# Genomics-enhanced contact tracing enabled the characterization of SARS-CoV-2 transmission chains and their associated infection contexts in the general population during community transmission

Johannes Ptok[1,2#], Jonas Weber[1#], Moritz Pigulla[1,3#], Coco Röhl[1,3], Nils Lüschow[1], Dirk Nagels[3], Torsten Houwaart[1], Samir Belhaj[1,3], Lena-Sophie Schneider[1,3], Dominik Regorz[3], Pascal Kreuzer[3], Yara Fröhlich[3], Patrick Finzer[1,4], Emran Tawalbeh[3], Jessica Nicolai[1], Mygg Stiller[5], Jacqueline Blum[5], Christian Lange[5], Roland Geisel[5], Daniel Wind[1], Lisanna Hülse[1], Alona Tysha[1], Tobias Wienemann[1], Malte Kohns Vasconcelos[1,6], Katrin Hoffmann[4], Nadine Lübke[2], Sandra Hauka[2], Klaus Pfeffer[1], Jörg Timm[2], Lutz Ehlkes[3,&], Andreas Walker[2,&], Alexander T. Dilthey[1,7,&,*]

[1] Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2] Institute of Virology, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3] Düsseldorf Health Authority (Gesundheitsamt Düsseldorf), Düsseldorf, Germany
[4] Zotz | Klimas, Düsseldorf, Germany
[5] Medizinische Laboratorien Düsseldorf, Düsseldorf, Germany
[6] Department for Epidemiology, Institute for Medical Biometry and Epidemiology, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany
[7] Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

# These authors are joint first authors
& These authors jointly supervised this work
* Address for correspondence: dilthey@hhu.de

## Abstract

Understanding how pathogens spread in human populations is key to effective infection prevention, but for most pathogens in community transmission, population transmission chains and their associated infection contexts remain elusive. From February to December 2021, we carried out one of the highest-intensity SARS-CoV-2 genomic surveillance programmes in Europe and implemented genomics-enhanced contact tracing for the characterization of SARS-CoV-2 community transmission. Here, we report on the first retrospective analysis of the complete dataset, integrating data 32,380 cases, 49,906 forward and backward contact tracing records, 162 recorded institutional outbreaks, and 8,028 viral genome sequences (case sequencing coverage 24.5%). 84% of putative transmission detected by forward contact tracing were associated with private households (estimated attack rate 30%), followed by "recreational context" contacts (9% of putative transmissions, 13% transmission probability). Lower transmission probabilities in the 3% – 6.5% range observed for many contexts (e.g. workplaces, schools, residential care facilities) suggested effective infection prevention in these. All institutional outbreaks combined were associated with 5% of total cases; schools and kindergartens exhibited the highest (1.3 and 1.2, respectively), and care homes the lowest (0.2), number of outbreak-linked community cases per outbreak case. A sequencing-led search showed that routine contact tracing underestimated the number of cases associated with school outbreaks (18% increase in associated sequenced cases) and it detected previously unrecognized links between different outbreaks. Combining all available epidemiological data, we identified a putative infection source for 23.7% of sequenced cases; even when ignoring genetic constraints, however, 55.6% of all registered cases remained without a putative transmission source. The available genetic data indicated that a sequencing-led search for epidemiological connections across all sequenced cases could substantially contribute to the detection of transmission events not captured by routine contact tracing (upper bound 60.5% of sequenced cases with a putative infection source, assuming that that all genetic links imply a discoverable epidemiological connection). Different indicators of the ability to characterize SARS-CoV-2 transmission showed no evidence of saturation towards the achieved case sequencing proportion of 24.5%, suggesting that higher sequencing rates would have been desirable. In conclusion, we demonstrated the comprehensive characterization of SARS-CoV-2 transmission in different contexts. Our study informs the implementation and application of genomics-enhanced contact tracing, which could become an instrument for more efficient infection prevention during future pandemics and help navigate difficult societal trade-offs.

# Introduction

Understanding when, where and between whom pathogens are transmitted is the foundational question of infectious disease epidemiology. Knowledge of pathogen transmission patterns, relevant infection contexts and context-specific transmission can contribute to the implementation of effective infection prevention measures. However, during community transmission, the specific source of individual infections and the context they were acquired in often remains elusive to both the affected individual and the public health system. Reasons for this include the fact that there are often multiple potential sources for each individual infection event, which are not easily disambiguated; that some transmissions take place between individuals not known to each other, limiting the efficacy of classical contact tracing; and resource constraints at the level of public health systems, which limit the ability to comprehensively track and investigate transmission chains and their associated infection contexts at the population level.

Pathogen genome sequencing can play an important role in the characterization of pathogen transmission patterns [Harris et al. 2013, Andersen et al. 2015, Mate et al. 2015, Faria et al. 2016, Mellmann et al. 2016, COVID-19 Genomics UK (COG-UK) 2020, Lane et al. 2021, Hjorleifsson et al. 2022, McCrone et al. 2022]. When combined with classical contact tracing, pathogen genome sequencing can enable the tracking of individual transmission chains in the population during community transmission ("genomics-enhanced contact tracing"; Seemann et al. [2020], Hjorleifsson et al. [2022], Walker et al. [2022], Sivertsen et al. [2024], Bludau et al. [2025]). In this context, pathogen genome sequencing can be used to validate or disambiguate epidemiological links detected by routine contact tracing; in addition, it can enable the discovery of links missed by routine contact tracing by directing in-depth epidemiological investigation efforts (e.g., based on in-depth case interviews) to case pairs with otherwise unexplained high genetic similarity [Houwaart et al. 2022]. The implementation of genomics-enhanced contact tracing requires high contact tracing and case sequencing rates; of note, neither pathogen sequencing by itself nor other modeling approaches (e.g. of individual-level mobility patterns, Pullano et al. [2020]) are sufficient to enable the tracking of individual transmission chains and their associated infection contexts.

In 2021, the city of Düsseldorf, the capital of Germany's most populous federal state, implemented a system for genomics-enhanced contact tracing of SARS-CoV-2, which combined best-practice classical contact tracing for each registered case with one of the highest-intensity SARS-CoV-2 genomic surveillance regimes in continental Europe [Houwaart et al. 2022, Walker et al. 2022]. Here, we report on the first retrospective analysis of the complete dataset (Figure 1), with the specific aim of characterizing to which extent the combined contact tracing and genetic data enable insights into the structure of SARS-CoV-2 transmission chains and their associated infection contexts in the general population.

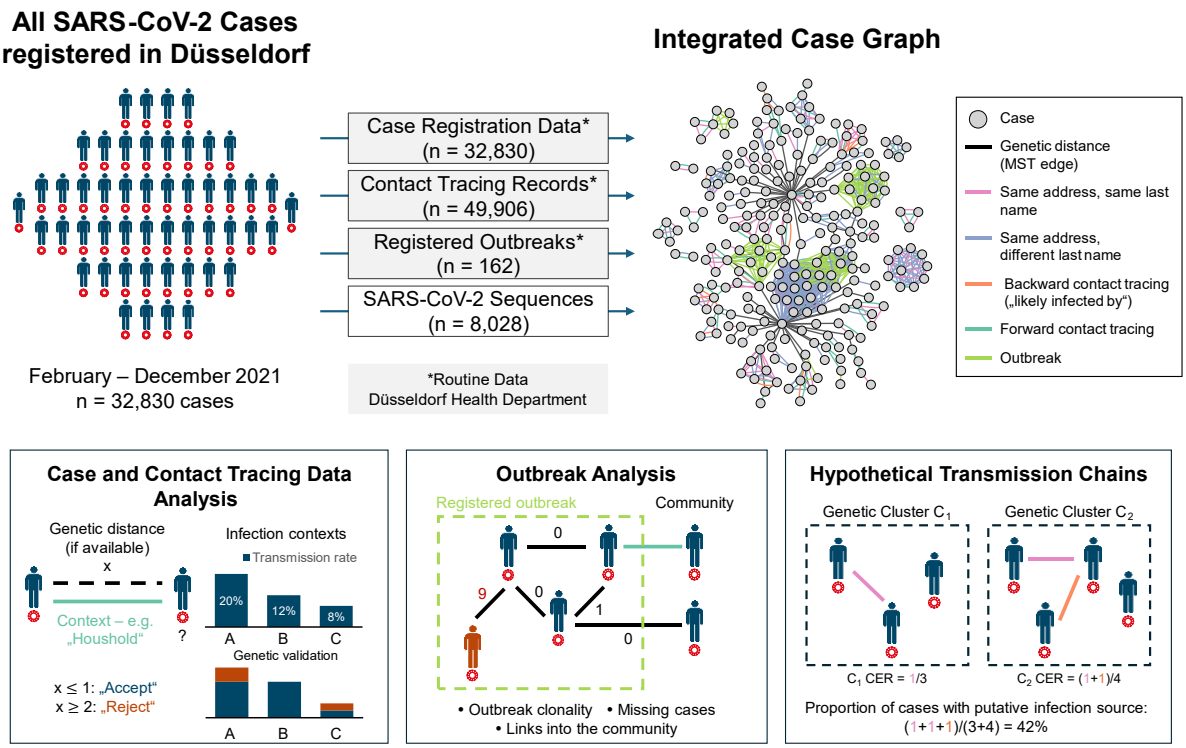


**Figure 1: Characterization of the community transmission of SARS-CoV-2 based on genomics-enhanced contact tracing.** Cases and their epidemiological connections, determined based on Düsseldorf Health Authority data, were represented in the form of an Integrated Case Graph. Genetic distance edges between sequenced cases are included for illustrative purposes; pairwise genetic distances are not stored as part of the graph. Based on the integrated dataset, we mapped infection contexts and their associated transmission probabilities; investigated all registered institutional outbreaks; and attempted to identify a putative infection source for each registered case through hypothetical transmission chains (CER: cluster explanation rate). Across analyses, we used genetic data to evaluate whether epidemiological links between cases were likely associated with transmission.

# Results

## Implementation of genomics-enhanced contact tracing of SARS-CoV-2 in Düsseldorf

Between February and December 2021, the city of Düsseldorf implemented genomics-enhanced contact tracing for SARS-CoV-2. Best-practice classical contact tracing for all diagnosed SARS-CoV-2 cases from Düsseldorf was carried out by Düsseldorf Health Department, based on Germany's Law on the Prevention of Infection ("Infektionsschutzgesetz"); this involved (i) the registration of all detected cases and their personal data (n = 32,830 cases); (ii) contact tracing for all registered cases, with a minimum of three contact attempts using available landline or mobile phone numbers by Düsseldorf Health Authority personnel before contact tracing was deemed unsuccessful (n = 48,131 forward and n = 875 backward contact tracing records); and (iii) the registration and investigation of suspected outbreaks in institutional settings such as schools (n = 162 outbreaks with 919 associated cases). Rapid SARS-CoV-2 sequencing for as many local SARS-CoV-2 isolates as possible was carried out in collaboration between two large commercial diagnostic labs, Düsseldorf Health Authority, and Heinrich Heine University [Houwaart et al. 2022, Walker et al. 2022]. The collected sequencing data (Figure 2A; Supplementary Table 1) comprised n = 11,349 collected viral genome sequences, of which n = 8,763 passed quality control with <= 3,000 undefined characters. The median turnaround time from sample arrival at Heinrich Heine University to availability of a first sequence was 63 hours

(tracked for a subset of the samples collected between April and June 2021; Supplementary Figure 1). All personally identifiable data remained at Düsseldorf Health Authority.

## Unified representation of all epidemiological and genomic data in the Integrated Case Graph

To enable the analysis of the collected data in a unified framework, we developed a scalable data integration approach. Case, contact tracing and pathogen genome sequencing data were represented in the form of an Integrated Case Graph (Figure 1), in which cases were represented as nodes; epidemiological connections between cases were represented as edges; and pathogen genome sequences were attached to the nodes of their corresponding cases (Methods, Supplementary Figure 2). The included epidemiological connections were established based on i) forward and backward contact tracing interviews; ii) assignment of cases to the same outbreak; and iii) the registered address and last name data of cases, inducing an epidemiological edge between two case nodes if the cases were registered at the same address and shared, or did not share, the same last name (the corresponding edge types are referred to as "sameAddress-sameName" and "sameAddress-differentName", respectively). We established record linkage between cases and n = 8,763 SARS-CoV-2 genome sequencing records after quality control based on available metadata and manual curation if necessary; for cases with more than one sequenced viral isolate, a single sequence was selected, resulting in 8,028 cases with viral genomes available (Methods).

The Integrated Case Graph comprised 32,830 case nodes, 8,028 viral genome sequences, and 39,927 links between pairs of cases (edges). "sameAddress"-type edges accounted for the majority of connections between cases (65.6%), followed by forward contact tracing (20.5%), outbreaks (11.7%), and backward contact tracing (2.2%; Figure 2B). The average number of links of any type per case was 2.4; 22.9% of cases had no registered links. The overall proportion of sequenced cases was 24.5%; the proportion of sequenced cases varied between 7.9% and 50.4% on a weekly basis (Figure 2A, Supplementary Table 2) and the sequenced viral genomes were dominated by the Alpha, Delta, and Omicron variants [Rambaut et al. 2020], depending on the time of the year (Supplementary Figure 3).

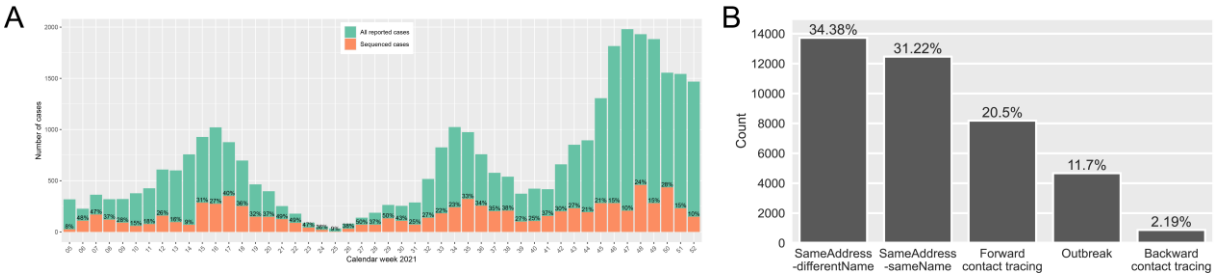


**Figure 2: SARS-CoV-2 sequencing and epidemiological connections represented in the Integrated Case Graph. A.** Shown is the number of registered and sequenced SARS-CoV-2 cases in Düsseldorf on a weekly basis between February and December 2021. **B.** Edges in the Integrated Case Graph, stratified by edge type.

## Routine contact tracing interviews enabled insights into infection contexts and their associated transmission probabilities

First, we asked whether routine contact tracing interviews were informative about relevant SARS-CoV-2 infection contexts and context-specific transmission probabilities. Contact classification was based on the official scheme that was in effect at the time, as published by the Robert-Koch-Institute (historical guidelines not available; current version available under https://www.rki.de/DE/Aktuelles/Publikationen/RKI-Ratgeber/Ratgeber/Ratgeber_COVID-19.html). For 45.5% of cases, at least one contact with an individual registered at a Düsseldorf address was recorded during forward contact tracing (Supplementary Figure 4 B, Supplementary Figures 5 A). 49%

of recorded contacts (Supplementary Table 3) took place in the "private household" context according to contact tracing records of Düsseldorf Health Authority (Figure 3A), followed by contacts in recreational contexts (12%), in schools (8%), in kindergartens (5.3%), and at work (3.3%). 20.7% of contacts were likely associated with a transmission event; for these, the individual that the contact occurred with was registered as a SARS-CoV-2 case within $\leq$ 14 days of the index case (Methods). The probability that a contact was associated with a likely transmission event (secondary attack rate) varied by context (Figure 3A; Supplementary Table 3). Contacts in the domestic context exhibited the highest context-specific transmission rate (30%), and comparatively high transmission rates were also observed in hospitals (15%) and for recreational contexts (13%). By contrast, lower transmission rates were observed in many institutional contexts, such as workplaces, schools, kindergartens, and residential care facilities (3-6%). With respect to the absolute number of n = 8,125 likely transmission events captured by forward contact tracing, household-associated contacts accounted for the large majority (84%), followed by contacts in the recreational context (9%); all corresponding case pairs were connected by an edge of type "forward contact tracing" in the Integrated Case Graph.

