THE LANCET
Microbe

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Derqui N, Koycheva A, Jie Zhou J, et al. Risk factors and vectors for SARS-CoV-2 household transmission: a prospective, longitudinal cohort study. *Lancet Microbe* 2023; published online April 6. https://doi.org/10.1016/S2666-5247(23)00069-1.

Supplementary content

Supplementary Methods

Supplementary Figures

Supplementary Tables

Additional References

Supplementary Methods

Laboratory assays

INSTINCT Upper Respiratory Tract (URT)-swabs processing and quantitative real-time PCR:

Throat and nostril were sampled with a single URT flocked swab which was placed in UTM (Copan Diagnostics, Murrieta, CA, USA; or MANTACC, Guangdong, China) and transported the same day (except for the day 4 swab which was stored at 2-8°C in the household refrigerator) by the study team to the Molecular Diagnostic Unit, Imperial College London (MDU) where it was stored at 4°C until processed the next working day. Viral RNA was extracted using the innuPREP Virus TS RNA 2.0 Kit on a CyBio Felix (Analytik Jena, Jena, Germany), following the manufacturer's instructions. A multiplex quantitative real-time PCR was used for the E gene and RNase P assay. Reactions were run with the TaqMan Fast Virus 1-Step Master Mix (Thermo Fisher Scientific, Waltham, MA, USA), with the addition of a SARS-COV-2 positive control and negative control. Samples with adequate RNAse P RNA and an E gene Ct <36.5 (which equates to 5 RNA copies per PCR reaction, i.e. 1 E gene RNA copy per μl RNA) were reported as SARS-CoV-2-positive[1]. A contact was deemed PCR-positive if they had any PCR-positive URT-swab in the first 14-days post-enrolment.

ATACCC URT-swabs processing and quantitative real-time PCR:

ATACCC samples were collected same-day via couriers and delivered to UK Health Security Agency (UKHSA) Colindale for RT-qPCR testing. Automated extraction of viral RNA from swabs was performed after aliquoting Universal Transport Medium (UTM) (Copan Diagnostics, Murrieta, CA, USA; or MANTACC, Guangdong, China)

into lysis buffer and adding internal control (IC). Samples were run on a triplex real-time PCR assay using TaqPath™ 1-Step Multiplex Master Mix (Thermo Fisher Scientific, Waltham, MA, USA) to amplify the targets over 40 cycles, which specifically detects SARS-CoV-2 in the ORF1Ab assay target, and Sarbecoviruses including SARS CoV-2 in the E gene target. SARS-CoV-2 positivity was assigned if either ORF1ab or E gene was detected at Ct<35, or if both targets were detected at Ct between 35 and 40. A contact was deemed PCR-positive if they had any PCR-positive URT-swab in the first 14-days post-enrolment (unless this was a single positive value with high Ct (>35), Figure 1).

Serology:

10mL of serum was collected, centrifuged at 2000g for 10 minutes and aliquoted into 500µl aliquots, which were frozen at -20°C. The serological samples were processed at the MDU, Imperial College London. Antibody (IgM and IgG) to SARS-CoV-2 receptor binding domain (anti-RBD) was measured using a two-step double antigen binding assay (DABA) (Imperial College London, London, UK) with recombinant S1 antigen on the solid-phase and labelled recombinant RBD as detector in the fluid-phase as previously described.[2] UK Patent Application No. 2011047.4 for "SARS-CoV-2 antibody detection assay" has been filed. The assay was developed and used in the studies prior to the vaccine rollout in the UK.

Samples with DABA≥1 were deemed positive. Participants were considered baseline seropositive if they had a positive serology result (DABA≥1) at day 0. Participants were defined as having seroconverted if they were seronegative (DABA<1) at day 0 but their DABA levels increased by 2.5 in the binding ratio (optical density sample-to-cut-off ratio) in the first 28 days post-enrolment if PCR-positive, or, to avoid

assessing serological responses to other exposures, in the first 14 days if PCR-negative. Participants that were vaccinated before or during the study period were excluded from the serology sub-cohort.

Surface-swabs and hand-swabs-samples processing and quantitative real-time PCR:

Environmental samples were analysed at Imperial College.[3] Surface samples were collected by trained nurses by swabbing approximately 25cm^2 areas of commonly touched communal surfaces as reported by the residents (Table S4), using flocked swabs moistened in Dulbecco's minimal essential medium (DMEM; Gibco, Waltham, MA, USA). Hand-swabs were collected by a trained nurse by swabbing participants' both hands with flocked swabs moistened in DMEM. Participants were advised not to change their hygiene habits and cleaning procedures prior to sample collection. The swabs were placed in VTM, transported using cold packs and stored at 4°C until processed. Viral RNA detection and absolute quantification was performed using RT-qPCR.

