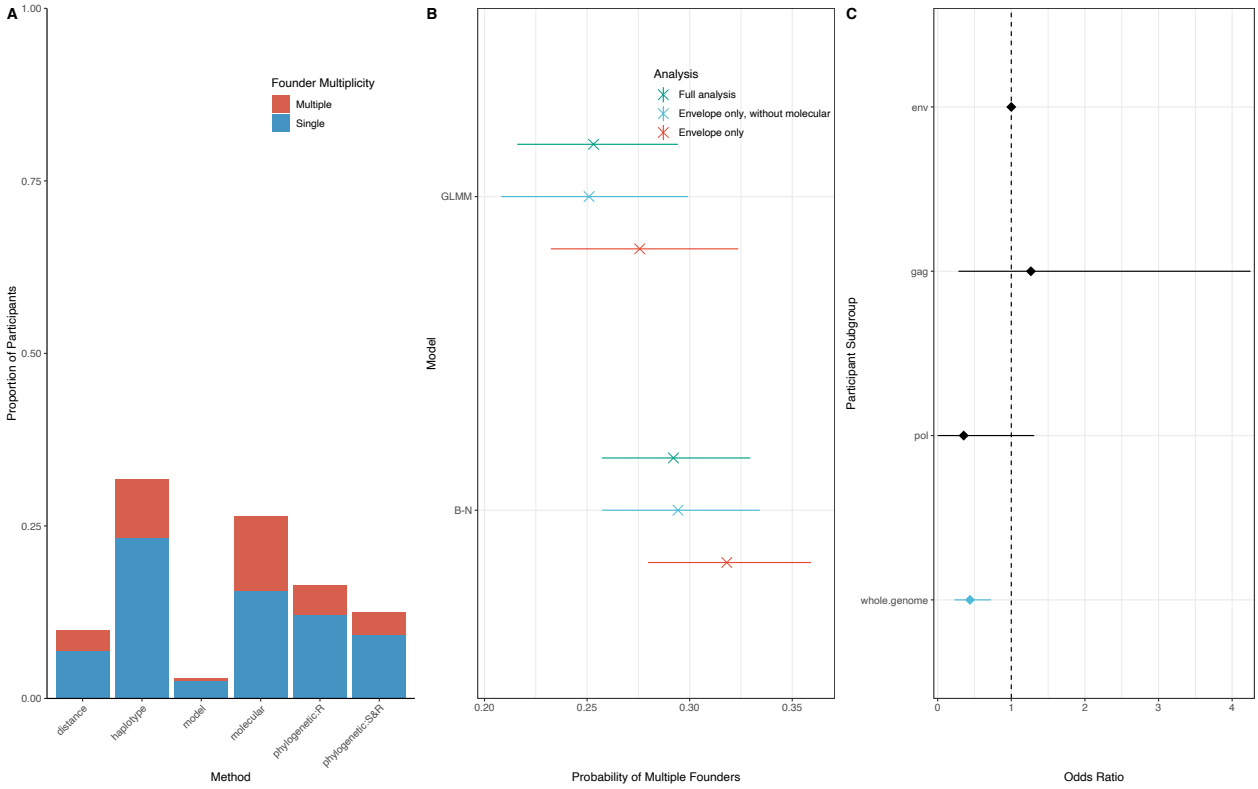


Figure S8: A) The founder variant multiplicity of 1315 individuals was analysed using the envelope genomic region, here segregated by method and indicating the prevalence of multiple founder infections. B) Pooled estimates calculated using one and two-step models from individuals for whom the envelope genomic region was analysed including/excluding molecular methods. C) Univariable analysis on dataset excluding molecular methods, reporting findings consistent with the main univariable analysis.

