Title:

Verification of emerging genomic mutations in *Mycobacterium tuberculosis* allows transmission chains to be distinguished in an epidemiological typing cluster extending over thirty years.

Short Title:

Emerging mutations in *M. tuberculosis* to support epidemiological investigations.

Authors:

Rina de Zwaan[1*], Gerard de Vries[3*], Ella Ubbelohde[1,2], Arnout Mulder[1], Miranda Kamst[1], Karin Rebel[4], Saskia Kautz[5], Kristin Kremer[2] , Richard M, Anthony[1] , Dick van Soolingen[1]

Affiliations:

[1] Tuberculosis Reference Laboratory, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

[2] KNCV Tuberculosis Foundation, The Hague, The Netherlands.

[3] Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

[4] TB Department, Municipal Public Health Services (GGD), (Utrecht)

[5] TB Department, Municipal Public Health Services (GGD), Rotterdam

* joint first author contributed equally

Abstract:

Whole genome sequencing (WGS) is able to identify epidemiological links between *Mycobacterium tuberculosis* isolates. Recent clustering can be ruled out using a pre-defined single nucleotide polymorphism (SNP) threshold. If WGS clusters grow significantly over time limited genetic variability hampers epidemiological investigations.

Newly emerging (informative) SNPs in isolates of an extended cluster growing for more than 30 years to >150 cases in the Netherlands were analysed. WGS data was analyzed from 61 sequencing files from 54 patients. Genomic positions that varied within the cluster isolates were carefully screened for minority populations in other isolates from the cluster. A transmission scheme was generated on the basis of WGS data alone then compared to the epidemiological information available.

Fifty-two informative SNPs were identified, eight of which were also detected as mixed variants. One emerging SNP in *dnaA* (1199G>A R400H) has been observed in other transmitted strains and may be under selection. There was high concordance between the transmission chains suggested on basis of the newly emerging SNPs and scenario's identified using classical epidemiological cluster investigations.

Analysis of filtered SNPs accumulating in the genome of *M. tuberculosis* in large clusters contains information on transmission dynamics and can be used to support epidemiological investigations.

Introduction

In the Netherlands, all *Mycobacterium tuberculosis* complex isolates have been subjected to epidemiological typing by whole genome sequencing (WGS) at the national tuberculosis reference laboratory since 2016 [1]. Isolates within a 12 single nucleotide polymorphism (SNP) distance are reported as 'clustered' to the Municipal Public Health Services (MPHs) who perform interviews to identify and confirm epidemiological links [2]. However, most international WGS analysis pipeline's nowadays apply a stricter threshold of six SNPs to rule in a possible recent transmission [3-4]. Furthermore, SNP filtering in our pipeline is less strict than in most other pipelines, which results in an increased sensitivity to genetic variation and a parallel reduction in specificity (overcalling of SNPs). We estimate the increased SNP calling in our pipeline to be between 0 and 3 SNPs based on observations in international quality control exercises [1, 5].

In low transmission settings such as the Netherlands epidemiological clusters mostly consist of two to three cases, but occasionally large, longer-term clusters occur. In general, the detectable genetic variability between *M. tuberculosis* isolates in even large transmission clusters is limited [6-8]. Moreover, SNP variability between isolates from a single individual can be as large as between isolates from different individuals in a cluster [5, 9] Nonetheless, as the population structure of *M. tuberculosis* is clonal, in principle a higher resolution can be achieved by tracking truly emerging SNPs, as non-fixed alleles in a part of the isolates. This approach should enable a more precise reconstruction of transmission chains [10-13]. This was already suspected in 2010 by investigating single colony cultures of *M. tuberculosis* isolates associated with the large Harlingen outbreak in the Netherlands; the occurrence of new mutations in a part of the colonies of a sub set of the isolates was proven by PCR and

63 could be used to distinguish different sources of spread in the extending cluster [10]. Thus if
64 SNP calling is sufficiently accurate, a higher resolution in typing can be achieved by following
65 the accumulation of specific emerging SNPs and tracking them as fixed or non-fixed alleles in
66 isolates within transmission chains.

67 Here, we investigate WGS data of the second largest typing cluster in the Netherlands that
68 has been followed by three different molecular typing methods (initially IS*6110* RFLP, then
69 VNTR-MIRU 24, then WGS) over a time period of 30 years [2], with the objective to study the
70 accumulation of new SNPs and to test their utility to distinguish transmission forks introduced
71 by subsequent sources of spread. The predicted evolutionary branching of the strain was
72 compared to information available at MHSs on possible sources of spread in the cluster and
73 subsequent cases, in order to investigate the utility of the cluster sub-branching information
74 in the epidemiological investigation of MHSs.
75
76

77 **Methods**

78 Since 1993, DNA fingerprinting of all *M. tuberculosis* isolates has been performed in the
79 Netherlands, using RFLP typing from 1993-2008, VNTR-MIRU from 2004-2018 and WGS
80 from 2016 to the present. TB public health nurses systematically investigate possible
81 epidemiological links between clustered cases. This activity, the so called 'cluster
82 investigation', is prioritised for recent transmission defined as cases that occur within two
83 years after another case. A link is confirmed if the investigation reveals that the newly
84 clustered case was indeed in contact with another clustered and infectious case (e.g. a
85 family member, friend) or if the person was at a particular location at the same time with a