Hand-swabs were considered positive if there was any positive result within 14 days post-enrolment. Environmental surface-swabs were considered positive if they were positive at either sampling time (day 0 or 7).

Viral culture assay:

Viral culture was caried out on all PCR-positive environmental swabs with Ct values <33, comprising 10 out of 22 PCR-positive hand-swabs and 15 out of 33 PCR-positive surface-swabs. Swabs were maintained at 4°C for 36-72 hrs until culture was performed, reflecting the time for transportation and storage while RT-qPCR was performed to ascertain which samples were PCR-positive with Ct values <33. African green monkey kidney (VeroE6) cells expressing human angiotensin-

converting enzyme 2 (ACE2) and transmembrane protease serine 2 precursor (TMPRSS2) were kindly provided by MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow[4] and were used to culture virus from the swabs.

VeroE6 cells were maintained in DMEM (Gibco, Waltham, MA, USA) supplemented with 10% heat-inactivated fetal bovine serum (Gibco, Waltham, MA, USA) and penicillin/streptomycin (10 000 IU/mL and 10 000 µg/mL; Gibco, Waltham, MA, USA). For virus culture, 200 µL of samples was added to 24-well plates. On day 0 and after 5–7 days, cell supernatants were collected, and RT-qPCR to detect SARS-CoV-2 was performed as described above.[1]

Samples with at least 1 log increase in copy numbers for the E gene (reduced Ct values relative to the original samples) after 5–7 days propagation in cells compared with the starting value were considered positive by viral culture.

SARS-CoV-2 variant detection:

INSTINCT primary cases' variants were determined by whole genome sequencing of RNA extracted from URT-swabs (method below). Through contact-tracing data, S-gene Target Failure (SGTF) status was obtained for the presumed primary cases of ATACCC contacts. The data was available only for specimens tested in one of the Lighthouse laboratories.[5] SGTF was used as a proxy for "Alpha" (B.1.1.7) infection, due to the high degree of concordance with sequencing data and the predominance of the strain at the time of analysis (October 2020 – March 2021).[6] Primary cases not identified as "Alpha" were considered "Pre-Alpha".

INSTINCT Whole Genome Sequencing (WGS) and analysis:

INSTINCT samples with a positive RT-qPCR result were submitted for Whole Genome Sequencing to assign lineages and generate phylogenetic trees. The

sample with the highest viral load from each individual was selected. Automated RNA extraction was performed using a CyBio FeliX (Analytik Jena, Jena, Germany) and innuPREP Virus TS RNA Kit 2.0 (Analytik Jena, Jena, Germany) according to the manufacturer's instructions, with a sample volume of 200 µl, without carrier RNA and with an elution volume of 50 µl. cDNA synthesis was then performed using the LunaScript RT SuperMix Kit (NEB, Ipswich, MA, USA) according to the manufacturer's instructions with a total reaction volume of 20 µl and extracted sample volume of 5 µl. Libraries were generated using the EasySeq™ RT-PCR SARS CoV-2 (novel coronavirus) Whole Genome Sequencing kit v1 or v2 (Nimagen, Nijmegen, The Netherlands) according to the manufacturer's instructions. Samples were then pooled and purified with AMPure XP (Beckman Coulter, Brea, CA, USA) magnetic beads. Suitable quality of libraries was confirmed using a Tapestation (Agilent, Santa Clara, CA, USA) and concentrations were measured using the Qubit 1x dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit 4 Fluorometer (ThermoFisher Scientific, Waltham, MA, USA). Pooled libraries were then diluted down to 55 pM. The final pool was then run on an iSeq 100 (Illumina, San Diego, CA, USA) with a total of 322 cycles (151 bp paired reads and 10 bp indices). Generated fastq files were processed using the EasySeq variant pipeline (v0.6.0)[7] which is a Nextflow[8] pipeline that uses fastp,[9] BWA MEM,[10] SAMtools,[11] BCFtools,[11] LoFreq,[12] mosdepth,[13] BEDtools,[14] SnpEff[15] and MultiQC[16] to QC, trim and assemble the reads (using reference sequence NC_045512.2) and then generate a consensus sequence and variant report before assigning a PANGO lineage[17] using pangolin (v3.1.16, lineages version 2021-10-18).[18]

Further sociodemographic and clinical data

Vaccination status of cases and contacts was obtained by linking case data to the National Immunisation Management System (NIMS) using a unique patient identifier (NHS number) or combinations of NHS number, forename, first initial, surname, date of birth, and postcode. Vaccination status was derived at the time of symptom onset or positive test (cases) or exposure (contacts).