To assess whether the epidemiological connections captured by forward contact tracing likely corresponded tp true transmission events, we evaluated genetic distances between the respective cases in the Integrated Case Graph. Genetic distances were available for 10.9% (n = 891) of forward contact tracing case pairs; of these, 80.2% had a genetic distance $\leq 1$, deemed compatible with a direct transmission event and thus classified as "accepted"; the other case pairs were conservatively classified as "rejected" (Figure 3C, Supplementary Figure 6; Supplementary Figure 7 A). The "accepted" case pairs exhibited lower case interval distributions than the ones classified as "rejected" (Supplementary Figure 8 A and B). Furthermore, to confirm that the observed distance distribution did not result from incidental genetic similarity of cases sequenced at approximately the same time, we generated a matched control dataset with randomly scrambled contact tracing links (Methods), which showed a markedly different distribution of genetic differences (average genetic distance 21, with only 0.6% of 1,011 randomized case pairs showing a genetic distance of ≤ 1; Supplementary Figure 9 D). At 80% and 88% of sequenced case pairs classified as "accepted", forward contacts in the "private household" and "recreational context" contexts were not significantly different with respect to genetic validation (Figure 3D; p = 0.42, Fisher's Exact Test). Backward contact tracing data were only available for 2.7% of cases; genetic distances were similar for cases connected by forward and backward contact tracing edges (Figure 3C), and manual context analysis within the subset of sequenced case pairs confirmed the importance of household- and recreational activity-related transmission (Supplementary Figure 10).

In summary, we could characterize relevant infection contexts and context-specific putative transmission probabilities based on routine contact tracing interviews; integration of pathogen sequencing showed that the large majority of contacts between cases were likely associated with transmission events. However, only 40% of cases in the Integrated Case Graph were connected to another case via a contact tracing edge, implying a substantial proportion of relevant epidemiological connections were not captured by contact tracing, and thus unknown to the index case.

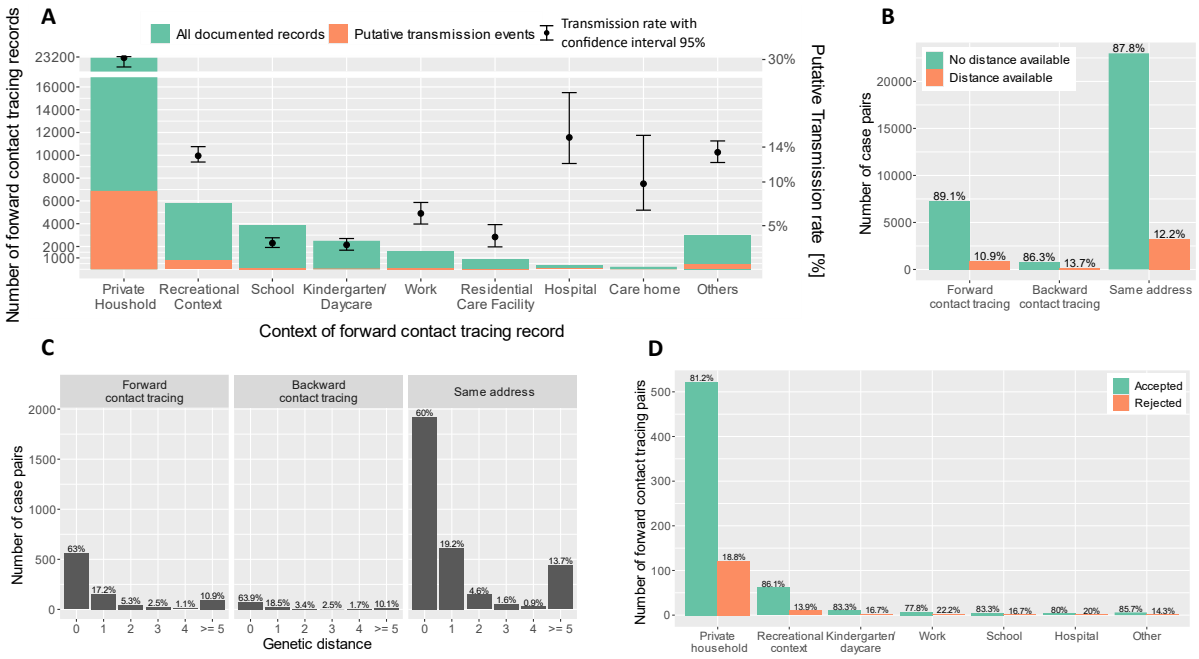


**Figure 3. Integrated genetic analysis of contact tracing, infection contexts, and address- and name-based proxy indicators. A.** Forward contact tracing contacts by context and associated putative context-specific transmission rates. Putative transmission rates were calculated for index cases registered until 17 December 2021. **B.** Number of case pairs connected by different types of links in the Integrated Case Graph; shown in red is the proportion of case pairs for which a genetic distance was, based on the sequencing status of the corresponding cases, available. **C.** Distribution of SARS-CoV-2 genetic distances for case pairs linked by different types of edges in the Integrated Case Graph, stratified by edge type and only shown for case pairs for which a genetic distance was available. **D.** Infection contexts captured by forward contact tracing (case pairs with genetic distances only). Shown is the number of case pairs that were connected by a forward contact tracing edge assigned to the corresponding putative context, stratified by whether the genetic distance between the connected cases was $\leq 1$ ("accepted") or $> 1$ ("rejected").

## Address-based proxy indicators of joint household membership pointed to domestic transmission of SARS-CoV-2 not captured by contact tracing interviews

We hypothesized that contact tracing interviews may not have captured all relevant contacts in the domestic context. As we did not have access to direct information on household membership, we relied on name and address information as proxy indicators. Edges of type "sameAddress-sameName" and "sameAddress-differentName", connecting cases in the integrated Case Graph with the same address and registration dates less than 14 days apart, linked 19,161 case pairs that were not already linked by contact tracing ("sameAddress-sameName": 12,467 case pairs; "sameAddress-differentName": 13,728 case pairs; Supplementary Figure 11). Applying the same threshold as for the analysis of contact tracing edges, 84% of n = 1,826 edges of type "sameAddress-sameName", and 73% of n = 1,382 edges of type "sameAddress-differentName", that connected two sequenced cases were classified as "accepted" (Supplementary Figure 12; difference in proportion "accepted" significant at $p < 0.001$, Pearson's Chi-squared test). Similar to contact tracing-implied links classified as "rejected", "sameAddress"-type links classified as "rejected" exhibited increased case interval distributions (Supplementary Figures 8). 27% of case pairs connected by an accepted "sameAddress"-type link reported symptom onset on the same day (compared to 7% for accepted forward contact tracing links), suggesting infection by a third member of the same household (see Supplementary Figure 13); consistent with this, cases connected by "sameAddress"-type often formed small clusters with more than 2 cases in the Integrated Case Graph (Supplementary Figure 14). Of note, 32 out of 3,761 connected components in the Integrated Case Graph associated with edges of type "sameAddress-

differentName" comprised more than 10 cases, suggesting that edges of this type may sometimes also capture transmission in residential facilities or large building complexes (Supplementary Figure 14).

In summary, a significant number of additional epidemiological links between cases, capturing putative contacts in the domestic or residential contexts, could be recovered from case registration data; compared to the data collected by contact tracing, the total number of address-based edges in the Integrated Case Graph exceeded the number of contact tracing-implied edges by more than a factor of 3. Analysis of genetic data showed that many of address-based epidemiological links were likely associated with direct transmission of SARS-CoV-2 or at least a share infection source.

## Institutional outbreaks were associated with approximately 5% of overall SARS-CoV-2 case load

To characterize transmission of SARS-CoV-2 in the context of institutional outbreaks and the contribution of outbreks to overall community transmission, we carried out an analysis of all n = 162 institutional outbreaks with 919 associated cases recorded in Düsseldorf during the study period (Supplementary Table 4). The number of cases varied by outbreak category; 49 school outbreaks accounted for 230 registered cases, followed by 13 care home outbreaks accounting for 161 cases, and 23 outbreaks in kindergartens/daycares accounting for 121 cases; care home outbreaks had the highest number of cases per outbreak (mean = 12.4; Figure 3B; Supplementary Table 5). To evaluate the clonality of outbreaks and whether the assignment of individual cases to outbreaks was correct, we used genome sequencing data. In 48 outbreaks more than one sequenced case was present; of these, 37.5% (n = 18 outbreaks) were classified as non-clonal because they contained at least one case that was not clonally related to the other outbreak cases (genetic distance threshold of $\leq 1$; Methods). Based on the same threshold, 15% (n = 31) of the 203 sequenced cases in the outbreaks with more than one sequenced case were found to be likely mis-assigned, i.e. not clonally related to the other cases in their corresponding outbreak (Methods).

As a first-order measure of the role of outbreaks in the wider transmission network of SARS-CoV-2, we determined the number of community cases connected to outbreaks (represented by any one of their cases) via links of type contact tracing and "sameAddress" in the Integrated Case Graph. 659 community cases were found to be epidemiologically linked to at least one outbreak, accounting, when combined with outbreak cases, for approximately 5% of cases in Düsseldorf over the considered period. Of the identified outbreak-linked cases, 22% (n = 150 / 659) could be evaluated genetically based on the availability of a genetic distance between the outbreak-linked case and any one outbreak case; of these, 65% (n = 98 / 150) were confirmed based on a genetic distance of $\leq 1$. Schools and kindergartens combined accounted for approximately two thirds of overall and sequencing-confirmed linked cases (Supplementary Table 5); these two outbreak types also had the highest number of linked community cases per outbreak case (5.9 and 6.2, summed over all school and kindergarten outbreaks; Supplementary Table 5, Supplementary Figure 15 L). By contrast, care homes exhibited the lowest proportion of linked community cases relative to registered outbreak cases (average = 0.2; Supplementary Table 5) and were therefore, despite high overall case count, associated with a relatively low absolute number of linked community cases. "sameAddress"-type links accounted for 1.7 times as many overall and approximately twice as many sequencing-confirmed linked community cases as contact tracing-implied links, respectively (Supplementary Table 5).

Only considering genetic similarity (Methods), 981 community cases were found to be linked to an outbreak; of these, 94% did not exhibit an epidemiological link to the outbreak they were genetically linked to. School and nightlife outbreaks were associated with the largest number of genetically linked community cases (n = 234 and n = 199, respectively; Supplementary Table 5); nightlife outbreaks also

had the highest number of genetically linked community cases when normalized by outbreak size (Supplementary Table 5, Supplementary Figure 15 M).

In summary, integrated analysis of outbreak and genetic data showed differences in terms of total outbreak-assigned cases and linked community cases between different outbreak types; the sequencing data also confirmed that the outbreak classification carried out by Düsseldorf Health Authority based on epidemiological data alone was largely correct. The overall contribution of institutional outbreaks to SARS-CoV-2 transmission was, however, limited, as outbreaks and their associated cases only accounted for 5% of overall cases.

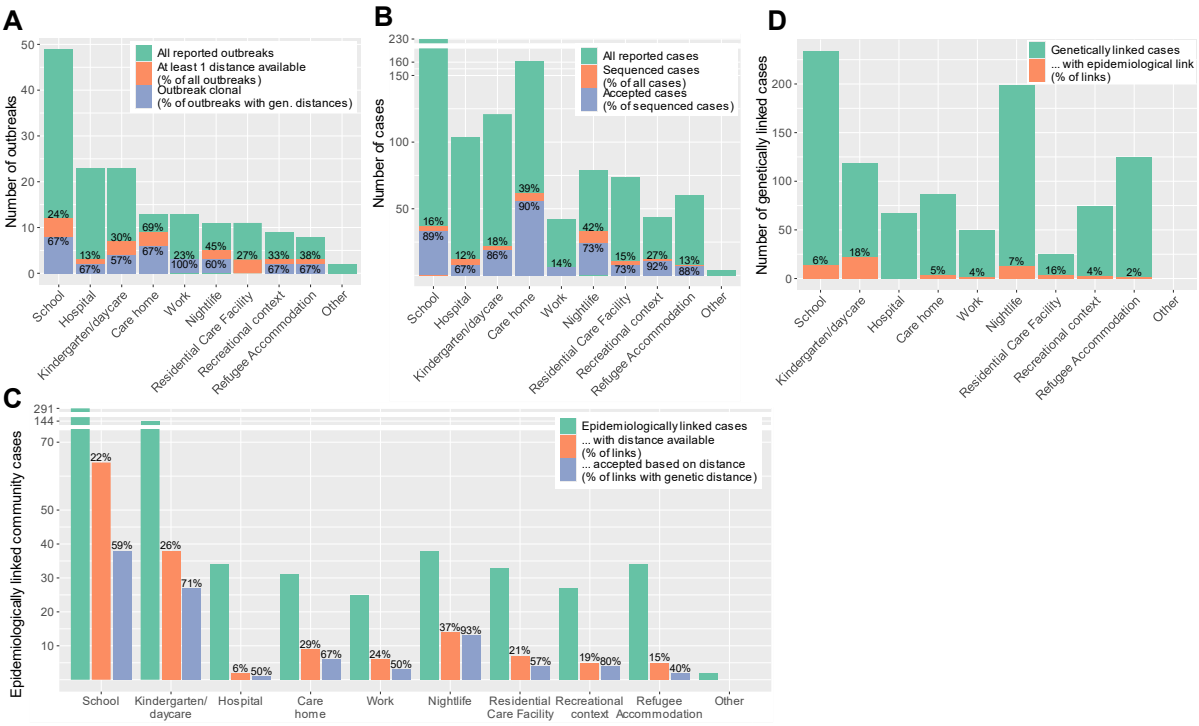


**Figure 4. Genomics-based investigation of SARS-CoV-2 outbreaks and their links into the community. A.** SARS-CoV-2 outbreaks by outbreak type; also shown is the proportion of outbreaks in the respective category with ≥2 sequenced cases and the proportion of outbreaks classified as clonal based on a genetic distance threshold of ≤1 (see Methods). **B.** SARS-CoV-2 cases in outbreaks by outbreak type; also shown is the proportion of sequenced cases in the respective outbreak category and the proportion of sequenced cases in the corresponding outbreak category classified as "accepted" (i.e., part of the largest genetic cluster in the outbreak, based on a clustering threshold of ≤1; Methods). **C.** Cases epidemiologically linked to outbreaks, by outbreak type; also shown is the proportion of epidemiologically linked cases in the corresponding outbreak category for which a genetic distance to their associated outbreak could be computed, and the proportion of such cases for which the computed distance was ≤1 ("accepted"). The genetic distance between a case epidemiologically linked to an outbreak and the outbreak was defined as the minimum genetic distance between the linked case and any case of the outbreak; computation of a such a distance was possible if the outbreak contained at least one sequenced case and if the linked case itself was sequenced. **D.** Cases genetically linked to outbreaks, by outbreak type; also shown is the proportion of genetically linked cases that also exhibited an epidemiological link to their corresponding outbreak.

## In-depth manual analysis of school outbreaks identified additional outbreak-linked cases and links between outbreaks

Prompted by the observed high degree of connectedness between school outbreaks and the local community, and the potentially important role of schools for SARS-CoV-2 transmission in general, we

characterized SARS-CoV-2 transmission within the n = 49 registered school outbreaks and their links to the community in detail (Supplementary Table 5; Supplementary Table 4).We (i) manually reviewed the complete Düsseldorf Health Department database records of all relevant cases and (ii) searched for additional outbreak-linked community cases based on genetic data (Methods; Supplementary Figure 16). According to routine contact tracing, the investigated outbreaks comprised 226 cases, of which 51 were sequenced and 4, based on genetic distances, falsely assigned (i.e. not part of the outbreak they were assigned to by routine contact tracing); as well as epidemiological links to 291 community samples, of which 38 were genetically validated (Figure 5A).