Evaluation of Publication Bias

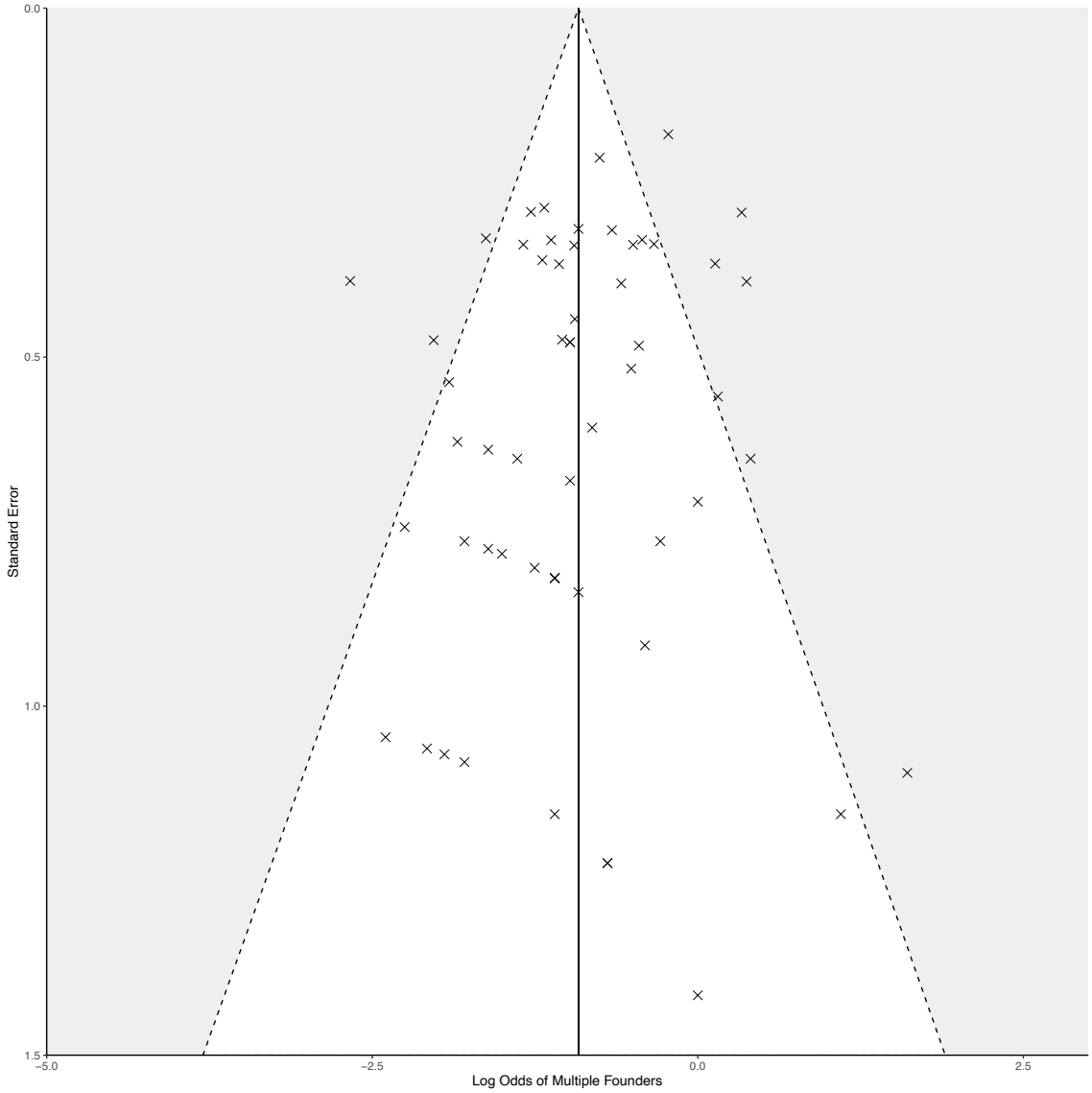


Figure S9: Funnel plot to visually evaluate the presence of publication bias. In the absence of publication bias, study estimates are distributed symmetrically with respect to the pooled estimate (vertical solid black line). Here, the log odds of an infection being initiated by multiple founders for 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 = -0·7495, df = 55, p = 0·4568.