86 clustered infectious case. Since 2009, results of cluster investigations are reported to the
87 National TB Register (NTR), including the identified source case.
88 A single large cluster, that existed throughout the whole epidemiological typing period since
89 1993 and with a substantial number of confirmed epidemiological links between clustered
90 cases, was selected for detailed study. We requested the NTR Registration Committee to
91 provide demographic data of cases in this cluster and the results of cluster investigations. All
92 isolates in this cluster had been typed with WGS since 2016. Additionally, patients from
93 before 2016 with a confirmed epidemiological link were selected for re-typing by WGS.
94 MPHs provided information from routine cluster investigation on the most likely place of
95 transmission of patients in the cluster with a confirmed epidemiological link. The laboratory
96 technicians were blinded to information on possible epidemiological links between the
97 selected isolates.
98 From the cluster, 61 WGS sequencing files of *M. tuberculosis* isolates from 54 individuals
99 were analysed. From four individuals the sequences from two different isolates were
100 included; two individuals had two isolates (<6 months apart) and two others had two
101 isolates as a result of a reactivation (>6 months apart). An additional three isolates were
102 included twice as duplicate controls for the WGS sequencing and data analysis procedure.
103
104 The isolates were sequenced using an Illumina (CA, USA) HiSeq™ 2500, NextSeq™ 500 or
105 NextSeq™ 550 sequencer. Unpaired reads were mapped against the H37Rv reference
106 genome using Bowtie2, (GenBank accession: AL123456.3). Breseq v0.28.1 was used to
107 detect SNPs and INDELs (insertions and deletions). The in-house pipeline called SNPs using a
108 ≥80% allele frequency cut off [1]. Called SNPs were excluded from the SNP distance
109 calculation when present in 'difficult to sequence' genomic regions according to an

110 exclusion list or when they were within 12 bp from another SNP [3]. All sequencing data
111 were subjected to our routine quality check, specifically an average coverage of 50x or
112 more, no contamination detected (i.e. <3 SNPs called in ribosomal genes *rrs* or *rrl* genes and
113 no more than five specific reads of another lineage in any of the 62 Coll SNPs [14]). Also 15
114 high confidence first line resistance mutations (supplementary Table S2) were screened for
115 mutant reads at <80% to exclude a mixed resistance genotype / emerging resistance.
116 The sequencing data generated is available online, NCBI accession numbers are listed in
117 supplementary table S1.
118
119 *Additional Analysis & WGS transmission scheme*
120 Any SNPs not present in all sequenced isolates from the cluster called by our pipeline vs the
121 reference genome were graded 'confident' or 'non-confident' based on visual inspection of
122 the SAMtools pile ups [15]. These SNPs were then screened for their distribution in the
123 Dutch WGS database (containing at the moment of screening 3,981 isolates from 53
124 different lineages [14]). Allele frequencies for any confident SNPs not present in all
125 sequenced isolates in the cluster were then examined in all the isolates from the cluster
126 sequenced. These allele frequencies were used to identify emerging SNPs; an allele
127 frequency over 95% was considered 'fixed' and below 5% as wild type, an allele frequency
128 between 5% and 95% was considered a mixed SNP.
129
130 A WGS transmission scheme was then produced based on the distribution of all 'confident'
131 (both fixed and mixed) SNPs based on the assumptions that;
132 (i) SNP acquisition is unique, i.e. the same SNP did not occur randomly twice.
133 (ii) Fixed SNPs are not lost in the transmission scheme.

134 (iii) Mixed SNPs are either transmitted or not transmitted to a secondary case.

135 The transmission scheme, generated from the WGS data was then compared to the
136 transmission inferences obtained on the basis of cluster investigations and the known
137 relationships between the isolates (including duplicates and multiple isolates from the same
138 individual).

139

140 *Ethical clearance*

141 The NTR Registration Committee approved our request and provided the WGS and
142 epidemiological data. As the routine collected data were pseudonymized, the study was not
143 subjected to the Dutch law for Medical Research Involving Human Subject Act (WMO).

144

145 **Results**

146 *Cluster description based on epidemiological information*

147 The selected cluster was identified in 1993 and composed of 157 tuberculosis cases by the
148 end of 2022 (overview in figure S1). In February 2003, infectious TB was diagnosed in a new
149 patient (case 5, Figure 1,transmission chain 1; red) who frequently visited pubs in one
150 specific street in Rotterdam. This patient was lost to follow-up during treatment and traced
151 in November 2003, being again highly infectious (case 23). Between 2003-2014, the cluster
152 increased by 121 patients; 80 (66%) were living in the city of Rotterdam, of which 30 (38%)
153 reported frequent visits to the pubs in the mentioned street. A new transmission chain
154 occurred in Amsterdam (Figure 1 transmission chain 7; dark blue Rotterdam, light blue
155 Amsterdam) with two highly infectious cases. Case 96 was diagnosed in 2009 by contact
156 investigation around a child with TB (case 94). A contact of this person developed
157 tuberculosis in 2012 (case 114) and had recurrent tuberculosis (case 129), more than two