Questionnaires assessed the presence/absence of the following, considered as comorbidities:

- Chronic cardiac disease, including congenital heart disease (not hypertension)
- Obesity (as defined by clinical staff)
- Chronic obstructive pulmonary disease (COPD)
- Asthma (physician diagnosed)
- Diabetic
- Chronic kidney disease
- Rheumatologic disorder
- Liver disease
- Malnutrition
- Dementia
- Chronic neurological disorder
- Cancer
- Chronic hematologic disease
- AIDS / HIV

We used the Index of Multiple Deprivation 2019 (IMD), the UK Government's official measure of relative deprivation at a small area level (average population of 1,500)

across England,[19] to assess household deprivation according to postcodes. IMDs are calculated from seven weighted domains of deprivation (income, employment, health, education, crime, barriers to housing and services, living environment), and further ranked and divided into deciles, where lower deciles denote higher levels of deprivation.

Statistical analysis

Univariable associations were performed with a modified χ^2 test accounting for household clustering as previously proposed[20] using the 'aod' package.[21] Household Infection Rate (SIR) was calculated using the package 'epikit'.[22] Adjusted relative risk estimates from multivariable, modified Poisson regression analyses were obtained using the 'gee' package,[23] by implementing a GEE model with a Poisson distribution, a log link and an exchangeable correlation[24] structure to account for household clustering. Robust standard errors were applied to the calculation of the confidence intervals using the 'pubh' package.[25] Cluster size (number of residents excluding the primary case) ranged from 1 to 5 (Figure S1). The estimates were adjusted for five significant demographic or household covariates: sex at birth, presence of comorbidities, vaccination status pre-enrolment, relationship to the primary case, and ratio of household residents to number of bedrooms. Primary case SARS-CoV-2 variant was not adjusted for as the data was unavailable for 197 contacts (47.5%). Supplementary figures were generated using the 'ggplot2'[26] and 'venneuler'[27] R packages.

For the genomic analyses, sequences were aligned using Clustal Omega[28] and the alignment was then used to generate a phylogenetic tree using IQ-TREE (v2.1.3).[29] Plots were generated using R and the 'ggtree' package.[30]

Supplementary Figures

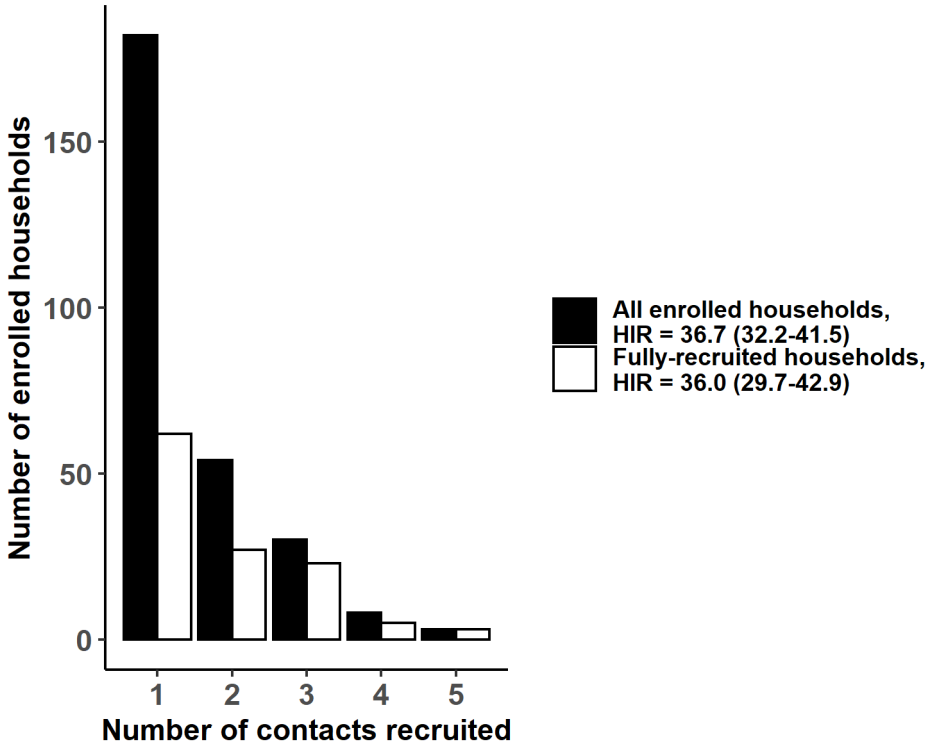

Figure S1: Number of enrolled households by household size. X-axis represents the number of contacts (residents excluding the primary case) recruited per household in: (i) all enrolled households, in black, (the number of contacts recruited may or may not be the total number of contacts in the household), and (ii) households in which all co-inhabitants were recruited as contacts, in white. Y-axis represents the number of households with that many recruited contacts. 200 contacts belonged to fully recruited households (in white), of which 72 were PCR-positive and 128 PCR-negative, giving a HIR = 36.0 (29.7 – 42.9). Abbreviations: PCR, polymerase chain reaction; HIR, household infection rate.