Binned Residuals Plot

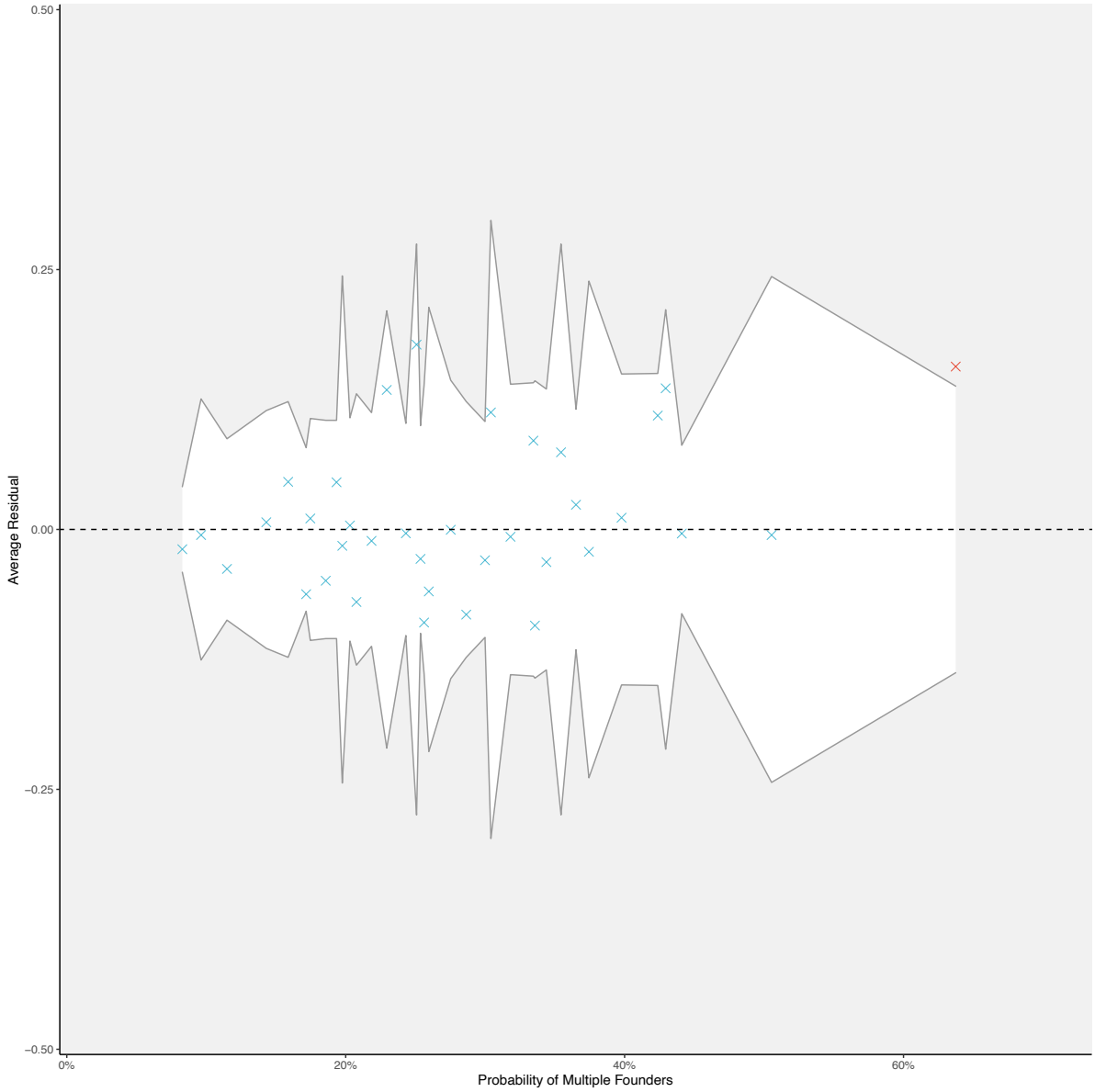


Figure S10: Binned residuals from the select multivariable model. 97% of the average residuals across each bin fall within the 95% confidence intervals (white area), indicating a good model fit.