158 years after completing the first treatment. Several new transmission locations were
159 identified by cluster investigation, such as a pub in another town (3 cases; not shown), a
160 penitentiary institute (3 cases), a coffee shop (5 cases), a school (4 cases, Figure 1,
161 transmission chain 6; purple) and a hospital (one case of nosocomial transmission). Case 153
162 developed tuberculosis nine years after being diagnosed with a TB infection in a contact
163 investigation and having received preventive treatment (Figure 1, transmission chain 5;
164 green).
165
166 *Cluster description based on WGS data (SNPs)*
167 A total of 773 SNPs vs the reference genome was called by our pipeline in all the sequenced
168 isolates from the selected cluster, 706 of these SNPs were shared by all 61 WGS files
169 (available online Table S1) investigated from the cluster on basis of WGS and uninformative.
170 All duplicate controls showed exactly the same confident SNPs as fixed, wild type or mixed.
171 Of the remaining 67 SNPs 53 were manually graded as confident, 14 as non-confident (Table
172 1). One SNP (Table 1, SNP 64), which was confidently graded, was present in all members of
173 the cluster but not always reliably called by our pipeline; this SNP was thus considered
174 uninformative and not used in the construction of the transmission scheme. The final
175 analysis was thus based on 52 SNPs (Table 1).
176 Subsequent comparison of the distribution of all 67 variable SNPs in the entire Dutch WGS
177 database showed that all but three of the 14 SNPs manually graded as non-confident SNPs
178 were randomly distributed throughout the database, and present in more than 100
179 unrelated isolates spread over at least seven different lineages, as a result of poor
180 mapping/sequencing at these positions (Table 1). The non-confident SNPs e.g. included
181 three SNPs in the *fhaA gene* (Table 1, SNPs 57-59) which had already been removed from

182 our routine cluster analyses by the pipeline based on the fact they were within 12 base pairs

183 from each other.

184 In total 46 of the 52 SNPs manually classified as 'confident' were exclusive to the cluster in

185 the Dutch WGS database. Five SNPs (Table 1; SNPs 8, 17, 18, 26, 29) were only shared by

186 one or two isolates from one different lineage in the Dutch WGS database. One confidently

187 graded SNP in *dna*A (Table 1, SNP 9) was shared by eight other isolates from four different

188 lineages.

189

190

191

SNP ID	Genome Position*	base change	Gene_amino acid change	Visual classification	No. isolates containing the SNP as called by the pipeline (allele frequency >80%)		No. of lineages [14] with members containing the SNP
					Isolates sequenced from the Cluster (n=58)	Dutch WGS database 2016-2022 (n=3981)	
1	2294333	A>G	Rv2047c_V65A	confident	1	0	1
2	443776	C>T	Rv0365c_W141*	confident	1	0	1
3	594270	G>A	cmaA2_Y170Y	confident	1	0	1
4	1943404	C>T	fadB3_P249L	confident	39	0	1
5	2319487	A>G	cobN_F423L	confident	39	0	1
6	4096187	G>T	Rv3658c_C251*	confident	39	0	1
7	4407263	G>A	parA_A90V	confident	39	0	1
8	4407980	G>A	gid_P75S	confident	1	2	2
9	1199	G>A	dnaA_R400H	confident	3	8	5
10	2173355	C>A	lppF_E147D	confident	1	0	1
11	963637	A>G	mog_H83R	confident	1	0	1
12	1004224	C>G	Rv0901_P90A	confident	1	0	1
13	3070852	A>C	vapB21/vapC42	confident	16	0	1
14	532562	T>C	Rv0443_V56A	confident	1	0	1
15	1044	G>A	dnaA_M348I	confident	3	0	1
16	4274903	A>G	Rv3811_I36V	confident	2	0	1
17	852736	A>C	phoR_H114P	confident	1	1	2
18	1254678	T>C	prpD_C42R	confident	1	2	2
19	2930021	G>A	SpeE /VapB41	confident	1	0	1

20	260800	C>T	lipW_R11H	confident	7	0	1
21	2243579	T>G	Rv1998c_H48P	confident	4	0	1
22	2188810	T>G	Rv1937_G105G	confident	1	0	1
23	2245506	C>T	Rv2000_R100C	confident	1	0	1
24	1113659	T>A	alaV/Rv0997	confident	1	0	1
25	2861465	C>T	aroD_E43K	confident	1	0	1
26	3412720	G>A	nrdE_H483Y	confident	1	1	2
27	528904	T>C	groEL2_V99V	confident	1	0	1
28	1130192	T>C	ispE_M1A	confident	1	0	1
29	1129791	C>A	ksgA_R214S	confident	1	1	2
30	3083748	T>A	Rv2776c_T186S	confident	2	0	1
31	2387113	C>T	Rv2125_A274V	confident	1	0	1
32	3538299	A>C	Rv3169_E354D	confident	1	0	1
33	4382137	C>T	Rv3896c_V239I	confident	1	0	1
34	4406935	G>T	parA_L199L	confident	1	0	1
35	249329	G>T	Rv0209_A98S	confident	1	0	1
36	899862	C>T	cpsY_R490H	confident	1	0	1
37	955882	T>A	fadA_I269N	confident	1	0	1
38	3541015	G>A	Rv3172c_A117V	confident	1	0	1
39	259417	G>A	Rv0216_G169S	confident	1	0	1
40	989495	C>T	citA_V123I	confident	4	0	1
41	1303690	C>G	fbiC_P254A	confident	4	0	1
42	520120	C>A	sodC_T174N	confident	1	0	1
43	3527101	G>A	nuoN_G438R	confident	1	0	1
44	1956891	C>T	Rv1730c_G119S	confident	1	0	1
45	2310534	C>T	ppm1_L74L	confident	1	0	1
46	2469060	C>T	Rv2204c_V76V	confident	1	0	1
47	1676559	C>T	Rv1486c_E109K	confident	1	0	1
48	3479870	C>T	moaD1_V57V	confident	1	0	1
49	402665	G>A	aspC_G166G	confident	1	0	1
50	2617123	G>T	mmpL9_G811*	confident	1	0	1
51	3196574	G>A	Rv2887_P48P	confident	1	0	1
52	4405126	C>G	Rv3916c_E14D	confident	1	0	1
53	1472517	T>C	rrs_0	non-confident	1	0	1
54	1472537	C>T	rrs_0	non-confident	1	0	1
55	1472581	A>C	rrs_0	non-confident	1	0	1
56	1480972	T>C	Rv1319c_E510E	non-confident	1	829	27
57	24704	T>C	fhaA_Q247Q	non-confident	3	159	23
58	24707	C>G	fhaA_E246D	non-confident	3	160	23
59	24716	A>G	fhaA_G243G	non-confident	3	183	24
60	2347521	T>C	Rv2090_L50P	non-confident	11	106	8
61	3232703	G>A	amt/ftsY	non-confident	19	560	23
62	3232759	G>A	amt/ftsY	non-confident	19	963	30
63	2339605	A>G	Rv2082_P299P	non-confident	33	401	7