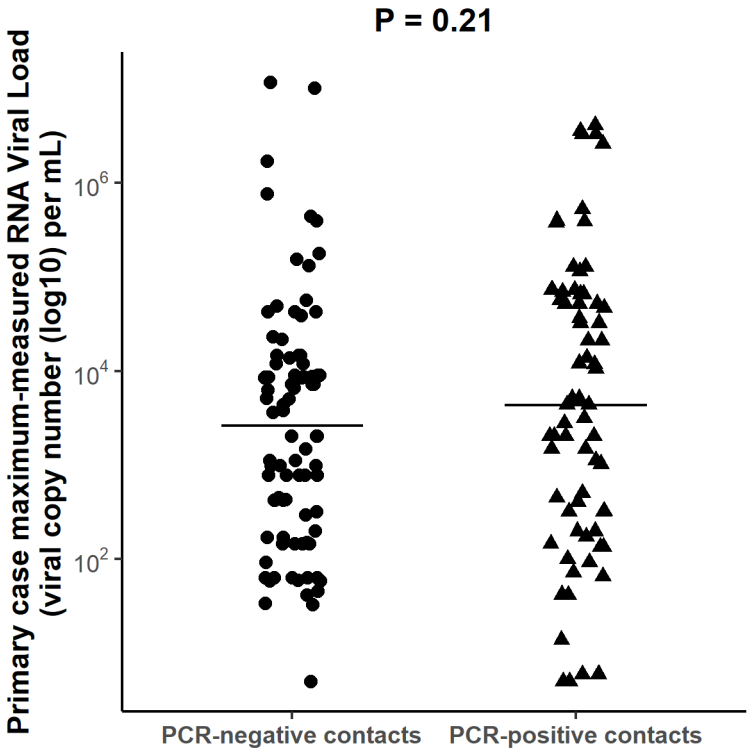

Figure S2: Primary case URT maximum-measured RNA viral load does not correlate with contacts' risk of infection. Maximum-measured primary case upper respiratory peak viral load, expressed as viral copy numbers per mL, among PCR-positive versus PCR-negative contacts. *P*-value indicated for Mann-Whitney test. Abbreviations: PCR, polymerase chain reaction; URT, upper respiratory tract.