First, we carried out a search for additional outbreak-linked cases. In some instances, free-text database entry fields were used to record the results of contact tracing interviews; we evaluated whether contacts recorded in these (i.e., in a non-structured manner and thus not amenable to automated extraction) increased the number of outbreak-linked community cases, but found only a single instance of a non-structured contact tracing record implying a link between a community case and a school outbreak that was not already represented in the Integrated Case Graph (via an edge of any type between the contact and any case of the associated outbreak). We turned to the analysis of genetic data for the identification of additional outbreak-linked cases and found 20 sequenced cases (Figure 5A) that were not linked to a school outbreak in the Integrated Case Graph or mentioned in free-text contact tracing records, but for which a genetic link to a school outbreak (genetic distance threshold ≤ 1) could be epidemiologically confirmed through manual review of case database records, typically based on information found in other free-text entry fields. For example, in the analysis of outbreaks 0107 and 0108 (Figure 5B), Case 24 was classified as outbreak-linked because it genetically clustered with the outbreaks and because the case was found to be a pupil at the school at which the outbreaks occurred, confirming an epidemiological link. In addition, we found 7 non-sequenced cases that were classified as outbreak-linked because they represented the “missing epidemiological link” between an outbreak and a genetically linked third case (Figure 5A, Methods); such as Case 23, which represented the only detected epidemiological link that could account for the observed genetic link between Case 38 and outbreaks 0107 and 0108 (Figure 5B).

Second, we evaluated whether the assignment of cases to outbreaks was complete. We focused on n = 58 sequenced cases that were not registered as school outbreak cases according to the original outbreak documentation carried out by Düsseldorf Health Authority, but genetically and epidemiologically linked to a school outbreak (including the 20 genetically identified outbreak-linked cases described above). Of the 58 investigated cases, 11 met the criteria for being classified as outbreak cases – for example Cases 23 and 24, as pupils at the school at which outbreaks 0107 and 0108 occurred (Figure 5B). Correct registration of the 11 cases would have increased the number of sequenced outbreak cases by 22% (compared to a baseline of 51 sequenced outbreak cases, of which at least 4 were misclassified according to genetic data; Figure 5A and Supplementary Table 5).

Third, we investigated whether there were connections between separately registered school outbreaks and found that 8 school outbreaks were epidemiologically and genetically connected to at least one other school outbreak (Supplementary Table 6; Supplementary Note 1). Examples include the genetic and epidemiological links between outbreaks 0107 and 0108 (Figure 5B), which were two separately registered multi-class outbreaks at the same school, as well as the detected links between the separately registered outbreaks 0212, 0231, and 0231 at the same school. Due to the detected epidemiological links, these, like the other outbreaks found to be linked in our analysis, should have been registered as single outbreaks.

In summary, the number of cases associated with school outbreaks was higher than initially estimated (18% increase when measured within the set of sequenced cases, Figure 5B); the absolute number of

school outbreak-associated cases, however, remained relatively low. Relevant epidemiological connections that were not captured by the automated data extraction approach used to construct the Integrated Case Graph could be recovered from Düsseldorf Health Authority database records based on the analysis of sequencing data.

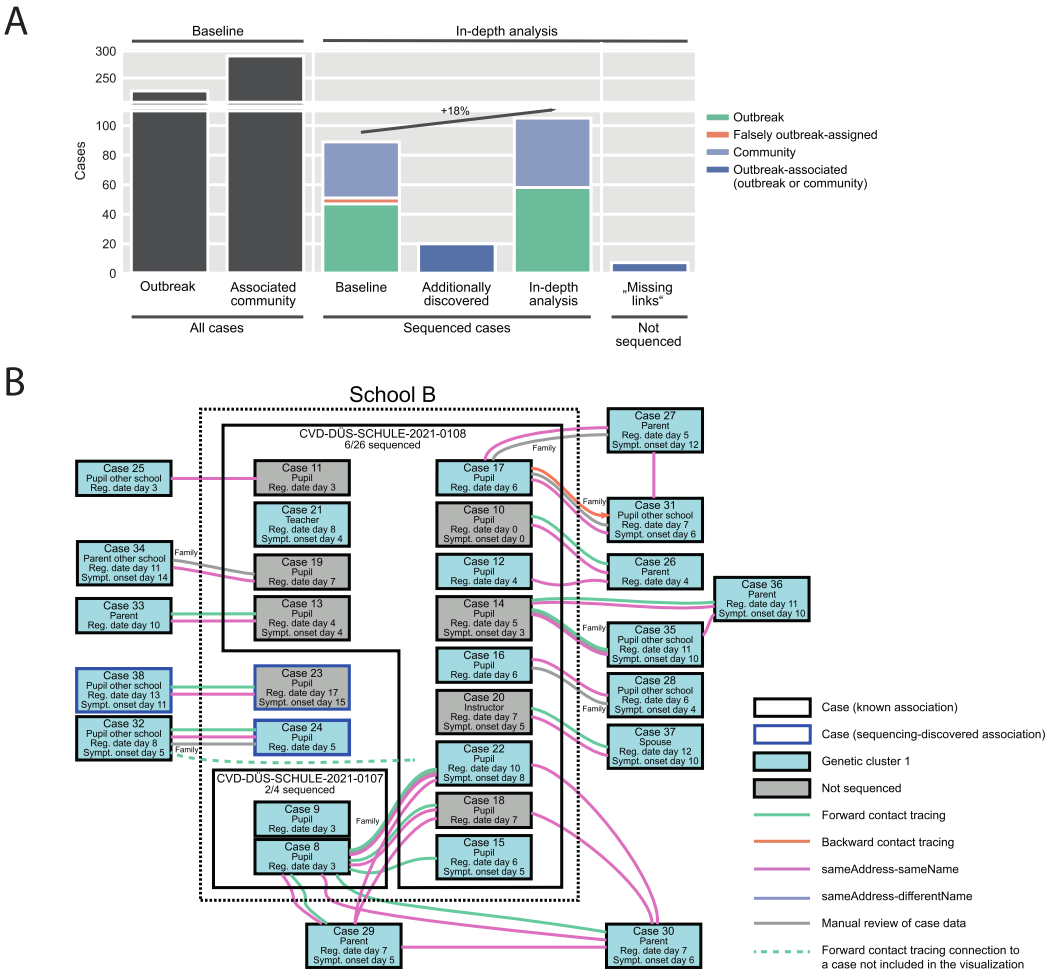


**Figure 5. In-depth analysis of school outbreaks. A.** School outbreak-associated cases before and after in-depth manual analysis. Black bars show the total number of school outbreak-associated cases before manual analysis and independent of sequencing status; colored bars show the results of in-depth analysis for the subset of sequenced outbreak-associated cases, as well as the number of non-sequenced cases that were classified as outbreak-linked because they represented the required "missing epidemiological link" between an outbreak and a genetically linked third case. **B.** Visualization of two school outbreaks (0107 and 0108), registered at the same school, and of the epidemiological links between the associated cases. All sequenced cases (boxes with blue background) were part of the same genetic cluster (see Supplementary Note 1 for a full genetic distance matrix); for outbreak 0108, only a subset of cases is shown. Cases 23, 24, and 38 were classified as outbreak-associated based on the analysis of genetic data; Case 23 is an example of a "missing epidemiological link" that was classified as outbreak-associated because it represented the only detected epidemiological link that could account for the genetic link between Case 38 and the two shown outbreaks.

## Integrated analysis of SARS-CoV-2 transmission identified a putative infection source for almost a fifth of sequenced cases

Having analyzed SARS-CoV-2 transmission in specific epidemiological and institutional contexts, we next set out to evaluate how well we could characterize transmission within the complete dataset of all registered SARS-CoV-2 cases from Düsseldorf in an integrated fashion. We clustered cases by genetic similarity; constructed hypothetical transmission chains for each cluster based on the

epidemiological connections represented in the Integrated Case Graph; and measured for how many cases we found a putative infection source in the constructed hypothetical transmission chains. In constructing hypothetical transmission chains, we only allowed for links between cases that were consistent with respect to the symptom onset dates of the connected cases; that had a genetic distance smaller than or equal to the chosen clustering threshold; and we ensured that each case in the constructed chains had at most one incoming edge, corresponding to the hypothetical infection source of the case (Methods).

We first clustered using a genetic similarity clustering threshold $t_g = 0$ and obtained n = 1,206 clusters of genetically related cases with $\geq 2$ cases that comprised 59.82% (n = 4,802) of all sequenced cases (Supplementary Table 7). The 1,206 clusters had an average size of 3.98 cases and cluster sizes remained relatively constant over the course of the year (Supplementary Figure 17). Hypothetical transmission chains constructed for the computed clusters implied a putative infection source for 31.4% of cases in the clusters (referred to as "Cluster Explanation Rate" or "CER"); for 18.8% of all sequenced cases; and for 4.6% of all registered cases (Figure 6A; Supplementary Table 8, entries with $t_g = t_d = t_j = 0$). The constructed hypothetical transmission chains were dominated by edges of type "sameAddress", followed by "forward contact tracing" edges (Figure 6B); 89% of putative transmissions were household-associated (Supplementary Table 9; "sameAddress"-type edges plus "household"-labeled contact tracing edges). The computed proportion of cases with an infection source was most sensitive to the removal of "sameAddress"-type of edges (Supplementary Figure 18). At the level of individual clusters, CERs of 0 and 0.5 were most frequently observed; larger clusters generally exhibited slightly lower CERs; and the highest CER values were observed for clusters detected in May and June 2021 (Supplementary Figure 19).

We then allowed for the inclusion of intermediate non-sequenced cases in the constructed hypothetical transmission chains and assessed the impact of the chosen genetic clustering threshold (Supplementary Table 8). Allowing for up to two non-sequenced intermediate cases in the constructed hypothetical transmission chains (i.e. setting parameter $t_j = 2$; see Methods), we observed a slightly increased CER of 32.1% and a slight increase in the proportion of sequenced cases with a potential infection source (19.2%). Increasing the threshold $t_g$ used for genetic clustering to 1 led, as expected, to a decreased number of genetic clusters (n = 1,047) and an increased proportion of sequenced cases comprised in the genetic clusters (75.26%). Based on the computed clusters, we observed a slightly decreased CER (31.5%; $t_j = 2$), but an increased proportion of sequenced cases with a potential infection source (23.7%). The inclusion of non-sequenced intermediate cases and modifying the threshold used for genetic clustering therefore only had modest effects; in all considered instances, the CER remained at approximately 30% and the proportion of sequenced cases with a putative infection source at approximately 20%.

Last, we asked how the ability to characterize SARS-CoV-2 transmission was limited by the availability of epidemiological information, and to which extent sequencing may contribute to the discovery of additional epidemiological connections. To obtain an upper bound on the proportion of cases that a putative infection source could be identified for based purely on known epidemiological connections, we treated the complete Integrated Case Graph of all n = 32,830 cases as one large cluster of genetically identical cases and found that the proportion of cases for which a potential infection source could be identified for when ignoring genetic constraints was 44.4% (with 89% of putative transmissions household-associated; Supplementary Table 9). More than half of registered cases thus remained without a putative infection source even based on all epidemiological connections represented in the Integrated Case Graph. As observed for school outbreaks, sequencing, however, may enable the discovery of additional epidemiological connections. To obtain an upper bound on the

potential impact of the sequencing-led discovery of epidemiological links, we re-computed hypothetical transmission chains assuming that all genetically linked samples were also epidemiologically connected; that is, for the purposes for this analysis, we assumed that an epidemiological connection could be found whenever the genetic distance between cases was below the chosen clustering threshold. For $t_g = 0$, we observed that the proportion of sequenced cases, and of all cases, with a defined infection source increased to 43.9% and 10.7%, respectively; the computed hypothetical transmission chains implied an infection source for 2,018 cases for which the directionalized Integrated Case Graph contained no genetically supported incoming epidemiological links. A further increase in the proportion of cases with a putative infection source was observed for $t_g = 1$ (60.5% of sequenced and 14.8% of all cases; Supplementary Table 8). While it is unlikely that epidemiological connections exist for all genetically linked cases, the sequencing-led discovery of unrecognized epidemiological connections may thus substantially contribute to understanding SARS-CoV-2 transmission chains. Similar results for all analyses were obtained for a "maximum symptom onset difference" parameter value $t_d = 3$.

In summary, integrated analysis enabled the construction of hypothetical transmission chains, implying putative infection sources for approximately a third of cases in genetic clusters and a fifth of all sequenced cases. Even when ignoring genetic data, however, more than half of all registered cases remained without a putative infection source, indicating that a substantial number of relevant epidemiological connections remained unknown.

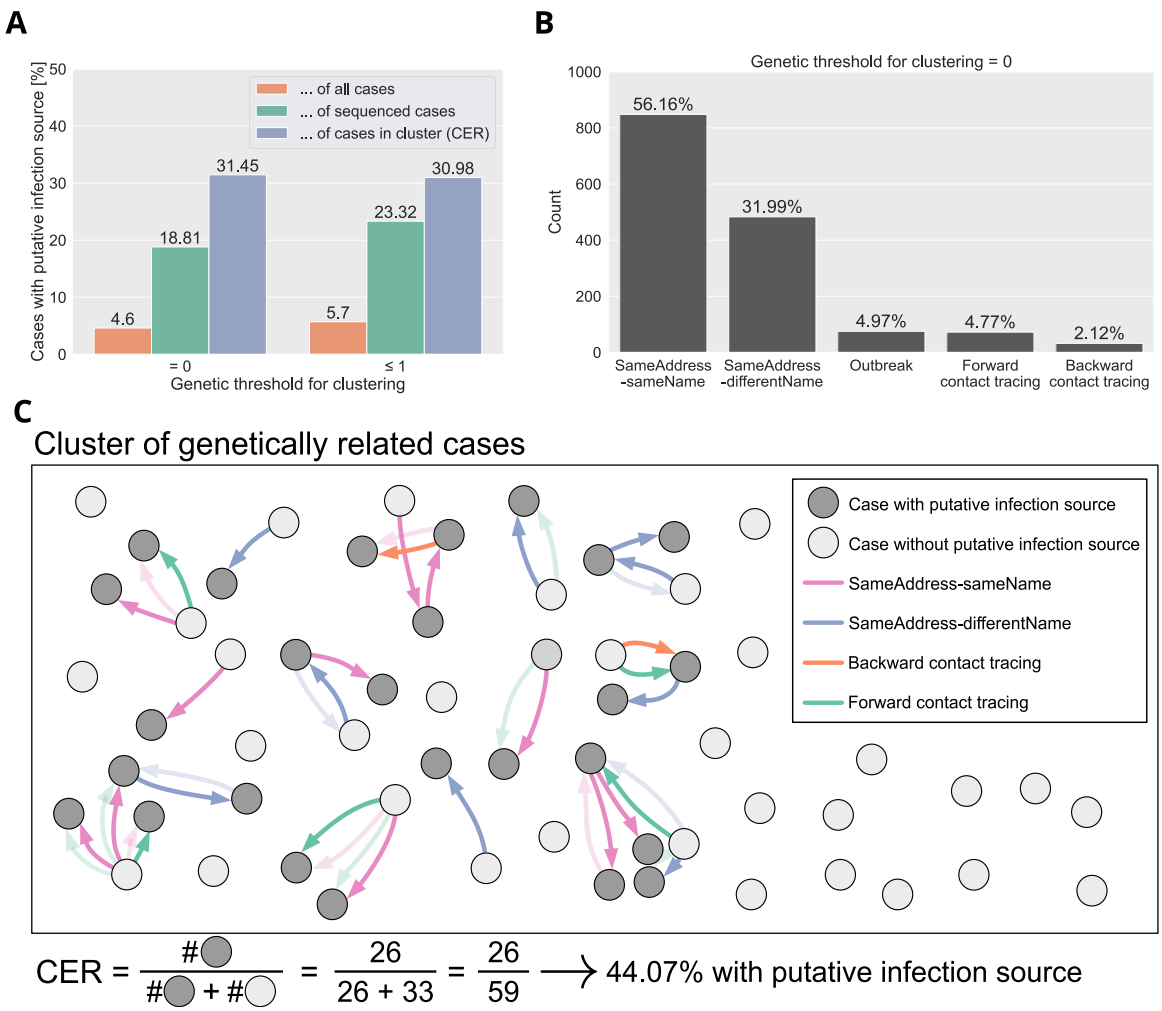


**Figure 6: Integrated characterization of SARS-CoV-2 community transmission. A.** Proportion of cases with a putative infection source, stratified by case sequencing and cluster membership status and based on hypothetical transmission chains (see Methods) computed based on genetic clustering similarity thresholds $t_g$ of 0 and 1 and. **B.** Edges contained in the computed hypothetical transmission chains for $t_g = t_j = 0$ and $t_d = 0$, stratified by edge type. **C.** Visualization of a real cluster of genetically related cases (cluster "27" of the $t_g = 1$ $t_d = 0, t_j = 0$ cluster computation; see Supplementary Table 7) and illustration of the CER computation. Nodes correspond to cases and edges to epidemiological connections of different types (after directionalization). Edges rendered with full opacity are part of the hypothetical transmission chain computed for the cluster; faded edges were not selected. We manually added 4 edges to illustrate that some case pairs maintain bidirectional links after directionalization.