64	2982956	A>G	Rv2665_P86P	confident	50	486	1
65	234477	T>G	Rv0197_Y749*	non-confident	55	2546	51
66	454295	T>C	Rv0376c_P26P	non-confident	55	3584	53
67	1779370	G>C	Rv1573_T19T	non-confident	57	1830	32

Table 1. Overview of all SNP diversity detected in the cluster, the visual grading of the SNPs and the distribution of the diverse identified in the clustered isolates in the NL database of 3981 isolates. . * = relative to the reference genome H37Rv (AL123456.3)

The accumulation of the 52 confident SNPs in a proportion of the isolates was used to inform the WGS-based transmission scheme (Figure 2) based on the assumptions outlined in the methods.

There were 44 confident SNPs that were not present in all isolates in the cluster and were fixed (>=95% of reads) in all the files in which they were observed. The remaining eight confident SNPs were present in at least one file as a mixed SNP (<95% of the reads). Six of the eight mixed SNPs were fixed in at least one isolate (Figure 2, grey [mixed] and black [fixed] symbols). The remaining two SNPs were only found as mixed in only a single isolate and never fixed (SNPs 22 and 32).

The duplicate controls were identical and placed together on the WGS transmission tree with pair 96 both containing mixed SNP 9 (Figure 2, triangle symbol) at 72% and 84% respectively.

Only two of the eight mixed SNPs were present as a mixed SNP in more than one sequencing file; one duplicate control (SNP 9 in sample 96 mentioned above) and one (SNP 15, Figure 2 round symbols) in two different sequencing files. The two files with a shared mixed SNP 15 were serial isolates from the same individual (with a reactivation) (Figures 1 and 2; 114, 129). These results were derived from isolates collected in September 2012 (SNP 15; 8%) and March 2015 (SNP 15; 68%) from the same patient. Four files also contained a fixed SNP 15(135, 139, 154 and 154*); two of these were duplicate controls (154, 154*). The

215 first isolate in which SNP 15 was mixed (September 2012) also contained a mixed SNP 14
216 (17%). SNP 14 was fixed in one additional isolate (134) without SNP 15. In the second isolate
217 (Figure 2 129; March 2015) two other additional mixed SNPs were present (SNP 17; 54%,
218 SNP 18; 58%). One isolate (Figure 2; 135) contained both SNPs 17 and 18 fixed along with
219 SNP 15.
220
221 *Comparison of the epidemiological and WGS based transmission schemes.*
222 The comparison of the information from the WGS analysis and the epidemiological
223 investigations included 33 confirmed epi-links (68 cases) and two relapses. The serial
224 isolates case 5 and its relapse (case 23) were identical, while the isolates from case 114 and
225 its relapse (case 129) contained mixed SNPs; (SNP 14; 17%, and SNP 15; 8%, in the first
226 episode SNP 15; 68%, SNP 17; 54% and SNP 18; 58% in the second isolate (Figure 2).
227 WGS failed for one isolate selected for WGS analysis (Figure 1, case 27 shaded oval). Out of
228 the remaining 33 of the confirmed epi-links and two repeat isolates screened for SNPs, 16
229 were identical, 11 contacts gained one SNP, four contacts gained two SNPs and two contacts
230 gained three new confident SNPs (Table 2). There were two epi-linked isolates (cases 85 and
231 97) that based on the WGS scheme lost a unique fixed SNP identified in the isolate of
232 putative source case 77 (Table 2, Figures 1 and 2). In retrospect it was concluded that these
233 two cases (85, 97) were related cases to case 77 but could also very well have been infected
234 by case 69, who was the source of infection of case 77, which would be consistent with the
235 WGS transmission scheme (Figure 2; supplementary table S1).
236
237

238 **Table 2.** Distribution of SNP distances between isolates considered to have an

239 epidemiological link based on contact investigations undertaken by the MHSs.

SNPs between source and secondary case		Number of Epi-links
SNPs gained	SNPs Lost	
0	0	16
1	0	10
2	0	5
3	0	2
0	1	1
1	1	1

240

241 All cases that occurred since 2015 in patients residing in Amsterdam and surrounding areas

242 contained SNPs 4-7 and for most epidemiological links were identified (Figure 1 and 2,

243 transmission chain 7 dark and light blue lines).