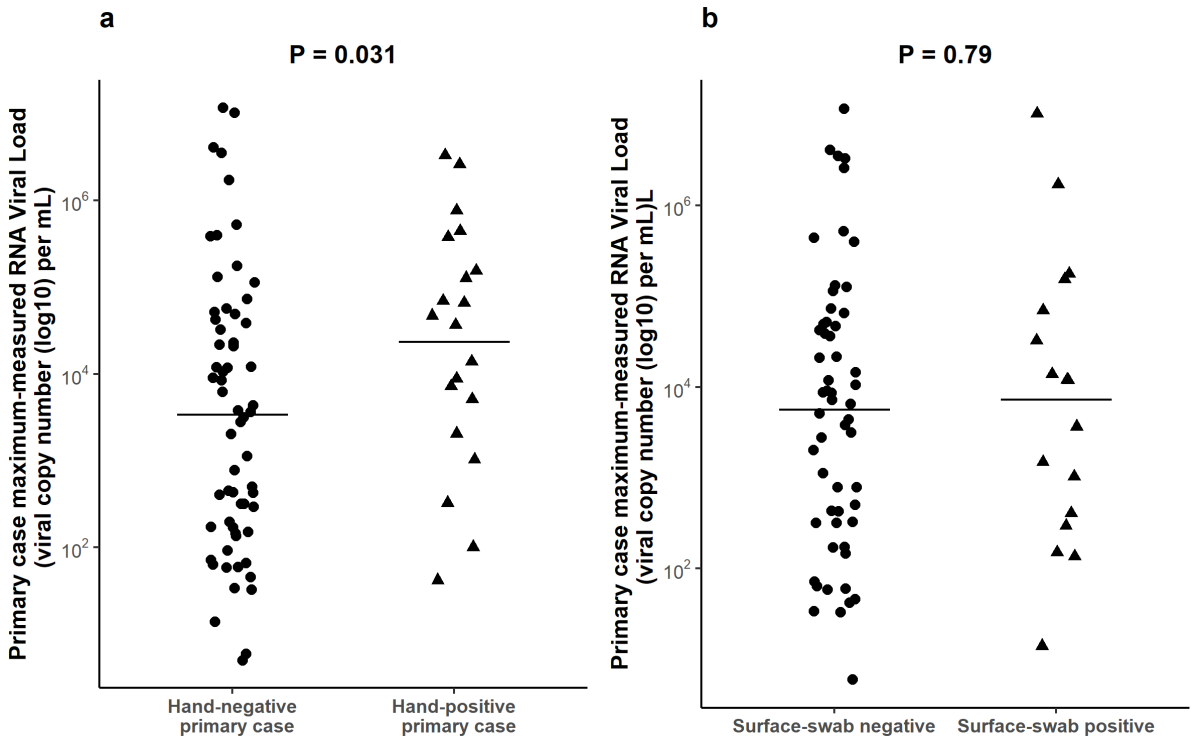

Figure S3: Primary cases' maximum-measured URT RNA viral load in relation to their own respective hand-swab status, and environmental-swabs SARS-CoV-2 PCR results. Measured primary case upper respiratory peak viral load, as viral copy numbers per mL, among **(a)** hand-swab-positive versus hand-swab-negative primary cases, and **(b)** surface swab-positive versus surface swab-negative households. *P*-values indicated for Mann-Whitney test. Abbreviations: PCR, polymerase chain reaction; URT, upper respiratory tract.

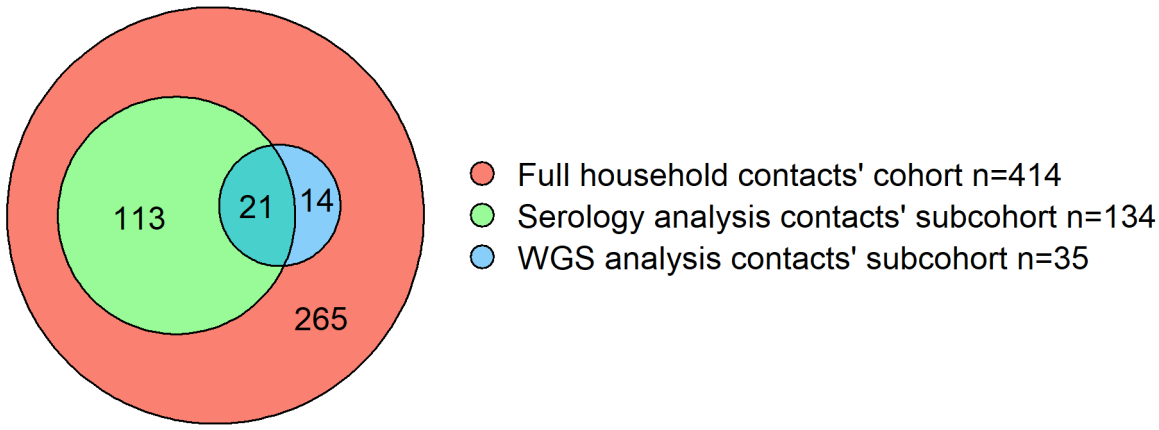

Figure S4: Overlap between full household contacts' cohort (n=414), serology analysis contacts' subcohort (n=134) and WGS analysis contacts' subcohort (n=35). Number of contacts included in each analysis, and their overlap: n=265 household contacts belonged to the full household contacts' cohort but not to the serology analysis or WGS subcohorts, n=113 contacts belonged to the serology subcohort (and hence the full household contacts' cohort) but not to the WGS, n=14 contacts were in the WGS analysis subcohort (and the full household contacts' cohort) and not in the serology subcohort, and n=21 contacts belonged to all three cohorts. Abbreviations: WGS, Whole Genome Sequencing.

Supplementary Tables

Table S1: Baseline demographic and clinical characteristics of household contacts who were household-visited versus those remotely recruited.

	Household contacts, No. (%)							
	Only household-visited (n=231)				Only remotely recruited (n=183)			
Characteristics	**Total (n=231)**	**PCR-positive (n=82)**	**PCR-negative (n=149)**	**χ2 test *P*-value**	**Total (n=183)**	**PCR-positive (n=70)**	**PCR-negative (n=113)**	**χ2 test *P*-value**
Sex at birth								
Male	108 (47)	46 (56)	62 (42)	0.060	92 (50)	29 (41)	63 (56)	0.076
Female	123 (53)	36 (44)	87 (58)		91 (50)	41 (59)	50 (44)	
Age*								
Median (IQR)	32 (23 – 50)	28 (23 – 47)	35 (23 – 51)	0.77	39 (27 – 49)	34 (26 – 49)	40 (27 – 49)	0.48
Age group								
≤16 years	36 (16)	13 (16)	23 (15)	0.18	9 (5)	6 (9)	3 (3)	0.17
17-29 years	70 (30)	23 (28)	47 (31)		54 (29)	16 (23)	38 (34)	
30-49 years	66 (29)	31 (38)	35 (23)		75 (41)	29 (41)	46 (41)	
50-69 years	57 (25)	14 (17)	43 (29)		43 (23)	19 (27)	24 (21)	
≥70 years	2 (1)	1 (1)	1 (1)		2 (1)	0 (0)	2 (2)	
Ethnicity								
White	179 (77)	63 (77)	116 (78)	0.99	146 (80)	59 (84)	87 (77)	0.30
Non-White	40 (17)	14 (17)	26 (17)		33 (18)	10 (14)	23 (20)	
Missing	12 (5)	5 (6)	7 (5)		4 (2)	1 (1)	3 (3)	
Weight (BMI)								
Underweight	5 (2)	1 (1)	4 (3)	0.12	3 (2)	1 (1)	2 (2)	0.064
Normal	80 (35)	22 (27)	58 (39)		83 (45)	25 (36)	58 (51)	
Overweight	57 (25)	26 (32)	31 (21)		54 (29)	19 (27)	35 (31)	
Obese	24 (10)	6 (7)	18 (12)		35 (19)	20 (29)	15 (13)	