## Transmission characterization was limited by the proportion of sequenced cases

Last, we investigated the relationship between the proportion of sequenced cases and the ability to characterize SARS-CoV-2 transmission (Figure 7, Supplementary Figure 20). Specifically, we considered how changing sequencing rates affected analyses in three key areas, downsampling the set of available SARS-CoV-2 sequences when actual pairwise genetic distances were required, or randomly increasing the proportion of sequenced cases otherwise. We found that the ability to genetically validate epidemiological edges in the Integrated Case Graph monotonically increased with the proportion of sequenced cases; for example, a case sequencing proportion of approximately 75% was required for genetic evaluation of 50% of forward contact tracing links (Figure 7A). With the respect to the investigation of outbreaks, the proportion of outbreaks that could be analyzed for clonality (based on > 1 sequenced cases), as well as the absolute number of community cases found to be

genetically linked to outbreaks, exhibited a strong and monotonic dependence on the overall proportion of sequenced cases (Figure 7 B and C). Finally, we also observed that the proportion of sequenced cases with an identified putative infection source increased with case sequencing coverage; that is, the probability to find the putative infection source for an individual case increased with the overall number of sequenced cases in a "network effect"-type manner, with no evidence of saturation towards the empirically achieved sequencing rate of 24.5% (Figure 7D). All considered indicators therefore suggested that a higher case sequencing rate would likely have enabled further improvements in the ability to characterize SARS-CoV-2 transmission.

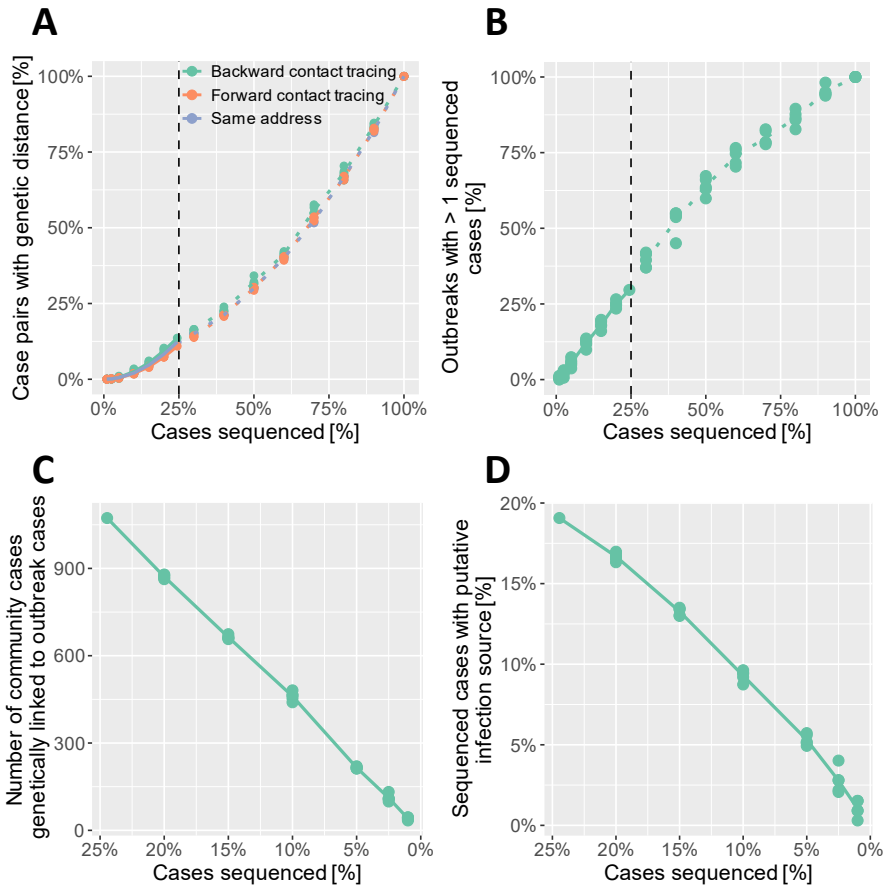


**Figure 7: Influence of the proportion of sequenced cases on the ability to characterize SARS-CoV-2 transmission. A.** Proportion of edges in the Integrated Case Graph that could be genetically validated at different case sequencing rates, shown for different edge types. **B.** Proportion of registered outbreaks that could be genetically validated (based on > 1 sequenced outbreak case). **C.** Number of community cases genetically linked to outbreaks. **D.** Proportion of sequenced cases with a putative infection source, based on hypothetical transmission chains re-computed for each replicate. Each considered case sequencing rate was represented by five replicates.

## Discussion

In 2021, we carried out one of the highest-intensity SARS-CoV-2 integrated genomic surveillance programmes in continental Europe and implemented genomics-enhanced contact tracing (Figure 1, Figure 2). The resulting data set of more than 30,000 cases, almost 50,000 contact tracing records, and 8,000 viral genomes represents a uniquely rich resource on the structure of SARS-CoV-2 community transmission. Here, we used the collected data to characterize SARS-CoV-2 transmission in different contexts.

The probability that a contact with an infected individual led to a transmission event varied by context. Contacts in the domestic context and in the recreational context were associated with transmission probabilities of 30% and 13%, respectively; many contexts, however, were associated with lower probabilities in the 3% – 6.5% range (e.g. workplaces, schools, residential care facilities). With respect to the absolute number of transmission events, most detected putative SARS-CoV-2 transmission events took place in or around households. Household-associated transmission accounted for 89% of links between cases in the constructed hypothetical transmission chains, and the importance of household-based transmission was supported independently by forward and backward contact tracing and address-based proxy indicators of joint household membership. Compared to the household context, the absolute contribution of other types of infection contexts to overall case load was relatively low. This also included the combined effect of institutional outbreaks, which were associated with only 5% of overall cases.

What are the implications with respect to the efficacy of infection prevention measures? The high proportion of cases associated with transmission in the domestic context – in which effective infection prevention, e.g. by isolation or using masks, was not always easy to implement and impossible to enforce by public health authorities – implicitly attested to the high effectiveness of infection prevention in other contexts. Consistent with this, we also observed high context-specific transmission probabilities for the household context. Lower transmission probabilities for certain contexts, in which infection prevention was institutionally enforced, such as schools or workplaces, suggested the efficiency of infection prevention measures in these. For hospitals and care homes for elderly people, the increased transmission probabilities observed for contacts in these contexts (15% and 10%, respectively) suggested that hygiene and social distancing measures within these institutions were perhaps not strictly enforced or adhered to by patients or residents, many of which also exhibited increased individual vulnerability to SARS-CoV-2. Specifically for care homes, the combination of high outbreak case numbers and low numbers of linked community cases was notable and suggested that the rules designed to prevent viral importation (e.g., strict access requirements) may also have limited viral export to the community.

Our data remained ambiguous with respect to the controversial role of schools during the SARS-CoV-2 pandemic [Lordan et al. 2021, Amodio et al. 2022, von Bismarck-Osten et al. 2022, Neil-Sztramko et al. 2024]. On the one hand, based on the epidemiological relationships captured by contact tracing, school outbreaks and their linked community cases only accounted for a small fraction of overall cases (1.6%). On the other hand, our sequencing-based search for additional outbreak-related cases showed that classical contact tracing did not capture all relevant epidemiological relationships and that the number of cases linked to school outbreaks was therefore, in fact, underestimated. While the number of additional cases linked to school outbreaks detected by sequencing was small in absolute terms, its contribution to the overall number of sequenced outbreak-associated cases was substantial (18% increase), and, at a case sequencing coverage of 24.5%, we almost certainly missed relevant cases. Our analysis thus showed that genomics-enhanced contact tracing was informative about the role of schools during the SARS-CoV-2 pandemic and that a more quantification of the impact of schools on overall SARS-CoV-2 transmission would likely have been possible, but required a higher case sequencing rate.

With respect to the implementation of contact tracing, our study showed, first, the accuracy and relevance of classical contact tracing. The large majority of epidemiological links captured by forward or backward contact tracing likely corresponded to true transmission events and the outbreak documentation carried out by Düsseldorf Health Authority was generally correct. Second, name and address data were sufficient to identify additional transmission-related epidemiological relationships

in domestic or residential contexts not captured by contact tracing interviews. Indicators of joint household membership, which were defined based on data provided by diagnostic laboratories and which did not depend on case interviews, provided an important independent data source on SARS-CoV-2 transmission in the population. Third, classical contact tracing likely missed an important fraction of relevant epidemiological connections, as evident from the fact that 55.6% registered SARS-CoV-2 cases remained without putative infection source based on all epidemiological connections represented in the Integrated Case Graph even when ignoring genetic constraints. Fourth, as shown here for school outbreaks and previously for nightlife-associated transmission chains Walker et al. [2022], sequencing could improve the performance of contact tracing by enabling a targeted search for epidemiological connections between genetically linked cases. Enabling the evaluation of known epidemiological connections as well as informing the discovery of unknown links, the role of pathogen genome sequencing in the context of genomics-enhanced contact tracing is therefore two-fold.

What proportion of sequenced cases should implementations of genomics-enhanced contact tracing aim for and what proportion of genetic links should be followed up with an epidemiological investigation? Simulation and downsampling experiments showed no saturation of different indicators of the ability to characterize SARS-CoV-2 transmission towards the case sequencing proportion of 24.5% achieved in our study. Based on the complete dataset, we found that the proportion of sequenced cases with a putative transmission source may increase to 43.9% if each genetic link with distance 0 corresponded to an epidemiological connection and if each such link could be investigated. While the true number of epidemiological links that could be discovered is almost certainly smaller, there is likely substantial potential for the sequencing-led discovery of additional epidemiological connections, as illustrated here for the in-depth analysis of school outbreaks. Genomics-enhanced contact tracing may significantly contribute to the understanding of pathogen transmission in areas not easily captured by classical contact tracing; we therefore recommend aiming at as high a proportion of cases for sequencing and epidemiological follow-up investigation as possible. While potentially costly, falling sequencing costs and the increasing availability of artificial intelligence models for the collection and analysis of unstructured data may facilitate future implementations of genomics-enhanced contact tracing. By enabling a comprehensive characterization of context-specific transmission probabilities and a reliable quantification of the contribution of different contexts to overall case load, genomics-enhanced contact tracing may provide the necessary information to navigate difficult societal trade-offs that arise in pandemic situations and enable an in-action as well as after-action evaluation of non-pharmaceutical interventions.

In conclusion, we demonstrated the comprehensive characterization of SARS-CoV-2 transmission based on genomics-enhanced contact tracing. The ability to characterize infection contexts, institutional outbreaks, and transmission chains in an integrated manner could enable the implementation of more targeted infection prevention measures. Our study informs future implementations and applications of genomics-enhanced contact tracing.

Limitations of this study: A link between two cases in the constructed hypothetical transmission chains does not imply an actual transmission event; in particular when computed independent of genetic data, the proportion of cases with a potential infection source should be interpreted as an upper bound on the ability to explain individual cases with the available epidemiological data. Edges of type “sameAddress” are imperfect indicators of joint household membership and edges of type “sameAddress-differentName”, in particular, may also sometimes capture transmission in e.g. residential facilities or large building complexes (Supplementary Figure 14 B, Supplementary Figure 11). Non-Düsseldorf contacts could not be analyzed in our study; however, non-Düsseldorf contacts only accounted for 19% of overall contacts and their improved analysis would, given the limited role

of forward contact tracing-induced edges in the computed hypothetical transmission chains, likely not have led to a significant increase in the proportion of cases with a potential infection source or different conclusions with respect to infection contexts. The case sequencing coverage in our study was only 24.5%; increasing the proportion of sequenced cases while keeping the available contact tracing information constant might have led to an increase in CERs and in the proportion of cases with a potential infection source, but not beyond the 44.4% reported above for the complete Integrated Case Graph. Our study was based on data collected over a period of 11 months in 2021 and our results may not generalize to viral variants of SARS-CoV-2 not covered by the study period, which may exhibit different transmission properties. Our analysis did not incorporate a potential effect of vaccination on SARS-CoV-2 transmission. Finally, the generalizability of the results to other respiratory pathogens beyond SARS-CoV-2 is an important direction for future research.

## Supplementary Figure Legends

**Supplementary Figure 1: SARS-CoV-2 sequencing turnaround times.** Distribution of SARS-CoV-2 sequencing turnaround times, measured from SARS-CoV-2 sample arrival at Heinrich Heine University to first availability of a resolved SARS-CoV-2 genome with <3000 unresolved (i.e. "N") characters on the dashboard of the Integrated Genomic Surveillance System Düsseldorf, tracked for a subset of n = 891 samples that were collected between April and June 2022. Samples that were not successfully sequenced to the desired level of SARS-CoV-2 genome completeness were not included.

**Supplementary Figure 2: Data integration approach.** Shown are the different steps of data integration leading to the Integrated Case Graph. Case, contact tracing and outbreak data were exported from the SurvNet database hosted at Düsseldorf Health Department and transferred into an intermediate MS Access application. SARS-CoV-2 samples were labeled with PCR test IDs and sampling dates; based on these, the first and last names and dates of births of the tested individuals could be obtained and merged with Düsseldorf Health Department case registration data. All case- or sample-identifying data (incl. names, dates of birth, sampling dates, address data) are hypothetical. In a final step, if a case was sequenced more than once, a unique sequence was selected. See Methods for details.

**Supplementary Figure 3: SARS-CoV-2 variant distribution over time.** Pango lineages [Rambaut et al. 2020] of the sequenced SARS-CoV-2 genomes, stratified by week.

**Supplementary Figure 4: Forward and backward contact tracing. A.** Histogram over the number of registered forward contact tracing records per case. **B.** Histogram over the number of registered forward contact tracing records per case, limited to contacts registered (residential address) in the city Düsseldorf. **C.** Histogram over the number of "forward contact tracing"-type edges per case in the Integrated Case Graph. **D.** Histogram over the number of registered incoming or outgoing forward contact tracing links per case in the directionalized ($t_d = 3$; see Methods) version of the Integrated Case Graph. **E.** Histograms over the number of "backward contact tracing"-type edges per case in the directionalized version of the Integrated Case Graph, stratified by direction of the edge.