244

245 **Discussion**

246 This study shows that with proper verification of variable mutations identified in a large

247 cluster useful information was revealed with respect to the micro-evolution of the strain.

248 The resulting branching divided the large cluster in several satellite-clusters, which is

249 relevant for monitoring the development of the outbreak and steer public health actions.

250 The inferences based on a detailed WGS analysis were almost entirely compatible with the

251 possible scenarios developed by the MHSs based on information extracted from

252 epidemiological investigations. The initial outbreak occurred in Rotterdam, the second

253 largest city in the Netherlands, and was associated with known risk factors like frequent

254 visits to pubs in one area. From 2009 onwards, cases emerged in a sub-cluster in

255 Amsterdam. Over time the outbreak strain acquired four confident SNPs, which were

256 present in isolates linked to subsequent cases in Amsterdam, but not in the initial cases in
257 Rotterdam.
258 We classified all 67 SNPs that were variably present in the cluster using two independent
259 approaches; firstly visual inspection of the sequencing reads (SAMtools Pileups [15]) and
260 secondly an analysis of distribution of the SNPs in the WGS database in the Netherlands
261 comprising 3,981 diagnostic isolates (Table 1). Of the 14 visually non-confident called SNPs,
262 11 were also identified as non-informative based on their presence in a large number of
263 genetically diverse isolates. A proportion of these SNPs were found in genomic regions that
264 have been previously recognized as poorly mapping [16]. The remaining three SNPs in *rrs*
265 (Table 1; SNP 53-55) were considered unreliable, presumably as a result of contamination
266 with other bacteria. Finally, one confidently called SNP (64; Rv2662_P86P) was found in 486
267 other isolates in the WGS database, all belonging to one lineage and was likely present in all
268 members of this lineage but not called by our pipeline in a small proportion of lineage
269 4.1.2.1 isolates due to poor mapping of a A>G substitution directly next to a string of G
270 nucleotides (Table 1).
271 Six of the 52 remaining confident SNPs were not unique to the cluster but present in one,
272 two and in one case eight additional isolates in the WGS database of 3,981 unique isolates.
273 Positions under selection in clonal organism such as *M. tuberculosis* are subject to
274 homoplasy, most notably mutations associated with drug resistance [17]. These six SNPs are
275 thus potential candidates for positions under selection. Strikingly, SNP 9 [dnaA_R400H]
276 emerged in the cluster and was identified in eight unrelated isolates in diverse lineages,
277 present in the WGS database (Table 1). Additionally, a second SNP also emerged in this
278 cluster in the *dna*A gene (M348I), this SNP was only identified in isolates from the cluster
279 studied (Table 1, SNP 15). Mutations in *dna*A have previously been observed in spontaneous

280 *M. tuberculosis* drug resistant laboratory mutants [11], and have been associated with
281 increased survival during INH treatment and more robust growth at the MIC of isoniazid
282 [18]. Furthermore, *dna*A mutations have also been observed in more efficiently transmitted
283 variants in another cluster of isolates that was subject to detailed SNP analysis [12]. Notably,
284 both *dna*A mutations observed in this cluster were initially identified as mixed SNPs and
285 transmitted to other individuals. Furthermore, the *dna*A_M348I mutation was also present
286 in two isolates from a treatment failure / reactivation case (cases 114, 129) isolated more
287 than two years apart (Figure 2, transmission chain 7; blue) with at least four mixed SNPs
288 which were identified as fixed in other individuals who were presumably infected by this
289 case.
290 After filtering, either manually or screening our database for unexpected homoplasy, high
291 confidence informative SNPs duplicate controls were identical and serial/reactivation
292 isolates were all directly linked in the WGS scheme.
293 We identified the SNP variability within the cluster on the basis of SNP thresholds and then
294 filtered this short list of SNPs based on a screen for variability of these variable positions in a
295 larger database allowing “noisy” SNPs to be identified and filtered out of the analysis. In this
296 way true potentially epidemiologically informative SNPs within the cluster were identified.
297 The resulting short list of confident informative positions could then be analysed in all other
298 isolates sequenced from the cluster in detail with the aim of identifying the isolates in the
299 cluster in which the SNPs were emerging/not fixed [10-12, 19]. This approach in cancelling
300 sequence noise using a large database to screen for epidemiologically informative variation,
301 adds to the epidemiological typing. Comparing the epidemiological data available to WGS
302 based transmission scheme supports the assumption that careful analysis can identify
303 epidemiologically informative variability in closely related sequences.

304

305 *Limitations*

306 Even with a permissive SNP calling algorithm for a significant proportion of individuals (20
307 of 54 ,37%) in the cluster analysis we were unable to identify any confident SNP variation
308 with at least one other case. Long read sequencing or the use of pan genome analysis may
309 allow the analysis of more variable regions of the genome that currently cannot be
310 confidently analysed using short read sequencing [16] allowing the detection of large
311 genome rearrangements as well as potentially allowing analysis of regions not present in the
312 reference genome currently used could potentially increase the typing resolution [20, 21]
313 within recently clustered isolates. We sequenced cultured isolates raising the possibility that
314 sub-populations of bacteria present in the sample may grow at different rates, be lost, or
315 additional mutations accumulate during culture [22]. Finally, we did not screen the whole
316 genomes for mixed SNPs to streamline the analysis, targeted only SNPs that were
317 variable/informative within the clustered isolates sequenced. This detailed analysis was
318 performed on a single larger cluster. The identification of emerging SNPs will only be
319 informative when the majority of cases in the cluster are identified and sequence data is
320 available for at least the majority of the potentially linked cases.