Missing	65 (28)	27 (33)	38 (25)	-	8 (4)	5 (7)	3 (3)	-
Presence of comorbidities								
Yes	38 (16)	6 (7)	32 (21)	0.0093	67 (37)	22 (31)	45 (40)	0.25
No	186 (80)	74 (90)	112 (75)		112 (61)	47 (67)	65 (57)	
Missing	7 (3)	2 (2)	5 (3)	-	4 (2)	1 (2)	3 (3)	-
COVID-19 vaccination pre-enrolment								
Yes	16 (7)	3 (4)	13 (9)	0.19	9 (5)	0 (0)	9 (8)	0.018
No	211 (91)	77 (94)	134 (90)		171 (93)	67 (96)	104 (92)	
Missing	4 (2)	2 (2)	2 (1)	-	3 (2)	3 (4)	0 (0)	-
By relationship status to the primary case								
Partner / Parent to child ≤16 years / child ≤16 years/ Sibling in same bedroom	104 (45)	55 (67)	49 (33)	<0.0001	70 (38)	30 (43)	40 (35)	0.0055
Parent to child >16 / child >16 / Sibling without shared bedroom	73 (32)	14 (17)	59 (40)		53 (29)	25 (36)	28 (25)	
Housemate / Residential employee	37 (16)	8 (10)	29 (19)		34 (19)	4 (6)	30 (27)	
Missing	14 (6)	5 (6)	9 (6)	-	11 (6)	5 (7)	6 (5)	-

Note: HIR in the household visited cohort was 35.5% (95% CI: 29.6 – 42.9) and 38.2% (31.5 – 45.5) in the remotely recruited group; $\chi 2$ test comparing both proportions gave a P-value = 0.63. BMI categories considered were underweight (<18.5), normal (18.5 to <25), overweight (25 to <30) and obese (≥30). Presence of comorbidities was considered if the participant had one or more conditions (listed in Supplementary Methods). Vaccination pre-enrolment was defined as having received one or more doses before day 0. P-values are given for each covariate $\chi 2$ test accounting for household clustering, applying for all categories where data is

available, except for median age (*), where Mann-Whitney test was used. Abbreviations: PCR, Polymerase Chain Reaction; IQR, Inter Quartile Range; BMI, Body Mass Index; HIR, household infection rate.

Table S2: Baseline demographic and clinical characteristics of primary cases who infected one or more of their recruited contacts versus those who did not.

	Primary cases, No (%)			
Characteristics	**Total (n=96)**	**With infected contacts (n=52)**	**No infected contacts (n=44)**	**χ2 test P-value**
Sex at birth				
Male	42 (44)	24 (46)	18 (41)	0.76
Female	54 (56)	28 (54)	26 (59)	
Age				
Median (IQR)	36 (25 – 47)	42 (32 – 50)	26.5 (22 – 38)	0.0005*
Missing	5	3	2	
Age group				
≤16 years	7 (7)	3 (6)	4 (9)	0.0015
17-29 years	28 (29)	7 (14)	21 (48)	
30-49 years	38 (40)	26 (50)	12 (27)	
50-69 years	17 (18)	13 (25)	4 (9)	
≥70 years	1 (1)	0 (0)	1 (2)	
Missing	5 (5)	3 (6)	2 (5)	
Ethnicity				
White	83 (86)	45 (86)	38 (86)	0.99
Non-White	12 (14)	6 (11)	6 (14)	
Missing	1 (0)	1 (2)	0 (0)	
Weight (BMI)				
Underweight	3 (3)	2 (4)	1 (2)	0.99
Normal	38 (40)	23 (44)	15 (34)	
Overweight	17 (18)	10 (19)	7 (16)	
Obese	15 (16)	9 (17)	6 (14)	
Missing	23 (24)	8 (15)	15 (34)	
Presence of comorbidities				
Yes	68 (71)	38 (73)	30 (68)	0.90
No	27 (28)	14 27)	13 (29)	
Missing	1 (2)	0 (0)	1 (2)	
Smoking habits				
Current	9 (9)	5 (10)	4 (9)	0.99
Former	18 (19)	10 (19)	8 (18)	
Never	64 (67)	35 (67)	29 (66)	
Missing	5 (5)	2 (4)	3 (7)	
COVID-19 vaccination pre-enrolment				
Yes	15 (16)	10 (19)	5 (11)	0.41
No	80 (83)	41 (79)	39 (89)	
Missing	1 (1)	1 (2)	0 (0)	