**Supplementary Figure 5: Forward contact tracing records over time. A**. Shown is the monthly total number of registered SARS-CoV-2 cases registered and the proportion of cases with at least one attached forward contact tracing record (independent of whether the contact was mapped onto another case or not). **B.** Distribution over the number of forward contact tracing records per case, limited to cases with at least one attached contact tracing record and stratified by month of the year. **C.** Same data as in panel B., with a truncated y-axis focusing on the region between 1 and 10 forward contact tracing records.

**Supplementary Figure 6: Distribution of genetic distances for contact tracing and "sameAddress"-type edges.** Distribution of SARS-CoV-2 genetic distances for case pairs linked by different types of edges in the graph, stratified by edge type and only shown for case pairs for which a genetic distance could be calculated.

**Supplementary Figure 7: Context-specific distribution of genetic distances for contact tracing edges.** Distribution of SARS-CoV-2 genetic distances for case pairs linked by forward (panel A) and backward (panel B) contact tracing edges in the graph, stratified by edge type and infection contexts, and only shown for case pairs for which a genetic distance could be calculated.

**Supplementary Figure 8: Case registration and symptom onset interval distributions for epidemiologically connected sequenced cases.** Shown are differences in case registration (right-hand side plots) and symptom onset (left-hand side plots) dates for pairs of epidemiologically connected sequenced case pairs, stratified by edge type in the Integrated Case Graph and by whether the connected cases exhibited a genetic distance of ≤1 ("accepted") or >1 ("rejected"). Panels **A.** and **B.**: Forward contact tracing, absolute date differences between case registration (panel A) and symptom onset (panel B) dates. Panels **C.** and **D.**: Backward contact tracing, date difference along the assumed directionality of infection as recorded during backward contact tracing (if two cases specified each other as assumed infection sources, two date differences per panel are included). Panels **E.** and **F.**: "sameAddress"-type edges (combining "sameAddress-sameName" and "sameAddress-differentName"), absolute date differences. Panels **G.** and **H.**: "sameAddress"-type edges, further stratified by whether the connected cases are also linked by a forward contact tracing edge; absolute date differences in case registration dates (panel G) and symptom onset dates (panel H). Panels **I.** and **J.**: Accepted "sameAddress"-type edges, further stratified by whether the connected cases are also linked by a contact tracing edge and by "sameAddress" edge subtype ("same name" and "not same name"); absolute date differences in case registration dates (panel I) and symptom onset dates (panel J).

**Supplementary Figure 9: Matched control dataset with scrambled contact tracing links. A.** Difference between case registration dates for pairs of sequenced cases connected by real forward or backward contact tracing links and scrambled contact tracing edges. **B.** Number of contacts per case, comparing cases that are part of pairs of sequenced cases connected by real or scrambled contact tracing edges. **C.** Vaccination status of cases, comparing cases that are part of pairs of sequenced cases connected by real or scrambled contact tracing edges. **D.** Genetic differences between cases, comparing cases that are part of pairs of sequenced cases connected by real or scrambled contact tracing edges.

**Supplementary Figure 10: Backward contact tracing infection contexts and genetic evaluation.** Genetic validation of infection contexts captured by backward contact tracing; shown is the number of case pairs that were connected by a backward contact tracing edge assigned to the corresponding putative context, stratified by whether the genetic distance genetic distance between the connected cases was ≤1 ("accepted") or >1 ("rejected"); edges for which no genetic distance was available were not included.

**Supplementary Figure 11: Overlap of "sameAddress"-type edges in the Integrated Case Graph with other edge types. A. Overlap with contact tracing edges.** Shown is the overlap between edges of type "sameAddress", "forward contact tracing", and "backward contact tracing" in the Integrated Case Graph. **B. Overlap with "outbreak"-type edges.** For both panels, two edges of different types were counted as overlapping if they connected the same two cases. The label "Same Address - same Name" refers to edges of type "sameAddress-sameName"; the label "Same address", to edges of type "sameAddress-differentName".

**Supplementary Figure 12: Genetic distances along "sameAddress"-type edges.** Comparison of genetic distances along the two types of "sameAddress"- type edges, shown for pairs of sequenced cases connected by the specified edge type ("sameAddress-sameName": orange bars, labelled as "SelbeAdresseNachname"; "sameAddress-differentName": green bars; labelled as "SameAddressDifferentName").

**Supplementary Figure 13: Domestic transmission of SARS-CoV-2 and reduced case interval distributions for "sameAddress"-type edges.** Illustration of how domestic transmission of SARS-CoV-2 can lead to low case intervals for cases connected by "sameAddress"-type edges. Left-hand side of the figure: Hypothetical transmission scenario; Case 1 transmits SARS-CoV-2 to three members of the same household (Cases 2, 3, and 4) over a short period of time. Cases 2, 3, and 4 will likely experience symptom onset at approximately the same time. Right-hand side of the figure: Representation of the four cases in the Integrated Case Graph; because the cases are members of the same household, the four cases are fully connected by pairwise "sameAddress"-type edges. The edges from Case 1 to Cases 2, 3 and 4 reflect true underlying transmission events and the observed

case intervals will likely reflect typical SARS-CoV-2 case intervals; by contrast, the edges between Cases 2, 3 and 4 will likely, while also exhibiting high genetic similarities, show reduced case intervals. In the example, 3 of the 6 considered edges reflect a shared infection context of the connected cases rather than direct transmission.

**Supplementary Figure 14: Size of "sameAddress"-type connected components. A.** Distribution of the size of connected components in the Integrated Case Graph when only considering edges of type "sameAddress-sameName". **B.** Distribution of the size of connected components in the Integrated Case Graph when only considering edges of type "sameAddress-differentName".

**Supplementary Figure 15: Integrated outbreak analysis, additional metrics. A.** Outbreak sizes, stratified by outbreak type; outbreak size was defined as the number of cases assigned to the outbreak by Düsseldorf Health Authority. **B.** Outbreak durations, stratified by outbreak type; the duration of an outbreak was defined as the time in days from the earliest to the latest case registration date of the cases assigned to the outbreak. **C.** Forward contact tracing records per outbreak (defined as the cumulative number of forward contact tracing records of all outbreak-assigned cases, independent of whether the contacts were mapped onto other cases or not), stratified by outbreak type. **D.** "Forward contact tracing"-type edges connecting outbreak cases to non-outbreak cases in the Integrated Case Graph, per outbreak and stratified by outbreak type. **E.** Like panel D but normalized by outbreak size. **F.** "Forward contact tracing"-type edges connecting outbreak cases to non-outbreak cases in the Integrated Case Graph, per outbreak and stratified by outbreak type and by whether the connected non-outbreak case had a genetic distance of ≤1 to the outbreak ("accepted"), whether it had a genetic distance ≥2 ("rejected"), or whether a genetic distance could not be computed ("No distance available"). The genetic distance between a linked non-outbreak case and an outbreak was defined as the minimum genetic distance between the linked case and any case of the corresponding outbreak, and a distance could be computed if the outbreak contained at least one sequenced case and if the linked case was sequenced. **G.** Like panel F but normalized by outbreak size. **H.** "sameAddress"-type edges connecting outbreak cases to non-outbreak cases in the Integrated Case Graph, per outbreak and stratified by outbreak type. **I.** Like panel H but normalized by outbreak size. **J.** Like panel F but for "sameAddress"-type edges. **K.** Like panel G but for "sameAddress"-type edges. **L.** Number of genetically linked community cases, per outbreak and by outbreak type; genetically linked community cases were defined as cases not linked to the analyzed outbreak that had a genetic distance of 0 to at least one case of the corresponding outbreak. **M.** Like panel L but normalized by outbreak size.

**Supplementary Figure 16: Manual investigation of outbreaks.** Shown is an outbreak comprising 6 cases, 4 of which are sequenced and 2 of which are not. 2 cases were likely erroneously assigned to the outbreak (shaded in red), as indicated by PCR test dates in one case and genetic distances to other outbreak cases in the other case. Two cases are linked to the outbreak via contact tracing; one of these is, based on belonging to the same institutional context as the other outbreak cases, classified as a "missing" outbreak case (shaded in dark blue), and the other case is classified as an "outbreak-linked" case (shaded in light blue); the association between the cases and the outbreaks is supported by sequencing data in both instances. Three additional cases not assigned to the outbreak exhibit a genetic distance of 0 to at least one outbreak case. The first of the three cases is, based on belonging to the same institutional context as the other outbreak cases, classified as a "missing" outbreak case (shaded in dark green). For the second of the three cases, an epidemiological connection to the outbreak that is not present in the contact tracing data is discovered and the case is classified as outbreak-linked based on the discovered connection (shaded in medium-intensity green). Finally, for the third case (shaded in medium-intensity green), an epidemiological connection to the outbreak that involves a non-sequenced additional case (shaded light green) is discovered. The third case and the non-sequenced additional case are therefore classified as outbreak-linked; in the case of the non-sequenced additional case, its classification as outbreak-linked is based on the fact that it represents the only detected connection that can explain the genetic identity between the third sample and the outbreak.

**Supplementary Figure 17: CER genetic cluster sizes. A.** Distributions of the size of genetic clusters computed for determining the CER, shown for genetic clustering thresholds $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). **B.** Mean sizes of genetic clusters computed for determining the CER, by month of the year and shown for genetic clustering thresholds $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). Clusters were assigned to months based on the average case registration date of the included cases. Error bars show 95% confidence intervals. Orange

bars in the background show the total number of SARS-CoV-2 registered in Düsseldorf in the corresponding month.

**Supplementary Figure 18: Contribution of different edge types to hypothetical transmission chains. A.** Number of edges contained in the hypothetical transmission chains constructed for determining the CER, by edge type and for $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). **B.** Sensitivity of CER to the removal of edges of different types; shown is the CER obtained when removing all edges of the specified type from Integrated Case Graph, for $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). All plots based on $t_d = -3$ and $t_j = 0$.

**Supplementary Figure 19: CER parameters and CER over time. A.** Distribution of per-cluster CER values, shown for genetic clustering thresholds $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). **B.** CER by cluster size, for genetic clustering thresholds $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). Shown is the average of per-cluster CER values for all clusters of the specified size in cases, with error bars indicating 95% confidence intervals. **C.** CER by month, for genetic clustering thresholds $t_g = 0$ (left subpanel) and $t_g = 1$ (right subpanel). Shown is the average of per-cluster CER values for all clusters assigned to the specified month, with error bars indicating 95% confidence intervals; clusters were assigned to months based on the average case registration date of the included cases. All plots in all panels based on $t_d = -3$ and $t_j = 0$.

**Supplementary Figure 20: Subsampling of proportion of sequenced cases. A.** Distribution of cluster sizes for different proportions of sequenced cases (x-axis) and $t_g = 0$ (left-hand panel) and $t_g = 1$ (right-hand panel). **B.** Absolute number of cases with a putative transmission source in the computed hypothetical transmission chains, for different proportions of sequenced cases (x-axis) and based on clusters computed for $t_g = 0$ (left-hand panel) and $t_g = 1$ (right-hand panel). **C.** Cluster Explanation Rate (CER), for different proportions of sequenced cases (x-axis) and for $t_g = 0$ (left-hand panel) and $t_g = 1$ (right-hand panel). **D.** Proportion of cases with a putative infection source based on the assumption that all genetic links correspond to an epidemiolocal link, for different proportions of sequenced cases (x-axis) and $t_g = 0$ (left-hand panel) and $t_g = 1$ (right-hand panel).

**Supplementary Figure 21: SARS-CoV-2 sample selection and assignment. A.** Pairwise genetic distances between SARS-CoV-2 genomes assigned to the same case, prior to the selection of at most one viral genome sequencing record for all cases. The red line shows the mean of the distribution. **B.** Scatterplot comparing, for each generated SARS-CoV-2 genome after quality control and before outlier removal, the sampling date of the diagnostic swab used for SARS-CoV-2 sequencing (Y axis) to the Düsseldorf Health Authority registration date of the case that the viral genome sequencing record was assigned to (see Methods). The diagonal (difference = 0) is shown as a black line; red lines show the variance. Before proceeding to the selection of at most one unique viral genome per case, SARS-CoV-2 sequencing records which exhibited a difference between case registration and sample collection dates of $< -21$ or $> 7$ days were removed (see Methods).

**Supplementary Figure 22: SARS-CoV-2 example transmission network**. Transmission network of 8 cases, comprising 5 sequenced cases, 2 cases which were registered but not sequenced, and 1 undetected case. Red arrows show transmissions in the (generally unknown) complete transmission network; blue connections show 4 case connection paths ($w_1$ – $w_4$) between sequenced cases in the complete network, which connect the sequence pairs most closely related in terms of the unknown full network. A hypothetical transmission chain for the 5 sequenced cases may imply a putative infection source for up to 80% (4 / 5) of cases if "jump edges" (see Methods) are included and if case connection path $w_4$ can be found in the known epidemiological connections.

**Supplementary Figure 23: Symptom onset and case registration dates. A.** Difference between the case registration and symptom onset dates of individual cases. Cases with missing symptom onset data were excluded. **B.** Symptom onset date differences for pairs of sequenced cases with genetic distance ≤1 and connected by edges of type "backward contact tracing"; shown is the difference in symptom onset dates along the assumed directionality of infection as recorded during backward contact tracing (i.e., if case x specified case y as their putative infection source, the plot shows the symptom onset of x minus the symptom onset of y; if two cases specified each other as assumed infection sources, two date differences are included). All case pairs comprising cases with missing symptom onset data were excluded. **C.** Transmission directionality probability distribution, based on a simple Bayesian model. Assume there are two cases a and b; either a infected b or b

infected a, and define the observed data $x$ the difference in symptom onset date between a and b, i.e. $x = \text{symptom_onset}(a) - \text{symptom_onset}(b)$. The probability that b infected a conditional on the observed data is $P(b \to a|\text{x}) = \frac{\Pr(b \to a) \times \text{P}(\text{x}|b \to a)}{\Pr(b \to a) \times \text{P}(\text{x}|b \to a) + \Pr(a \to b) \times \text{P}(\text{x}|a \to b)}$. Assuming uniform priors, we obtain $P(b \to a|\text{x}) = \frac{\text{P}(\text{x}|b \to a)}{\text{P}(\text{x}|b \to a) + \text{P}(\text{x}|a \to b)}$. We empirically estimate a probability distribution $P_{Backward}$ on the difference in symptom onset dates between the infected case and the infecting case observed in backward contact tracing; specifically, $P_{Backward}(d) = \frac{|(c_1, c_2) \in E'_B : \text{symptom_onset}(c_2) - \text{symptom_onset}(c_1)|}{|E'_B|}$, where $E'_B$ is the set of edges of type "backward contact tracing" in the Integrated Case Graph. We define $\text{P}(\text{x}|b \to a) = P_{Backward}(x)$. With respect to $\text{P}(\text{x}|a \to b)$, we note that x is, conditional on the assumption "$a \to b$" that a infected b, the difference in symptom onset dates between the infecting case and the infected case, i.e., the opposite of the required parameterization for $P_{Backward}(d)$. To obtain the difference in symptom onset dates between the infected case and the infected case under the assumption that a infected b, x can be multiplied by -1; we thus obtain $\text{P}(\text{x}|b \to a) = P_{Backward}(-x)$. This gives $P(b \to a|\text{x}) = \frac{P_{Backward}(x)}{P_{Backward}(x) + P_{Backward}(-x)}$.