Sensitivity Analyses for Meta-regression

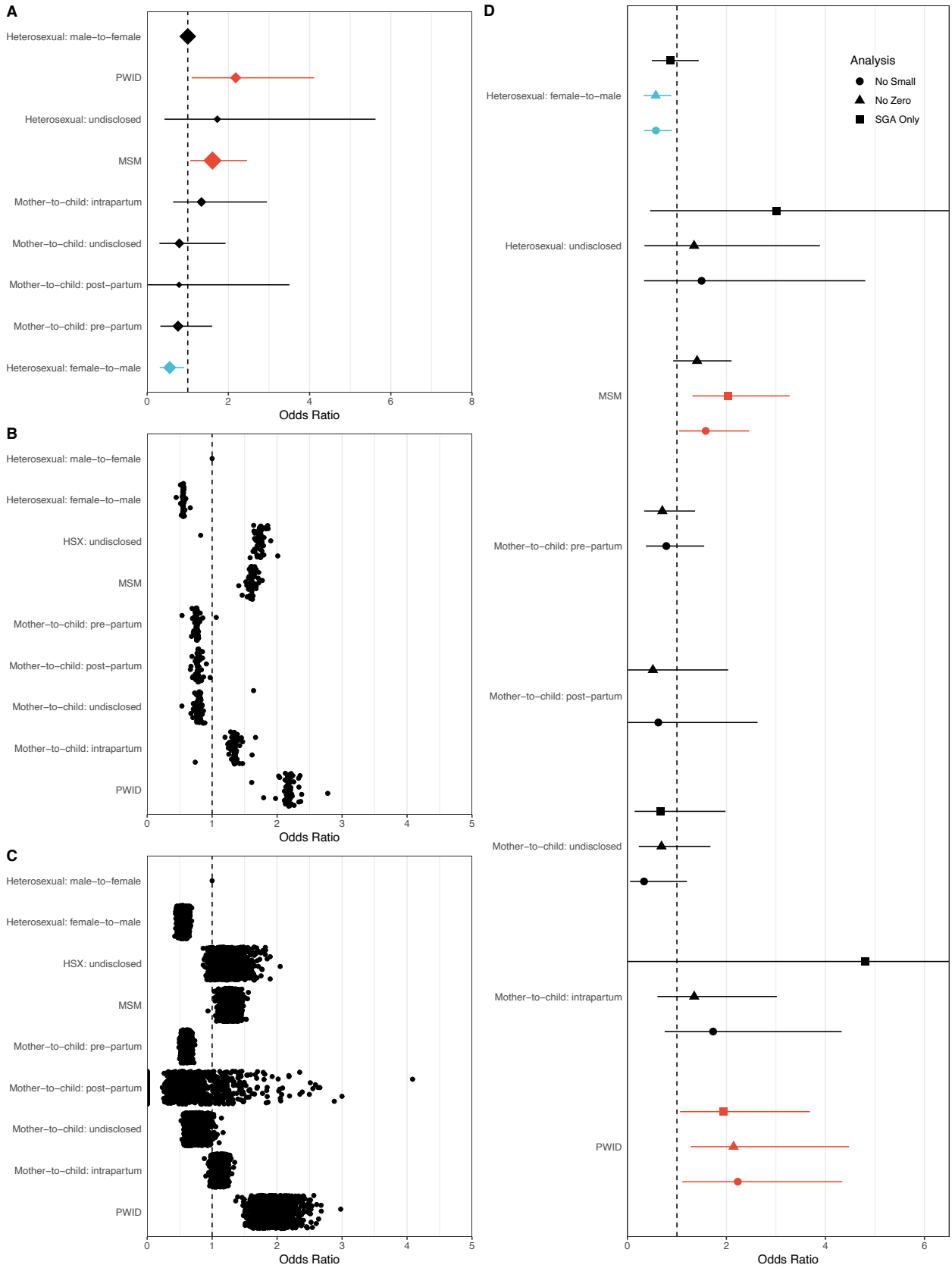


Figure S11: Odds ratios that an infection is initiated by multiple founders, stratified by route of transmission, as calculated in the main analysis (A), following the iterative exclusion of individual studies (B) and bootstrapped estimates recalculated from 1000 datasets in which the representative 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).

Supplementary References

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

Bertolli J, St. Louis ME, Simonds RJ, Nieburg P, Kamenga M, Brown C, Tarande M, Quinn T, Ou C-Y. 1996. Estimating the timing of mother-to-child transmission of human immunodeficiency virus in a breast-feeding population in Kinshasa, Zaire. Journal of Infectious Diseases 174:722–726.

Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, Bhattacharya T. 2010. Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC bioinformatics 11:532.

Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, Karita E, Manigart O, Mulenga J, Keele BF, Shaw GM, et al. 2009. Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1. PLOS Pathogens 5:e1000274.

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proceedings of the National Academy of Sciences 105:7552–7557.

Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, Pham KT, Goepfert PA, Michael Kilby J, Saag MS, et al. 2009. Modeling sequence evolution in acute HIV-1 infection. Journal of Theoretical Biology 261:341–360.

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

Meyerhans A, Vartanian J-P, Wain-Hobson S. 1990. DNA recombination during PCR. Nucleic acids research 18:1687–1691.

Novitsky V, Arnold C, Clewley JP. 1996. Heteroduplex mobility assay for subtyping HIV-1: improved methodology and comparison with phylogenetic analysis of sequence data. Journal of virological methods 59:61–72.

Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, et al. 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. Journal of virology 82:3952–3970.

Simmonds P, Balfe P, Peutherer JF, Ludlam CA, Bishop JO, Brown AJ. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. Journal of virology 64:864–872.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562.

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. Journal of statistical software 36:1–48.

Wickham H. 2012. reshape2: Flexibly reshape data: a reboot of the reshape package. R package version 1.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J. 2019. Welcome to the Tidyverse. Journal of open source software 4:1686.