Note: Primary cases with transmission were considered if one or more of their household contacts were PCR-positive, and without transmission if all their recruited contacts were PCR-negative for ≥3 days. BMI categories considered were underweight (<18.5), normal (18.5 to <25), overweight (25 to <30) and obese (≥30). Presence of comorbidities was considered if the participant had one or more conditions (listed in Supplementary Methods). Vaccination pre-enrolment was defined as having received one or more doses before day 0. P-values are given for each covariate $\chi 2$ test, applying for all categories where data is available, except for median age (*), where is given for Mann-Whitney test. Household clustering was not considered as all primary cases belonged to different households. Abbreviations: IQR, Inter Quartile Range; BMI, Body Mass Index.

Table S3: Viral detection on environmental surfaces and primary case hands, by hand-swab viral detection status in the full household contacts' cohort (n=414) and in the serology analysis contacts' sub-cohort (n=134).

	Full household contacts' cohort (n=414)					Serial serology sub-cohort (n=134)				
	Contacts' hand-swab SARS-CoV-2 PCR status, No. (%)				Multivariable analysis	Contacts' hand-swab SARS-CoV-2 PCR status, No. (%)				Multivariable analysis
	Total (n=211)	PCR-positive (n=22)	PCR-negative (n=189)	χ2 test *P*	Adjusted Relative Risk (95% CI)	Total (n=119)	PCR-positive (n=16)	PCR-negative (n=103)	χ2 test *P*	Adjusted Relative Risk (95% CI)
Primary cases' hand-swab SARS-CoV-2 PCR status										
Positive	33 (16)	7 (32)	26 (14)	0.12	3.11 (1.21-8.05)	19 (16)	5 (31)	14 (14)	0.29	2.35 (0.89-6.45)
Negative	123 (58)	11 (50)	112 (59)		Ref	65 (54)	9 (56)	56 (54)		Ref
Missing	55 (26)*	4 (18)	51 (27)	-	-	35 (29)*	2 (13)	33 (32)		-
Environmental surface-swabs SARS-CoV-2 PCR status										
Positive	20 (9)	8 (36)	12 (6)	<0.0001	3.77 (1.27– 11.14)	12 (10)	6 (50)	6 (6)	0.0038	3.62 (1.16-11.29)
Negative	147 (70)	12 (55)	135 (72)		Ref	85 (71)	8 (38)	77 (74)		Ref
Missing	44 (21)*	2 (9)	42 (22)	-	-	22 (19)*	2 (12)	20 (19)		-

Note: (*) Does not refer to unavailable hand-swab data, but rather unavailable primary case or environmental swabs data for all the contacts whose hand-swab was available. P-values are given for each covariate χ2 test accounting for household clustering, applying for all categories where data is available. Relative risk estimates from the multivariable analyses were calculated with a modified Poisson regression using robust standard errors, accounting for household clustering, and adjusting for contact's sex at birth, comorbidity status, vaccination pre-enrolment status, relationship status to the primary index and ratio number of residents in

the household to number of bedrooms. Abbreviations: PCR, Polymerase Chain Reaction; CI, Confidence Interval; Ref, reference covariate for regression.

Table S4: Total number, and PCR status, of sampled environmental surfaces from 103 households, by room location.

	Sampled environmental surfaces, No. (%)		
Environmental surfaces	**Total (n=344)**	**PCR-positive (n=33)**	**PCR-negative (n=311)**
General			
Doors/doors' handles	46 (100)	1 (2)	45 (98)
Bins	7 (100)	1 (14)	6 (86)
Stairs/Banister	4 (100)	0 (0)	4 (100)
Living room			
Remotes/TV controllers	14 (100)	3 (21)	11 (79)
Desk/Tables	9 (100)	0 (0)	9 (100)
Kitchen			
Fridge/fridge handles/doors	56 (100)	7 (12)	49 (88)
Kettles/Water filter	49 (100)	7 (14)	42 (86)
Kitchen counter/kitchen tap/sink	75 (100)	5 (7)	70 (93)
Toaster/Microwaves/Oven	10 (100)	6 (60)	4 (40)
Bathroom			
Bathroom sink/tap	17 (100)	2 (12)	15 (88)
Toilet/toilet flush	57 (100)	1 (2)	56 (98)

Note: Average number of surfaces sampled per household: 3.34. Environmental surface-swabs were considered positive if they were positive at either sampling timepoint, i.e. day 0 or day 7. Abbreviations: PCR, Polymerase Chain Reaction.