**Supplementary Figure 24: Graph directionalization parameter $t_d$.** The plots show, for two different values (0 and 3) of the "maximum symptom onset difference" directionalization parameter $t_d$ ("Directionalization cut-off"), the number of edges in the Integrated Case Graph that were converted into one ("Directed edges") or two directed edges ("Undirected edges") in the directionalized version of the Integrated Case Graph, stratified by edge type (panel **A**) or by the genetic distance between the connected cases (panel **B**; only for case pairs consisting of two sequenced cases). Directionalization captures the concept of directionality of transmission and ensures that the symptom onset dates of cases connected in hypothetical transmission chains are compatible with the directionality of hypothetical transmission. In the process of directionalization, undirected edges of the form $(c_1, c_2)$ are converted into either a single directed edge of the form $(c_1, c_2)$ or $(c_2, c_1)$, or into a pair of directed edges of the form $(c_1, c_2)$ and $(c_2, c_1)$. Setting $t_d = 0$ implies that all connections (with the exception of links induced by backward contact tracing, see below) between two cases in the Integrated Case Graph become "directed" (i.e., are represented by a single directed edge) whenever there is any difference in symptom onset dates, and that all considered hypothetical transmissions always follow the direction of increasing symptom onset. The directionalization of edges of type "backward contact tracing" is carried out based on the individual cases' assessment of their assumed infection source; a "backward contact tracing" edge from case $c_2$ to case $c_1$ in the directionalized graph is created whenever $c_1$ specified $c_2$ as their assumed infection source, independent of symptom onset dates (see Supplementary Figure 8 for the distribution of differences in symptom onset dates between sequenced cases linked by "backward contact tracing" edges).

# Supplementary Table Legends

**Supplementary Table 1: Sequenced SARS-CoV-2 samples.** The table includes, for each sequenced SARS-CoV-2 sample, a pseudonymized identifier ("Sample_ID"); the sample collection date; a binary indicator indicating whether the sequence was attached to a case node in the Integrated Case Graph; the number of undefined bases (Ns); GISAID, ENA and COGDAT upload identifiers; sample-to-sequence turnaround time (from arrival of sample at Heinrich Heine University to availability of the first viral genome sequence with <3000 "N" characters on the Integrated Genomic Surveillance System Düsseldorf website; not available for all samples); and Pango lineage classification results.

**Supplementary Table 2: Total and sequenced cases per week.** Total registered and sequences cases as well as case sequencing rate, by week.

**Supplementary Table 3: Forward contact tracing contexts.** For each forward contact tracing context, the table shows information on the number of recorded contacts, underlying index cases (the index case is the individual with whom the contact tracing interview was carried out), putative transmission events (i.e. edges in the Integrated Case Graph), putative transmission events with genetic distances available, (i.e. edges with a genetic distance), putative transmission events that were genetically validated (i.e. edges classified as "accepted"), and the corresponding putative transmission and genetic validation rates. For column "Number of putative transmission events", instances in which cases A and B specified each other as contacts during forward contact

tracing contributed a count of 2; for column "Number of putative transmission events (unique edges in integrated graph)", such instances only contributed a count of 1 (as, prior to directionalization, edges in the Integrated Case Graph are undirected).

**Supplementary Table 4: Outbreak analysis results for individual outbreaks**. With the exception of one additional column specifying the month in which the earliest case of each outbreak was registered, all included columns are also found in Supplementary Table 5 and the definitions given in Supplementary Table 10 apply (with the modification that no agglomeration by outbreak type is carried out for any of the metrics; in this table, each outbreak is – when applying the definitions given in Supplementary Table 10 – effectively treated as a separate category).

**Supplementary Table 5: Outbreak analysis results agglomerated by outbreak type.** See Supplementary Table 10 for a detailed legend of the included fields.

**Supplementary Table 6: Identified links between the outbreaks.** The table shows, for all pairs of outbreaks, whether a link between the two outbreaks was identified. For each linked pair of outbreaks, the table specifies the number of epidemiological links between the two outbreaks (i.e. of edges connecting any case from the one outbreak to any case of the other outbreak in the Integrated Case Graph) and which types of edges these were induced by ("Forward", "Backward," "SameAddress"); the number of genetically linked cases (i.e. the number of unordered pairs of sequenced cases from both outbreaks with a genetic distance of 0; "Genetically linked"); and whether the link was confirmed manually ("Manual analysis", encoded as a binary true/false variable).

**Supplementary Table 7: Genetic clusters computed for determining the CER.** List of genetic clusters. For each cluster, the table contains the ID of the cluster; the size of the cluster in cases; the time in days ("length") between the earliest and the latest case registration dates of the cases assigned to the cluster; the mean of the case registration dates of the cases assigned to the clusters; and the cluster-specific CER values, computed for the Cartesian product of $t_j \in \{0,1,2\}$ and $t_d \in \{0,1\}$. The prefix of the cluster ID indicates whether the specified cluster is part of the cluster set computed using a genetic similarity threshold $t_g$ of 0 or 1.

**Supplementary Table 8: CER results.** The table shows the results of the CER computation for the complete directionalized case graph $G'$, i.e. summing over all genetic clusters. Shown are CER results for all combinations of parameter values controlling the CER computation process and for all permutations of edge types (indicated by an "X" in the corresponding edge type columns; i.e. the absence of an "X" indicates that the corresponding edge type was ignored when computing the CER). "Direction cutoff": Maximum symptom onset difference" parameter ($t_d$, "direction cutoff") multiplied by -1, determining which differences in dates of symptom onset were allowed between immediately linked cases in the constructed hypothetical transmission chains. "Clusters max distance": genetic clustering parameter ($t_g$). "Jump distance": "jump distance" parameter ($t_j$) that controled how many intermediate non-sequenced cases ("jump edges") any link between two cases in the constructed hypothetical transmission chains might comprise. "TrueSelbeAdresse": Edges of type "sameAddress-differentName". "SelbeAdresseNachname": Edges of type "SameAddressAndLastname". "SurvnetAngestecktBei": Backward contact tracing edges. "Outbreak": Edges of type "outbreak". "SurvnetKontaktperson": Forward contact tracing edges. "Jump edges count": Number of "jump edges" that represented intermediate non-sequenced cases available based on the specified parameterization (independently of whether these were used in the computed hypothetical transmission chains). "Used edges": Total number of edges in the hypothetical transmission chains computed for the specified parameterization. The four last columns show the number of "jump edges" used in the hypothetical transmission chains for the specified parameterization and the proportion of these edges relative to the total number of edges in these chains, stratified by how many intermediary non-sequenced cases the utilized edges comprised.

**Supplementary Table 9: Hypothetical transmission chains context analysis.** The table shows the number of edges contained in the computed hypothetical transmission chains, stratified by edge type and (when available) further transmission context or outbreak type information. The columns "Forward" and "Backward" contain the number of forward and backward contact tracing edges for which no specific context was available. All computations based on $t_j = t_d = 0$. The row labeled "On cluster" shows results for hypothetical transmission chains that were computed incorporating genetic distances with $t_g = 0$; the row labeled "On entire graph

ignoring sequences" shows results for hypothetical transmission chains that were computed based purely on epidemiological connections, i.e. ignoring genetic constraints.

**Supplementary Table 10: Outbreak analysis results column legend.** Definition of fields in Supplementary Table 4. The given definitions also apply to Supplementary Table 5, with the modification that no agglomeration by outbreak type is carried out in Supplementary Table 5; i.e. in table Supplementary Table 5, each outbreak is – according to the definitions given for Supplementary Table 4 – effectively treated as its own category.

**Supplementary Table 11: Controlled forward contact tracing context vocabulary.** Vocabulary terms used for recording forward contact tracing infection contexts, in their German original and English translation.

# Methods

## Implementation of genomics-enhanced contact tracing in Düsseldorf

Genomics-enhanced contact tracing was implemented using the Integrated Genomic Surveillance System Düsseldorf (IGSD), which has been fully described elsewhere [Houwaart et al. 2022, Walker et al. 2022]. Briefly, the high-level aims of the IGSD were to monitor viral evolution in the Düsseldorf area and to investigate whether a combination of high-density genomic surveillance with classical epidemiological approaches could enable an improved understanding of SARS-CoV-2 transmission in the population at large. The IGSD was implemented as a collaboration between Heinrich Heine University Düsseldorf (HHU), Zotz | Klimas (ZK) and Medizinische Laboratorien Düsseldorf (MLD), two large commercial diagnostic laboratories from Düsseldorf, and Düsseldorf Health Authority. Viral samples provided by ZK and MLD, as well as viral samples collected from patients treated at Düsseldorf University Hospital, were rapidly Nanopore-sequenced using HHU's SARS-CoV-2 sequencing platform; SARS-CoV-2 sequences generated by MLD were also integrated.

## Case registration and contact tracing carried out by Düsseldorf Health Authority

SARS-CoV-2 case data collection and contact tracing, using the SurvNet application [Faensen et al. 2006, Krause et al. 2007] for data storage and management, were carried out by Düsseldorf Health Authority under the mandate of the Law on the Prevention of Infection (Infektionsschutzgesetz (IfSG) based on the classification of SARS-CoV-2 as a notifiable pathogen in Germany since 2020. This involved (i) the reporting to Düsseldorf Health Authority of all Düsseldorf residents tested positive for SARS-CoV-2 by diagnostic laboratories; (ii) the registration of the personal data of all registered cases by Düsseldorf Health Authority; (iii) contact tracing interviews with all Düsseldorf cases conducted by Düsseldorf Health Authority, the participation in which was mandatory; (iv) the reporting of all suspected outbreaks (see "Retrospective investigation of outbreaks" for the applied definition of an outbreak) and their involved cases to Düsseldorf Health Authority. Contact tracing was based on the official contact tracing guidelines in place at the respective time, as published by the Robert-Koch-Institute. It always involved a forward contact tracing component and also typically a backward contact tracing component. Forward contact tracing was carried out for quarantine and isolation purposes, aiming at the identification of all individuals that a case had been in contact with around a time window of 2 days before symptoms onset (if symptomatic) or 2 days before the case's positive diagnostic test (if asymptomatic), irrespective of the potential directionality of transmission. Personal data of unvaccinated contacts, and of vaccinated contacts registered up until 20 April 2021, were always entered into SurvNet in a structured manner (and thus represented as edges in the Integrated Case Graph, see below); on 21 April 2021, mandatory quarantine requirements were lifted for fully vaccinated contacts, and whether and in which form (structured or using free-text entry fields) personal data of fully vaccinated contacts were entered into SurvNet after that date was decided by the individual carrying out the contact tracing interview. Of note, the number of registered contacts per case remained relatively similar over the course of 2021, except for November and December

2021 (Supplementary Figure 5). During backward contact tracing, the infected individual was asked whether they could identify whom they thought they had likely acquired their infection from. Backward contact tracing data were usually entered into SurvNet in a structured manner. During contact tracing interviews, responses by the interviewed individuals were assessed for plausibility and likelihood that the contact led to a transmission; contacts that were deemed unlikely to have led to a transmission (e.g., between two masked individuals in a well-ventilated area) were not recorded. Contact tracing also involved an assessment of the context in which the detected contacts had taken place in; for structured forward contact tracing records, context information was stored in a structured manner as part of the contact tracing records using a controlled vocabulary (Supplementary Table 11); otherwise, context information was sometimes stored in free-text entry fields.

## Genetic distance matrix

A matrix of genetic distances between all analyzed sequences was calculated as previously defined [Houwaart et al. 2022]. Briefly, based on a multiple sequence alignment (MSA) of the included sequences computed using MAFFT [Katoh and Standley 2013] following GISAID [Shu and McCauley 2017] instructions, the distance $d(x, y)$ between samples x and y was defined as the count of differences between the entries of $x$ and $y$ in the MSA, (i) ignoring leading or trailing "gap" characters, (ii) accounting for International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes, (iii) treating subsequent non-matching gaps columns as a single difference, (iv) ignoring deletions in the MSA entry of x aligned to 'N' characters in y and vice versa, and (v) not counting any differences found between the beginning of the MSA and the 20th A/C/G/T character of either sequence and the end of the MSA and the 20 last A/C/G/T characters of either sequence.

## Integrating genomic and viral genome sequencing data into a case graph

Integration of genomic, case, and contact tracing data was carried out using a custom Microsoft Access database application deployed within the protected IT infrastructure of Düsseldorf Health Authority. All personally identifiable information remained at Düsseldorf Health Authority. See Supplementary Figure 2 for a schematic of the data integration process.

As a first step, case and contact tracing data for SARS-CoV-2 cases from Düsseldorf were exported from the SurvNet application and imported into the custom Access database. Routinely collected data included (a) case-identifying information, such as name, date of birth and registered address; (b) forward contact tracing data; forward contact tracing data comprised the SurvNet database ID of the corresponding case (i.e. specifying the individual with whom the contact tracing interview was carried out), as well as the first and last names, the date of birth, and the residential status (Düsseldorf resident or not) of the registered contacts; the exported forward contact tracing data records, however, did not comprise information on the subsequent infection or case registration status of the contacts (see next section); (c) backward contact tracing data ("InfectedBy" field), only available for a small proportion of cases; backward contact tracing records comprised the SurvNet case ID of the individual with whom the contact tracing interview was carried out (the infected individual) as well as the SurvNet case ID of the individual who was registered as the putative source of the infection (i.e. both individuals were necessarily registered as cases); (d) outbreak data, specifying, for each registered outbreak, details on the specific institutional context as well as a list of case IDs assigned to the outbreak.

As a second step, the following post-processing steps were carried out. (i) The total number of registered forward contact tracing records on a per-case basis was tabulated and attached to each case record, separately for all registered contacts and registered contacts from Düsseldorf. (ii) For each registered forward contact tracing contact from Düsseldorf, the case database was queried for

cases with matching first and last names and a matching date of birth; for each match, a case-resolved forward contact tracing record, comprising the case IDs of both the interviewed individual and of the registered contact, as well as the database ID of the contact tracing record, was created and stored in the database (with additional filters on the maximum difference between the registration dates of any two cases so-connected applied later, see below; these date-based filtering steps also resolved instances in which a contact tracing record matched more than one case record, e.g. because the same individual tested SARS-CoV-2-positive more than once); (iii) For each case ("query case"), other cases with case registration dates within a 30-day window around the registration date of the query case (≤ 30 days absolute date difference) and (a) with the same registered address as the query case or (b) sharing the same registered address and the same last name as the query case were identified and stored as "SameAddress" and "SameAddressAndLastname" records, comprising the case IDs of the linked cases, in the database. (iv) For each outbreak, the set of all pairs of cases assigned to the outbreak was determined and each pair was stored as a record of type "SameOutbreak", comprising the case IDs of the two member cases.

As a third step, SARS-CoV-2 genome sequencing records were matched with cases, based on the following procedure. (i) All sequenced samples, and therefore all sequenced SARS-CoV-2 genomes, were labelled with a sample collection date and a PCR test ID by the diagnostic laboratory providing the sample. (ii) Based on these data, we queried data sheets provided by the diagnostic laboratories for the first and last name and the date of birth of the individuals that had provided the sequenced sample. (iii) We matched these data against the SurvNet case registration records; if exactly one match covering all 3 criteria was identified, the viral genome sequence was attached to the SurvNet case record; otherwise, manual curation was carried out to account e.g. for last name spelling mistakes in one of the two data sources. (iv) For all viral genome sequences attached to case records, the difference between case registration date and sample collection date was calculated, and all n = 36 viral genome sequences with a difference < -21 or > 7 were removed (Supplementary Figure 21)

As a fourth step, for all n = 533 cases $c$ with more than one assigned viral genome sequencing record, a unique record was selected based on the following procedure. (i) If all sequences attached to a case had genetic distance ≤1 between each other, we selected the sequence with the smallest number of undefined nucleotides ("N" characters). (ii) Otherwise, define the set $C_{Contacts-Sequenced}$ as the set of other cases that were linked to case $c$ via contact tracing and that had attached viral genome sequencing data (there we no instances of cases in $C_{Contacts-Sequenced}$ that had more than one attached viral genome sequencing record). (iii) For each viral genome sequencing record $s$ attached to $c$ (more than one by definition), define $compatible(c, s)$ as the set of cases in $C_{Contacts-Sequenced}$ that had an associated viral genome sequencing record with genetic distance $\leq 1$ to $s$ (see Supplementary Figure 21). (iv) Determine $\arg\max_{s} compatible(c, s)$ and attach the corresponding viral genome sequencing record $s$ to $c$.