321

322 *Conclusion*

323 In conclusion, although clustering can be effectively ruled out on the basis of SNP distance
324 thresholds alone we demonstrate there is additional information available in micro
325 variability within the WGS data from isolates in a large cluster that can be useful to identify
326 possible links and target epidemiological investigations. We confirm mixed SNPs, if present,
327 may be fixed in subsequent isolates and strongly suggest a transmission direction and direct

328 epidemiological link as supported by the epidemiological information available from the
329 cluster studied and recently proposed by others [19, 23, 24]. Screening for these mixed SNPs
330 in a long established cluster can therefore provide epidemiologically useful information.
331 Strategies to automate and validate this type of SNP analysis could be implemented. Marker
332 SNPs characteristic of specific transmission chains [25] can be identified in larger clusters
333 and are potentially more precise than SNP thresholds to identify epidemiologically linked
334 isolates.

Figure legends:

Figure 1 Transmission scheme based on epidemiological information. Cases sequenced with an epidemiological link confirmed by MHSs are present as ovals, with arrows showing the direction of transmission. Relapse cases in the same individual are linked with a solid black line. Oval borders indicate the location of the case; solid thick border Rotterdam; dashed thick border Amsterdam; thin border; other towns or villages in the Netherlands. The arrow colour indicates a distinct epidemiological transmission chain. Dotted arrows indicate situations in which transmission was considered likely. The cases and epidemiological transmission chains are presented using the same colours in Figure 2 (except case 27 shaded, for which WGS data was unanalysable due to contamination of the culture). The year and number of isolates from the cluster is indicated on the left hand side (and in table S1).

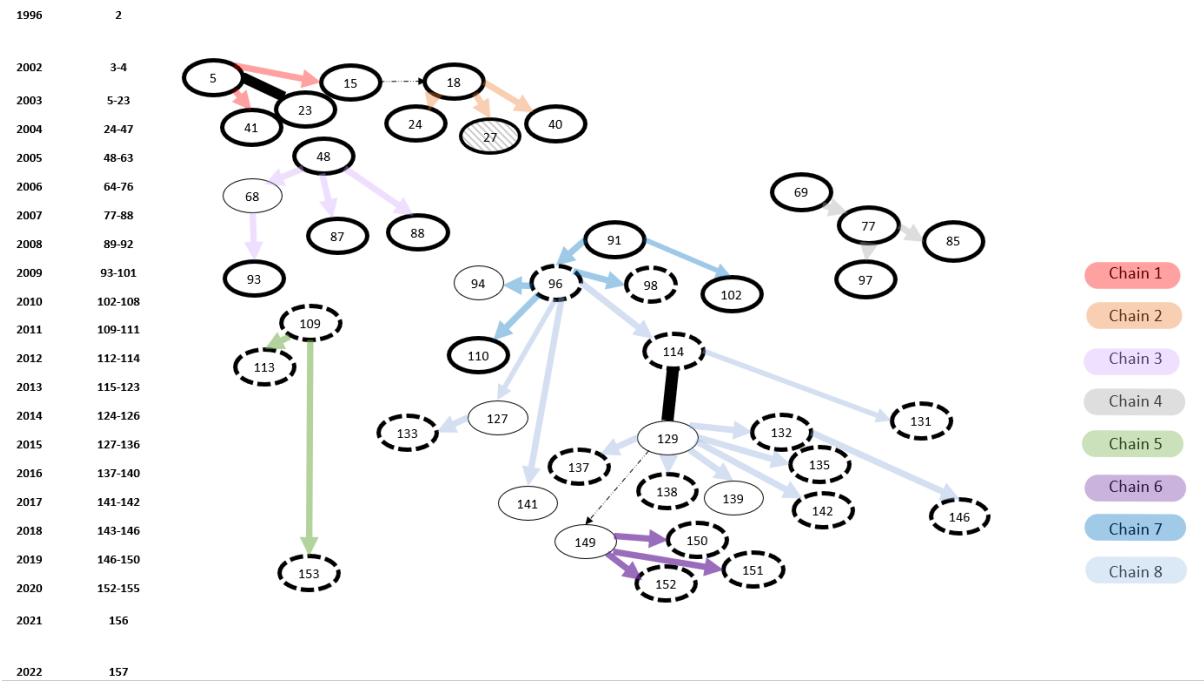
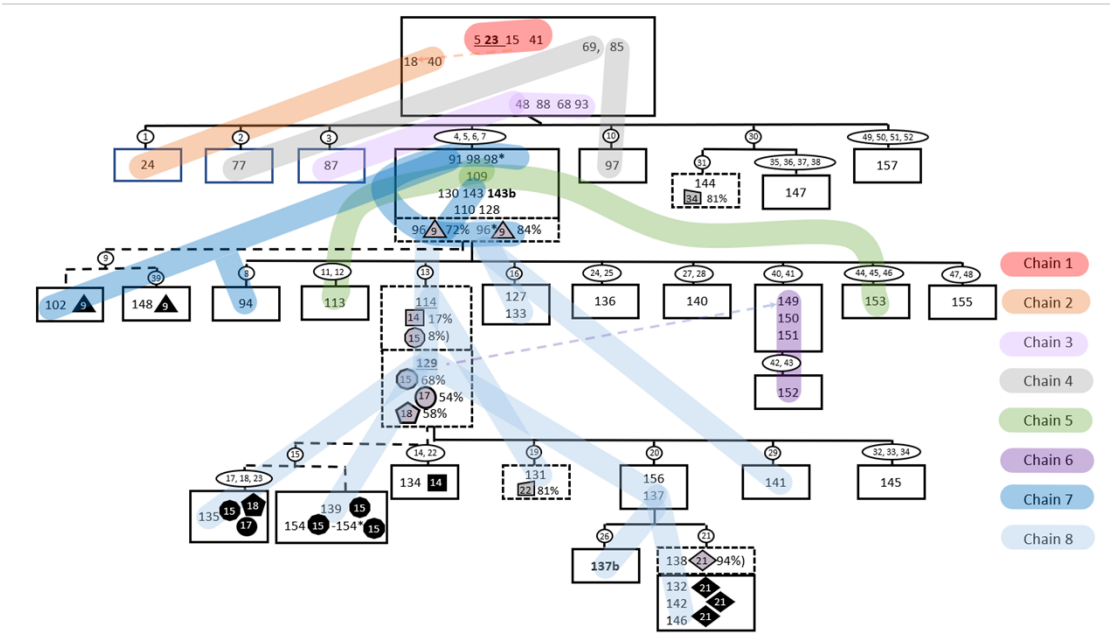