Additional References

1 Rowan AG, May P, Badhan A, *et al.* Optimized protocol for a quantitative SARS-CoV-2 duplex RT-qPCR assay with internal human sample sufficiency control. *J Virol Methods* 2021; **294**. DOI:10.1016/J.JVIROMET.2021.114174.

2 Rosadas C, Khan M, Parker E, *et al.* Detection and quantification of antibody to SARS CoV 2 receptor binding domain provides enhanced sensitivity, specificity and utility. *J Virol Methods* 2022; **302**: 114475–114475.

3 Cordery R, Reeves L, Zhou J, *et al.* Transmission of SARS-CoV-2 by children to contacts in schools and households: a prospective cohort and environmental sampling study in London. *Lancet Microbe* 2022; published online Aug. DOI:10.1016/s2666-5247(22)00124-0.

4 Rihn SJ, Merits A, Bakshi S, *et al.* A plasmid DNA-launched SARS-CoV-2 reverse genetics system and coronavirus toolkit for COVID-19 research. *PLoS Biol* 2021; **19**: e3001091.

5 Investigation of novel SARS-CoV-2 variant - Variant of Concern 202012/01: Technical briefing 2. Public Health England, 2020.

6 Investigation of novel SARS-CoV-2 variant - Variant of Concern VOC_202012_01: Technical Briefing 5. Public Health England, 2021.

7 Coolen JPM, Wolters F, Tostmann A, *et al.* SARS-CoV-2 whole-genome sequencing using reverse complement PCR: For easy, fast and accurate outbreak and variant analysis. *Journal of Clinical Virology* 2021; **144**: 104993.

8 di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology 2017 35:4* 2017; **35**: 316–9.

9 Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018; **34**: i884–90.

10 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; published online March 16. https://arxiv.org/abs/1303.3997v2 (accessed Dec 16, 2021).

11 Danecek P, Bonfield JK, Liddle J, *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021; **10**: 1–4.

12 Wilm A, Aw PPK, Bertrand D, *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012; **40**: 11189–201.

13 Pedersen BS, Quinlan AR. mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* DOI:10.1093/bioinformatics/xxxxxx.

14 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–2.

15 Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80–92.

16 Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016; **32**: 3047–8.

17 Rambaut A, Holmes EC, O'Toole Á, *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology 2020 5:11* 2020; **5**: 1403–7.

18 O'Toole Á, Scher E, Underwood A, *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* 2021; **7**. DOI:10.1093/VE/VEAB064.

19 English indices of deprivation 2019: mapping resources - GOV.UK. https://www.gov.uk/guidance/english-indices-of-deprivation-2019-mapping-resources (accessed Sept 7, 2021).

20 Donner A, Banting D. Adjustment of Frequently Used Chi-square Procedures for the Effect of Site-to-Site Dependencies in the Analysis of Dental Data. *http://dx.doi.org/101177/00220345890680091201* 1989; **68**: 1350–4.

21 Lesnoff M, Lancelot R. aod: Analysis of Overdispersed Data. 2012. https://cran.r-project.org/package=aod (accessed Dec 9, 2022).

22 Kamvar ZN, Schumacher D. epikit: Miscellaneous Tools for the 'R4Epis' Project. 2020. https://CRAN.R-project.org/package=epikit.

23 Carey VJ, Ported to R by Thomas Lumley and Brian Ripley., Files src/dgedi.f and src/dgefa.f are for LINPACK authored by Cleve Moler. gee: Generalized Estimation Equation Solver. 2019. https://CRAN.R-project.org/package=gee.

24 Gallis JA, Turner EL. Relative Measures of Association for Binary Outcomes: Challenges and Recommendations for the Global Health Researcher. *Ann Glob Health* 2019; **85**: 1–12.

25 Athens J. pubh: A Toolbox for Public Health and Epidemiology. 2022.

26 Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016. https://ggplot2.tidyverse.org.

27 Wilkinson L. venneuler: Venn and Euler Diagrams. 2011. https://CRAN.R-project.org/package=venneuler.

28 Sievers F, Wilm A, Dineen D, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011; **7**. DOI:10.1038/MSB.2011.75.

29 Minh BQ, Schmidt HA, Chernomor O, *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530–4.

30 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8**: 28–36.