Finally, the integrated dataset was exported as an Integrated Case Graph in which (a) each registered case from Düsseldorf, together with non-identifiable metadata such as the number of total forward contact tracing contacts and forward contact tracing contacts from Düsseldorf, was represented as a node; (b) connections between pairs of cases, i.e. forward and backward contact tracing relationships, "sameAddress"-type information, as well assignment of two cases to the same outbreak, were represented as undirected edges; and (c) SARS-CoV-2 genome sequencing records of cases were attached to nodes in the case graph. Edges of type "sameAddress-sameName" were created based on records of type "SameAddressAndLastname"; edges of type "sameAddress-differentName" were created for all sample pairs that appeared in "SameAddress"-type records, but not in "SameAddressAndLastname"-type records. All edges $(c_1, c_2)$ of type "backward contact tracing"

between cases $c_1$ and $c_2$ were annotated with the information on whether $c_1$ specified $c_2$, and $c_2$, $c_1$, as their assumed infection sources (note that these two cases were not mutually exclusive). All edges between pairs of cases with an absolute difference in case registration dates of more than 14 days were removed [Ryu et al. 2022]. A distance matrix for all SARS-CoV-2 genomes attached to the nodes of the graph was exported with the graph.

## Infection context analysis

To characterize context-specific transmission probabilities, we carried out a separate database export of all forward contact tracing records stored in SurvNet, comprising, for all exported records, the SurvNet database ID of the index case (i.e. specifying the individual with whom the contact tracing interview was carried out); the residential status (Düsseldorf resident or not) of the contact; the database ID of the contact tracing record; and the contact context (determined by Düsseldorf Health Authority staff during contact tracing; see Supplementary Table 11). We filtered the exported records for index cases with a case registration date $\leq$ 17 December 2021 and represented in the Integrated Case Graph and removed all records with contacts not registered at a Düsseldorf address. A recorded contact was classified as associated with a potential transmission event if and only if (i) a matching case-resolved forward contact tracing record (see previous section; matching was carried out based on the case ID of the index case and the database ID of the contact tracing record) was identified and (ii) the identified case-resolved forward contact tracing record was represented by an edge of type "forward contact tracing" in the Integrated Case Graph after application of all filtering steps; in all instances satisfying the two conditions, the individual that the contact was registered with had also tested SARS-CoV-2 positive in a period of 14 days around the registration date of the index case, compatible with a direct transmission event. Context-specific transmission probabilities were then calculated by dividing the number of contacts associated with a potential transmission event by the total number of registered contacts for each context.

For backward contact tracing records, export of infection context information from SurvNet in a structured manner was not possible. We therefore employed a manual classification approach, based on review of the data stored in SurvNet (involving all structured case data as well as free-text notes taken during routine contact tracing interview, if available). Each backward contact tracing contact was, based on the definitions given below, assigned to one of the following categories: (i) “Private household”: Contact between two individuals in the context of a private household, involving at least one overnight stay; (ii) “Recreational activity”: Contact associated with a recreational activity, such as taking a walk, attending a family party, visiting a friend’s place (without an associated overnight stay), or contact during participation in a shared hobby (e.g. football); (iii) “School”: Contact in the context of a primary or secondary school, including contacts between students and teachers; (iv) “Kindergarten/Daycare”: Contact in the context of a kindergarten or daycare nursery (German “KiTa”), including contacts between children and daycare personnel; (v) “Workplace”: The involved individuals interacted in a work context; (vi) “Asylum/homeless care home”: The individuals had overlapping stays at a care home for asylum seekers or homeless individuals; (vii) “Elderly care home”: Contact in the context of a care home for the elderly, including contacts between care home personnel and inhabitants; (viii) “Hospital”: Contact in the context of a hospital, including contacts between patients and staff; (ix) “Assisted living care home”: Contact in the context of a care home for assisted living for non-elderly individuals (German “betreutes Wohnen”), including contacts between inhabitants and staff.

## Generation of a scrambled-edges control dataset for analysis of contact tracing genetic distances

To investigate whether genetic distances observed for case pairs connected by contact tracing edges may be the result of incidental genetic similarity of cases sequenced at approximately the same time, we generated a matched control dataset with randomized connections between sequenced cases registered at approximately the same time; the genetic distances along the randomized edges could then be compared to the genetic distances observed for real contact tracing edges. To generate the randomized dataset, we applied the following procedure. We identified all 3,862 cases that were sequenced and that either had a forward or backward contact tracing edge in the Integrated Case Graph. We created candidate randomized case pairs by first generating all 14,915,044 possible ordered case pairs from the identified 3,862 cases and removed all pairs that exhibited a registration date difference of more than 14 days. Furthermore, we removed all pairs that were connected by real forward or backward contact tracing links in the Integrated Case Graph, as well as all pairs linking a case to itself and case pairs $(a, b)$ for which forward or backward contact tracing edges $(a, x)$ and $(b, x)$ – where $x$ denotes a joint shared contact of $a$ and $b$ – existed. From the set of remaining 1,616,999 candidate randomized case pairs, we selected the 1,011 pairs that exhibited the highest degree of similarity (see Supplementary Figure 9) to the 1,011 real case pairs connected by edges of type "forward contact tracing" or "backward contact tracing" (and with a genetic distance available) in the Integrated Case Graph, taking into account i) vaccination status, ii) date of index case registration, iii) mean case registration date, iv) distance between the case registration dates, v) number of registered contacts. The matching procedure was implemented using a custom R script and R packages "lubridate", "stringr" and "data.table".

## Retrospective investigation of outbreaks

For all outbreaks recorded by Düsseldorf Health Authority, routinely collected outbreak data, including the type of institution, the involved cases and the duration of the outbreak, were extracted from the database of Düsseldorf Heath Department and combined with the data represented in the Integrated Case Graph for analysis. An outbreak was defined as "non-clonal" if a graph defined by (i) the sequenced outbreak cases as nodes and (ii) inter-case genetic distances as edges contained, after removal of edges corresponding to genetic distances > 1, more than one connected component; an individual case was classified as erroneously assigned to the outbreak if it was not part of the largest connected component of the graph so-defined (i.e., all clonal outbreaks contained no erroneously assigned cases according to this definition). A community case (here defined as a case not part of the any other outbreak of the same category as the outbreak under investigation) was defined as connected to an outbreak if it was connected to any case of the considered outbreak in the Integrated Case Graph; for outbreaks with ≥ 1 sequenced cases and all outbreak-linked sequenced community cases, the corresponding links were classified as "sequencing-confirmed" if the genetic distance between the linked community case and any sequenced outbreak case was $\leq 1$. For outbreaks with ≥ 1 sequenced case, we also identified community cases that were only linked to the outbreak on a genetic basis, i.e. independent of epidemiological information; for this analysis, genetically linked community cases were defined as community cases that exhibited a genetic distance $\leq 1$, and that had a case registration date difference to any outbreak case of $\leq 14$ days. Two outbreaks A and B were classified as epidemiologically linked if any case of outbreak A was linked to any case of outbreak B in the Integrated Case Graph; two outbreaks A and B were classified as genetically linked if any sequenced case of outbreak A had a genetic distance $\leq 1$ to any sequenced case of outbreak B.

For school outbreaks, an in-depth analysis was carried out (Supplementary Figure 16), comprising the following steps and based on the definition of an outbreak as a group of cases connected by suspected

or confirmed transmission events in a shared defined (typically institutional) context. First, epidemiological identification of cases falsely assigned to the outbreak: An evaluation of all available test and symptom onset dates for all outbreak cases was carried out and cases for which the available test and symptom onset data were not compatible with the other outbreak cases were identified. Second, identification of missing cases based on contact tracing and genetic data: For each outbreak with ≥ 1 sequenced case, we identified all cases that (i) were linked to at least one outbreak case by structured forward contact tracing data; (ii) were not already assigned to the outbreak; (iii) exhibited a genetic distance of 0 to at least one case in the outbreak; (iv) belonged, based on the available case data, to the same (typically institutional) context as the outbreak (e.g., were students or teachers at the school at which the outbreak had taken place). Third, analysis of non-structured and backward contact tracing data: For a subset of outbreaks (identified by column "Analysis of non-structured contact information" in table Supplementary Table 4), contact tracing information entered in a non-structured manner into the SurvNet application (i.e., using free-text entry fields) as well as backward contact tracing data were also analyzed during the previous step. Fourth, genetically driven identification of missing or linked cases: For a subset of outbreaks with ≥ 1 sequenced case (identified by column "Outbreak-external links investigated based on individual case genetic distance checks" in table Supplementary Table 4), we identified cases that (i) were not already linked to the outbreak by either direct assignment to the outbreak or via any contact tracing data; (ii) exhibited a genetic distance of ≤ 1 and a PCR test date difference of ≤ 4 weeks to at least one outbreak case; (iii) met one of the following three criteria of linkage to the outbreak: "missing outbreak case", defined as belonging to the (institutional) context of the outbreak; "direct epidemiological linkage", defined as a direct epidemiological link (e.g. shared household, family connection, workplace) to at least one outbreak case; "indirect epidemiological linkage", defined as a direct epidemiological link to another case X with a direct epidemiological link to at least one outbreak case. For connections of the latter (indirect epidemiological) type, we furthermore required that "X" was either non-sequenced or that its viral genome was compatible (distance ≤ 1) to the outbreak; if "X" was not sequenced and found to represent the only connection between a sequenced case classified as "indirectly epidemiologically linked" and the outbreak, "X" was classified as an outbreak-linked case with indirect genetic support and further classified into a missing outbreak case or a linked community sample. Fifth, assessment of potential outbreak-associated transmission chains: We carried out a comprehensive manual assessment and analysis of potential infection chains associated with the investigated outbreaks. This led, for example, to the conclusion that the only sequenced case of outbreak CVD-DÜS-SCHULE-2021-0182 was likely a mis-assigned (see Supplementary Note 1). Sixth, identification of related outbreaks: Two outbreaks were classified as related if there was, according to the criteria defined above, at least one case from the one outbreak that exhibited direct or indirect genetic linkage to at least one case from the other outbreak and if the connection between the two outbreaks was also supported by a direct or indirect epidemiological connection between at least one case from the one outbreak to at least one case from the other outbreak.

## Cluster Explanation Rate

Given a single cluster C of N genetically related cases, and a set of epidemiological connections between them, the Cluster Explanation Rate (CER) measures the proportion of cases in the cluster that a putative infection source can be identified for based on a hypothetical transmission chain constructed from the epidemiological connections for the cases in the cluster. When generalized to multiple clusters, the CER is computed by summing over the number of cases with a putative infection source and the total number of cases across the considered clusters. Finally, at the level of a complete dataset, the CER is determined by first clustering all sequenced cases by genetic similarity and then

computing the CER of the clusters so-determined; at this level, the CER provides a measure of how well the collected epidemiological data explain overall observed genetic similarities.

Formally, let $G' = (V, E')$ be a directionalized and augmented version of the case graph $G = (V, E)$; directionalization and augmentation are required for ensuring that hypothetical transmission chains considered for computing the CER are consistent with respect to symptom and/or case registration dates and for including hypothetical transmission chains that involve intermediate non-sequenced cases (see below). All CER computations are based on $G'$. We first define the CER of a single cluster $C$, i.e. of a subset of nodes of the case graph $G'$, of genetically related cases; let $t_g$ denote the genetic distance threshold that was used for the clustering (defined in detail below). Let the set $E'(C)$ represent the set of genetically supported epidemiological connections between the cases in $C$; $E'(C)$ is defined as the subset of directed edges between $C$ in $G'$ that connect two cases with a pairwise genetic distance $\leq t_g$. A hypothetical transmission chain $h$ is defined as an acyclic subset of $E'(C)$ such that no node $c \in C$ has more than one incoming edge from the set $h$ (note that the definition allows for partial transmission chains that do not connect all nodes; epidemiologically, such transmission chains may be interpreted to include cases outside of the cluster $C$; Supplementary Figure 22; in addition, no constraints are placed on the number of outgoing edges, i.e. the set of considered transmission chains includes "branching" transmission chains or "networks"). The CER for cluster $C$ in case graph $G'$ is then defined as

$$\mathrm{CER}_{G'}(C) := \frac{\max\limits_{h \in H(C)} |h|}{|C|},$$

, where $H(C)$ is the set of all possible hypothetical transmission chains for the cases in cluster $C$ and their epidemiological connections $E'(C)$; $|h|$ is the size of the hypothetical transmission chain $h$ in edges; $|C|$ is the number of cases in $C$; and $\max\limits_{h \in H} |h|$ is computed according to the heuristic algorithm defined below.

For a set $W$ of multiple clusters of genetically related cases in case graph $G'$, we define the CER as

$$\mathrm{CER}_{G'}(W) := \frac{\sum_{C \in W}\left(\max\limits_{h \in H(C)} |h|\right)}{\sum_{C \in W}(|C|)}.$$

Finally, to define the CER for a case graph $G'$, the sequenced cases in $G'$ are clustered into groups of genetically related isolates, based on a similarity threshold $t_g$. To do so, we consider a graph that consists of the sequenced cases in $G'$ as nodes and that includes all pairwise genetic distances as undirected edges; we remove all edges corresponding to genetic distances $> t_g$; and, finally, we define the set $\mathrm{genetic_clusters}(G')$ as the set of remaining connected components with size > 1 in the graph. Any two cases $c_1$ and $c_2$ from the same cluster either have genetic similarity $\leq t_g$ or they are connected by a path of sequenced cases with intermediate pairwise distances between $\leq t_g$; furthermore, the computed clusters are non-overlapping. Based on the computed clusters, we define the CER for a case graph $G'$ as

$$\mathrm{CER}(G') = \mathrm{CER}_{G'}\big(\mathrm{genetic_clusters}(G')\big).$$

Directionalization, yielding a directed case graph $G' = (V, E')$, enables the definition of hypothetical transmission chains at the level of a directed graph between cases, capturing the concept of directionality of transmission and ensuring that all considered pairwise links between cases are consistent with respect to symptom onset (or, if unavailable, case registration) dates. In $G'$, a directed edge $(c_1, c_2)$ represents a possible hypothetical transmission event from $c_1$ to $c_2$. Directionalization is followed by potential augmentation of the graph with "jump edges", representing links in hypothetical

transmission chains between sequenced cases that involve intermediate non-sequenced cases. For directionalization, $E'$ is constructed by including, for each (undirected) edge $e = (c_1, c_2) \in E$, one or two directed edges, as defined below. For all edges except for edges of type "backward contact tracing", directionalization is controlled by the parameter $t_d$, specifying the allowed difference in symptom onset dates in days between linked cases (according to the definition below). Denote the date of symptom onset of case $c$ as $s(c)$; if the symptom onset date for case $c$ is unavailable, $s(c)$ is set to the case registration date of $c$ minus 3 days, based on an empirical analysis of the difference between symptom onset and case registration dates for samples for which both dates are available (Supplementary Figure 23). We distinguish between the following cases: (i) if $\mathrm{abs}(s(c_1) - s(c_2)) \leq t_d$, edges $(c_1, c_2)$ and $(c_2, c_1)$ are added to $E'$; (ii) if $s(c_1) - s(c_2) > t_d$, edge $(c_2, c_1)$ is added to $E'$; (iii) otherwise, $s(c_2) - s(c_1) > t_d$ is true, and edge $(c_1, c_2)$ is added to $E'$. Note that these definitions imply that one undirected edge in the original case graph may be represented by two directed edges in the directed case graph, if the observed symptom onset dates are consistent with transmission in both directions, depending on the threshold $t_d$. For undirected edges $(c_1, c_2)$ of type "backward contact tracing" in the original case graph $G$, a directed edge from $c_1$ to $c_2$ is added to $E'$ if case $c_2$ specified case $c_1$ as their assumed infection source during infection source during backward contact tracing; and a directed edge from $c_2$ to $c_1$ is added to $E'$ if case $c_1$ specified case $c_2$ as their assumed infection source during backward contact tracing. For augmentation with "jump distance" parameter $t_j = 0$, no additional edges are added to $G'$. For augmentation with $t_j = 1$, an edge from $c_1$ to $c_2$ is added to $E'$ for all instances of the form $(c_1, a, c_2)$, in which (i) $c_1$ and $c_2$ represent two sequenced cases; (ii) $a$ represents a non-sequenced case; (iii) there are no edges directly connecting $c_1$ to $c_2$; (iv) there is at least one edge from $c_1$ to $a$; (v) there is at least one edge from $a$ to $c_2$; (v) the genetic distance between $c_1$ to $c_2$ is $\leq (t_j + 1) * t_g$. Augmentation for $t_j = 2$ proceeds analogously for instances of the form $(c_1, a, b, c_2)$.