Figure 2: Transmission scheme based on WGS data. Numbers inside boxes indicate NTR cases (Figure 1), cases with zero fixed SNP differences are inside the same box. NTR case numbers appearing more than once indicate duplicate sequences generated from the same isolate are indicated by an *. Follow up isolates from the same NTR case are indicated in bold with a “b”. Reactivations have a different NTR case number but are underlined with the second episode in bold. Numbers inside white ovals indicate a fixed SNP (Table 1); unfixed SNPs inside shaded symbols; a grey symbol if the SNP was mixed and a black symbol when in subsequent cases where the SNP was fixed, they have their frequency indicated as a percentage after a comma inside the box. Solid lines linking boxes indicate the inferred transmission of fixed SNPs, dashed lines the inferred transmission of mixed SNPs. Coloured lines indicate isolates in epidemiologically linked transmission chains(Figure 1).



Acknowledgments

This project was funded by the Dream Fund project "No more pandemics" of the Dutch Postcode Lottery, and the Dutch Ministry of Health, Welfare and Sports. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication. We would also like to acknowledge the support of The Netherlands Organisation of Health, Research and Development (ZonMw project 541002006) which supported the development of the SNP screening algorithm. We thank the MHSs in the Netherlands for their highly valuable contribution to this paper and ongoing collaboration.

Contributions (CRediT)

Rina de Zwaan: Data curation, Formal Analysis, Writing review and editing.
Gerard de Vries: Conceptualization, Data curation, Methodology, Writing – review & editing.
Ella Ubbelohde: Data analysis, Writing – review & editing.
Arnout Mulder: Data curation.
Miranda Kamst: Project administration, Supervision.
Karin Rebel: Data curation.
Saskia Kautz: Data curation.
Kristin Kremer: Funding acquisition, Project administration, Supervision, Writing – review & editing.
Richard Anthony: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft.
Dick van Soolingen: Conceptualization, Funding acquisition, Writing -original draft, Writing – review & editing.

392 **References**

393 1. Jajou R, Neeling AD, Hunen RV, Vries GD, Schimmel H, Mulder A, Anthony R, Hoek
394 WV, Soolingen DV. Epidemiological links between tuberculosis cases identified twice
395 as efficiently by whole genome sequencing than conventional molecular typing: a
396 population-based study. PloS one. 2018 Apr 4;13(4):e0195413.
397 2. de Vries, G., van Beurden, K., Kamst-van Agterveld, M., Rebel, K., Veulings, Y., & van
398 Soolingen, D. (2020). Tuberculoseclusters in kaart gebracht door whole genome
399 sequencing.
400 3. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ,
401 Hawkey PM, Crook DW, Parkhill J. Whole-genome sequencing to delineate
402 Mycobacterium tuberculosis outbreaks: a retrospective observational study. The
403 Lancet infectious diseases. 2013 Feb 1;13(2):137-46.
404 4. Jajou R, Kohl TA, Walker T, Norman A, Cirillo DM, Tagliani E, Niemann S, de Neeling
405 A, Lillebaek T, Anthony RM, van Soolingen D. Towards standardisation: comparison
406 of five whole genome sequencing (WGS) analysis pipelines for detection of
407 epidemiologically linked tuberculosis cases. Eurosurveillance. 2019 Dec
408 12;24(50):1900130.
409 5. Anthony RM, Tagliani E, Nikolayevskyy V, de Zwaan R, Mulder A, Kamst M, Ködmön
410 C, van der Werf MJ, Cirillo D, van Soolingen D, ERLTB-Net members. Experiences
411 from 4 years of organization of an external quality assessment for *Mycobacterium*
412 *tuberculosis* whole-genome sequencing in the European Union/European Economic
413 Area. Microbiology Spectrum. 2023 Feb 14;11(1):e02244-22.
414 6. Nikolayevskyy V, Niemann S, Anthony R, Van Soolingen D, Tagliani E, Ködmön C, van
415 der Werf MJ, Cirillo DM. Role and value of whole genome sequencing in studying