The computed $\mathrm{CER}(G')$ depends on the parameter used for the genetic clustering ($t_g$), the "maximum symptom onset difference" parameter used for directionalization ($t_d$), and the "jump distance" parameter ($t_j$). To assess the impact of different parameter choices on the computed CER, all CER computations are carried out for all combinations (i.e., the outer product) of the parameter values $t_g = \{0,1\}$, $t_d = \{0,3\}$ and $t_j = \{0,1,2\}$. For $t_d$, a value of 0 implies constraining the direction of hypothetical transmission whenever there is any difference in symptom onset dates (i.e., a conservative choice); $-3$ was chosen empirically based on inspection of observed symptom onset dates for sample pairs connected by edges of type "backward contact tracing", which are intrinsically directed (Supplementary Figure 23). Directionalization results for different edge types and the two employed thresholds are visualized in Supplementary Figure 24.

$\max_{h \in H(C)} |h|$ for a cluster $C$ is computed according to the following procedure. Initialize the variable $h'$ as the empty set; $h'$ is populated progressively with edges from $E'(C)$, and, after termination, the set $h'$ represents a heuristic approximation of $\arg\max_{h \in H(C)} |h|$.

The following steps are carried out until termination. First, define $M$ as the set of nodes $c \in C$ that are neither the origin nor the destination of any of the edges currently in $h'$. Second, for each $c \in M$, define $\mathrm{reachable}(c)$ as the set of other nodes in $M$ that can be reached from $c$, employing paths that consist of the directed edges $E'(C)$ and that only involve intermediate nodes that are also part of $M$. $\mathrm{reachable}(c)$ can be computed using a depth-first search. Third, find $c_{max} = \arg\max_{c \in M} |\mathrm{reachable}(c)|$ by iterating through the current members of $M$. If $|\mathrm{reachable}(c_{max})| = 0$, terminate; otherwise, let $B$ be the set of edges underlying the paths connecting $c_{max}$ to $\mathrm{reachable}(c)$. If necessary, remove edges from $B$ so that $B$ does not contain any cycles and so that there is no node

in $\mathrm{reachable}(c)$ with more than one incoming edge from $B$. Finally, add the edges in $B$ to $h'$ and go back to the first step.

## Ethics statement

This study was approved by the ethics committee of the Medical Faculty of Heinrich Heine University Düsseldorf (#2020–839).


## Funding

We acknowledge funding by the Ministry for Work, Health and Social Affairs of the State of North Rhine-Westphalia ("24.04.01 Gen. u. IKA" and CPS-1-1A); the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; Netzwerk Universitätsmedizin, GenSurv/MolTraX), award number 01KX2021; by the Jürgen Manchot Foundation; and by the Deutsche Forschungsgemeinschaft (DFG) award 428994620.


## No competing interests

All authors have completed the ICMJE uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.


## Acknowledgements

We thank Gil McVean and Marc Janssen for helpful discussions. We also thank the contact tracing team of Düsseldorf Health Authority for their commitment and the countless extra hours.


## References


Amodio, E., M. Battisti, A. Kourtellos, G. Maggio and C. M. Maida (2022). "Schools opening and Covid-19 diffusion: Evidence from geolocalized microdata." Eur Econ Rev **143**: 104003.

Andersen, K. G., B. J. Shapiro, C. B. Matranga, R. Sealfon, A. E. Lin, L. M. Moses, O. A. Folarin, A. Goba, I. Odia, P. E. Ehiane, M. Momoh, E. M. England, S. Winnicki, L. M. Branco, S. K. Gire, E. Phelan, R. Tariyal, R. Tewhey, O. Omoniwa, M. Fullah, R. Fonnie, M. Fonnie, L. Kanneh, S. Jalloh, M. Gbakie, S. Saffa, K. Karbo, A. D. Gladden, J. Qu, M. Stremlau, M. Nekoui, H. K. Finucane, S. Tabrizi, J. J. Vitti, B. Birren, M. Fitzgerald, C. McCowan, A. Ireland, A. M. Berlin, J. Bochicchio, B. Tazon-Vega, N. J. Lennon, E. M. Ryan, Z. Bjornson, D. A. Milner, Jr., A. K. Lukens, N. Broodie, M. Rowland, M. Heinrich, M. Akdag, J. S. Schieffelin, D. Levy, H. Akpan, D. G. Bausch, K. Rubins, J. B. McCormick, E. S. Lander, S. Gunther, L. Hensley, S. Okogbenin, C. Viral Hemorrhagic Fever, S. F. Schaffner, P. O. Okokhere, S. H. Khan, D. S. Grant, G. O. Akpede, D. A. Asogun, A. Gnirke, J. Z. Levin, C. T. Happi, R. F. Garry and P. C. Sabeti (2015). "Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus." Cell **162**(4): 738-750.

Bludau, A., A. Jack, N. Fischer, J. Dreesman, C. Drosten, R. Egelkamp, L. Ehlkes, F. Feil, A. Grundhoff, H. Grundmann, P. Kreuzer, M. Monazahian, I. Overesch, D. Schmitt, M. Troger, A. von Reiswitz, J. Weber, A. Dilthey, C. Hornberg, S. Reuter and S. Scheithauer (2025). "Use of integrated genomic surveillance by local public health authorities: Recommendations based on a mixed-methods study of current adoption, applications and success factors, Germany, 2023." Euro Surveill **30**(13).

COVID-19 Genomics UK (COG-UK) (2020). "An integrated national scale SARS-CoV-2 genomic surveillance network." Lancet Microbe **1**(3): e99-e100.

Faensen, D., H. Claus, J. Benzler, A. Ammon, T. Pfoch, T. Breuer and G. Krause (2006). "SurvNet@RKI--a multistate electronic reporting system for communicable diseases." Euro Surveill **11**(4): 100-103.

Faria, N. R., E. C. Sabino, M. R. Nunes, L. C. Alcantara, N. J. Loman and O. G. Pybus (2016). "Mobile real-time surveillance of Zika virus in Brazil." Genome Med **8**(1): 97.
Harris, S. R., E. J. Cartwright, M. E. Torok, M. T. Holden, N. M. Brown, A. L. Ogilvy-Stuart, M. J. Ellington, M. A. Quail, S. D. Bentley, J. Parkhill and S. J. Peacock (2013). "Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study." Lancet Infect Dis **13**(2): 130-136.
Hjorleifsson, K. E., S. Rognvaldsson, H. Jonsson, A. B. Agustsdottir, M. Andresdottir, K. Birgisdottir, O. Eiriksson, E. S. Eythorsson, R. Fridriksdottir, G. Georgsson, K. R. Gudmundsson, A. Gylfason, G. Haraldsdottir, B. O. Jensson, A. Jonasdotti, A. Jonasdottir, K. S. Josefsdottir, N. Kristinsdottir, B. Kristjansdottir, T. Kristjansson, D. N. Magnusdottir, R. Palsson, L. le Roux, G. M. Sigurbergsdottir, A. Sigurdsson, M. I. Sigurdsson, G. Sveinbjornsson, E. A. Thorarensen, B. Thorbjornsson, M. Thordardottir, A. Helgason, H. Holm, I. Jonsdottir, F. Jonsson, O. T. Magnusson, G. Masson, G. L. Norddahl, J. Saemundsdottir, P. Sulem, U. Thorsteinsdottir, D. F. Gudbjartsson, P. Melsted and K. Stefansson (2022). "Reconstruction of a large-scale outbreak of SARS-CoV-2 infection in Iceland informs vaccination strategies." Clin Microbiol Infect **28**(6): 852-858.
Houwaart, T., S. Belhaj, E. Tawalbeh, D. Nagels, Y. Frohlich, P. Finzer, P. Ciruela, A. Sabria, M. Herrero, C. Andres, A. Anton, A. Benmoumene, D. Asskali, H. Haidar, J. von Dahlen, J. Nicolai, M. Stiller, J. Blum, C. Lange, C. Adelmann, B. Schroer, U. Osmers, C. Grice, P. P. Kirfel, H. Jomaa, D. Strelow, L. Hulse, M. Pigulla, P. Kreuzer, A. Tyshaieva, J. Weber, T. Wienemann, M. Kohns Vasconcelos, K. Hoffmann, N. Lubke, S. Hauka, M. Andree, C. J. Scholz, N. Jazmati, K. Gobels, R. Zotz, K. Pfeffer, J. Timm, L. Ehlkes, A. Walker, A. T. Dilthey, C.-O. I. German and C.-O. I. German (2022). "Integrated genomic surveillance enables tracing of person-to-person SARS-CoV-2 transmission chains during community transmission and reveals extensive onward transmission of travel-imported infections, Germany, June to July 2021." Euro Surveill **27**(43).
Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.
Krause, G., D. Altmann, D. Faensen, K. Porten, J. Benzler, T. Pfoch, A. Ammon, M. H. Kramer and H. Claus (2007). "SurvNet electronic surveillance system for infectious disease outbreaks, Germany." Emerg Infect Dis **13**(10): 1548-1555.
Lane, C. R., N. L. Sherry, A. F. Porter, S. Duchene, K. Horan, P. Andersson, M. Wilmot, A. Turner, S. Dougall, S. A. Johnson, M. Sait, A. Goncalves da Silva, S. A. Ballard, T. Hoang, T. P. Stinear, L. Caly, V. Sintchenko, R. Graham, J. McMahon, D. Smith, L. E. Leong, E. M. Meumann, L. Cooley, B. Schwessinger, W. Rawlinson, S. J. van Hal, N. Stephens, M. Catton, C. Looker, S. Crouch, B. Sutton, C. Alpren, D. A. Williamson, T. Seemann and B. P. Howden (2021). "Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study." Lancet Public Health **6**(8): e547-e556.
Lordan, R., S. Prior, E. Hennessy, A. Naik, S. Ghosh, G. K. Paschos, C. Skarke, K. Barekat, T. Hollingsworth, S. Juska, L. L. Mazaleuskaya, S. Teegarden, A. L. Glascock, S. Anderson, H. Meng, S. Y. Tang, A. Weljie, L. Bottalico, E. Ricciotti, P. Cherfane, A. Mrcela, G. Grant, K. Poole, N. Mayer, M. Waring, L. Adang, J. Becker, S. Fries, G. A. FitzGerald and T. Grosser (2021). "Considerations for the Safe Operation of Schools During the Coronavirus Pandemic." Front Public Health **9**: 751451.
Mate, S. E., J. R. Kugelman, T. G. Nyenswah, J. T. Ladner, M. R. Wiley, T. Cordier-Lassalle, A. Christie, G. P. Schroth, S. M. Gross, G. J. Davies-Wayne, S. A. Shinde, R. Murugan, S. B. Sieh, M. Badio, L. Fakoli, F. Taweh, E. de Wit, N. van Doremalen, V. J. Munster, J. Pettitt, K. Prieto, B. W. Humrighouse, U. Stroher, J. W. DiClaro, L. E. Hensley, R. J. Schoepp, D. Safronetz, J. Fair, J. H. Kuhn, D. J. Blackley, A. S. Laney, D. E. Williams, T. Lo, A. Gasasira, S. T. Nichol, P. Formenty, F. N. Kateh, K. M. De Cock, F. Bolay, M. Sanchez-Lockhart and G. Palacios (2015). "Molecular Evidence of Sexual Transmission of Ebola Virus." N Engl J Med **373**(25): 2448-2454.
McCrone, J. T., V. Hill, S. Bajaj, R. E. Pena, B. C. Lambert, R. Inward, S. Bhatt, E. Volz, C. Ruis, S. Dellicour, G. Baele, A. E. Zarebski, A. Sadilek, N. Wu, A. Schneider, X. Ji, J. Raghwani, B. Jackson, R. Colquhoun, A. O'Toole, T. P. Peacock, K. Twohig, S. Thelwall, G. Dabrera, R. Myers, C.-G. U.

Consortium, N. R. Faria, C. Huber, Bogoch, II, K. Khan, L. du Plessis, J. C. Barrett, D. M. Aanensen, W. S. Barclay, M. Chand, T. Connor, N. J. Loman, M. A. Suchard, O. G. Pybus, A. Rambaut and M. U. G. Kraemer (2022). "Context-specific emergence and growth of the SARS-CoV-2 Delta variant." Nature **610**(7930): 154-160.
Mellmann, A., S. Bletz, T. Boking, F. Kipp, K. Becker, A. Schultes, K. Prior and D. Harmsen (2016). "Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting." J Clin Microbiol **54**(12): 2874-2881.
Neil-Sztramko, S. E., E. Belita, R. L. Traynor, L. Hagerman, S. Akaraci, P. Burnett, A. Kostopoulos and M. Dobbins (2024). "What is the specific role of schools and daycares in COVID-19 transmission? A final report from a living rapid review." Lancet Child Adolesc Health **8**(4): 290-300.
Pullano, G., E. Valdano, N. Scarpa, S. Rubrichi and V. Colizza (2020). "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study." Lancet Digit Health **2**(12): e638-e649.
Rambaut, A., E. C. Holmes, A. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis and O. G. Pybus (2020). "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology." Nat Microbiol **5**(11): 1403-1407.
Ryu, S., D. Kim, J. S. Lim, S. T. Ali and B. J. Cowling (2022). "Serial Interval and Transmission Dynamics during SARS-CoV-2 Delta Variant Predominance, South Korea." Emerg Infect Dis **28**(2): 407-410.
Seemann, T., C. R. Lane, N. L. Sherry, S. Duchene, A. Goncalves da Silva, L. Caly, M. Sait, S. A. Ballard, K. Horan, M. B. Schultz, T. Hoang, M. Easton, S. Dougall, T. P. Stinear, J. Druce, M. Catton, B. Sutton, A. van Diemen, C. Alpren, D. A. Williamson and B. P. Howden (2020). "Tracking the COVID-19 pandemic in Australia using genomics." Nat Commun **11**(1): 4376.
Shu, Y. and J. McCauley (2017). "GISAID: Global initiative on sharing all influenza data - from vision to reality." Euro Surveill **22**(13).
Sivertsen, A., N. Mortensen, U. Solem, E. Valen, M. F. Bullita, K. A. Wensaas, S. Litleskare, G. Rortveit, H. M. S. Grewal and E. Ulvestad (2024). "Comprehensive contact tracing during an outbreak of alpha-variant SARS-CoV-2 in a rural community reveals less viral genomic diversity and higher household secondary attack rates than expected." mSphere **9**(8): e0011424.
von Bismarck-Osten, C., K. Borusyak and U. Schönberg (2022). "The role of schools in transmission of the SARS-CoV-2 virus: quasi-experimental evidence from Germany." Economic Policy **37**(109): 87-130.
Walker, A., T. Houwaart, P. Finzer, L. Ehlkes, A. Tyshaieva, M. Damagnez, D. Strelow, A. Duplessis, J. Nicolai, T. Wienemann, T. Tamayo, M. Kohns Vasconcelos, L. Hulse, K. Hoffmann, N. Lubke, S. Hauka, M. Andree, M. P. Daumer, A. Thielen, S. Kolbe-Busch, K. Gobels, R. Zotz, K. Pfeffer, J. Timm, A. T. Dilthey and C.-O. I. German (2022). "Characterization of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection Clusters Based on Integrated Genomic Surveillance, Outbreak Analysis and Contact Tracing in an Urban Setting." Clin Infect Dis **74**(6): 1039-1046.