416 tuberculosis transmission. Clinical Microbiology and Infection. 2019 Nov
417 1;25(11):1377-82.
418 7. Jajou R, Neeling AD, Hunen RV, Vries GD, Schimmel H, Mulder A, Anthony R, Hoek
419 WV, Soolingen DV. Epidemiological links between tuberculosis cases identified twice
420 as efficiently by whole genome sequencing than conventional molecular typing: a
421 population-based study. PloS one. 2018 Apr 4;13(4):e0195413.
422 8. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, Andersen
423 AB, Niemann S, Kohl TA. Tracing *Mycobacterium tuberculosis* transmission by whole
424 genome sequencing in a high incidence setting: a retrospective population-based
425 study in East Greenland. Scientific reports. 2016 Sep 12;6(1):33180.
426 9. Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, García-
427 de-Viedma D. Whole genome sequencing analysis of intrapatient microevolution in
428 *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis
429 transmission. The Journal of infectious diseases. 2014 Jan 1;209(1):98-108.
430 10. Schürch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, Smith NH, van
431 Soolingen D. The tempo and mode of molecular evolution of *Mycobacterium*
432 *tuberculosis* at patient-to-patient scale. Infection, genetics and evolution. 2010 Jan
433 1;10(1):108-14.
434 11. Bergval I, Coll F, Schuitema A, de Ronde H, Mallard K, Pain A, McNerney R, Clark TG,
435 Anthony RM. A proportion of mutations fixed in the genomes of in vitro selected
436 isogenic drug-resistant *Mycobacterium tuberculosis* mutants can be detected as
437 minority variants in the parent culture. FEMS microbiology letters. 2014 Dec
438 4;362(2):1-7.

12. Abascal E, Genestet C, Valera A, Herranz M, Martinez-Lirola M, Muñoz P, Dumitrescu O, García de Viedma D. Assessment of closely related *Mycobacterium tuberculosis* variants with different transmission success and in vitro infection dynamics. Scientific reports. 2021 May 26;11(1):11041.
13. Specht IO, Petros BA, Moreno GK, Brock-Fisher T, Krasilnikova LA, Schifferli M, Yang K, Cronan P, Glennon O, Schaffner SF, Park DJ. Inferring Viral Transmission Pathways from Within-Host Variation. medRxiv. 2023 Oct 15.
14. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nature communications. 2014 Sep 1;5(1):4812.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. bioinformatics. 2009 Aug 15;25(16):2078-9.
16. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. NATURE reviews microbiology. 2019 Sep;17(9):533-45.
17. Hazbón MH, Brimacombe M, Bobadilla del Valle M, Cavatore M, Guerrero MI, Varma-Basil M, Billman-Jacobe H, Lavender C, Fyfe J, García-García L, León CI. Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant *Mycobacterium tuberculosis*. Antimicrobial agents and chemotherapy. 2006 Aug;50(8):2640-9.
18. Hicks ND, Giffen SR, Culviner PH, Chao MC, Dulberger CL, Liu Q, Stanley S, Brown J, Sixsmith J, Wolf ID, Fortune SM. Mutations in dnaA and a cryptic interaction site

463 increase drug resistance in *Mycobacterium tuberculosis*. PLoS Pathogens. 2020 Nov
464 30;16(11):e1009063.
465 19. Ortiz AT, Kendall M, Storey N, Hatcher J, Dunn H, Roy S, Williams R, Williams C,
466 Goldstein RA, Didelot X, Harris K. Within-host diversity improves phylogenetic and
467 transmission reconstruction of SARS-CoV-2 outbreaks. Elife. 2023 Sep 21;12:e84384.
468 20. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC,
469 Glynn JR, Crampin AC, Alves A, Bessa TB. Recombination in pe/ppe genes contributes
470 to genetic variation in *Mycobacterium tuberculosis* lineages. BMC genomics. 2016
471 Dec;17:1-2.
472 21. Gómez-González PJ, Campino S, Phelan JE, Clark TG. Portable sequencing of
473 *Mycobacterium tuberculosis* for clinical and epidemiological applications. Briefings in
474 bioinformatics. 2022 Sep;23(5):bbac256.
475 22. Vargas Jr R, Farhat MR. Antibiotic treatment and selection for *glp*K mutations in
476 patients with active tuberculosis disease. Proceedings of the National Academy of
477 Sciences. 2020 Feb 25;117(8):3910-2.
478 23. Walter KS, Cohen T, Mathema B, Colijn C, Sobkowiak B, Comas I, Goig GA, Croda J,
479 Andrews JR. Signatures of transmission in within-host *Mycobacterium tuberculosis*
480 complex variation: a retrospective genomic epidemiology study. The Lancet Microbe.
481 2024.
482 24. Wyllie D, Do T, Myers R, Nikolayevskyy V, Crook D, Peto T, Alexander E, Robinson E,
483 Walker AS, Campbell C, Smith EG. *M. tuberculosis* microvariation is common and is
484 associated with transmission: analysis of three years prospective universal
485 sequencing in England. Journal of Infection. 2022 Jul 1;85(1):31-9.

25. de Neeling AJ, Tagliani E, Ködmön C, van der Werf MJ, van Soolingen D, Cirillo DM, Anthony RM. Characteristic SNPs defining the major multidrug-resistant *Mycobacterium tuberculosis* clusters identified by EuSeqMyTB to support routine surveillance, EU/EEA, 2017 to 2019. Eurosurveillance. 2024 Mar 21;29(12):2300583.

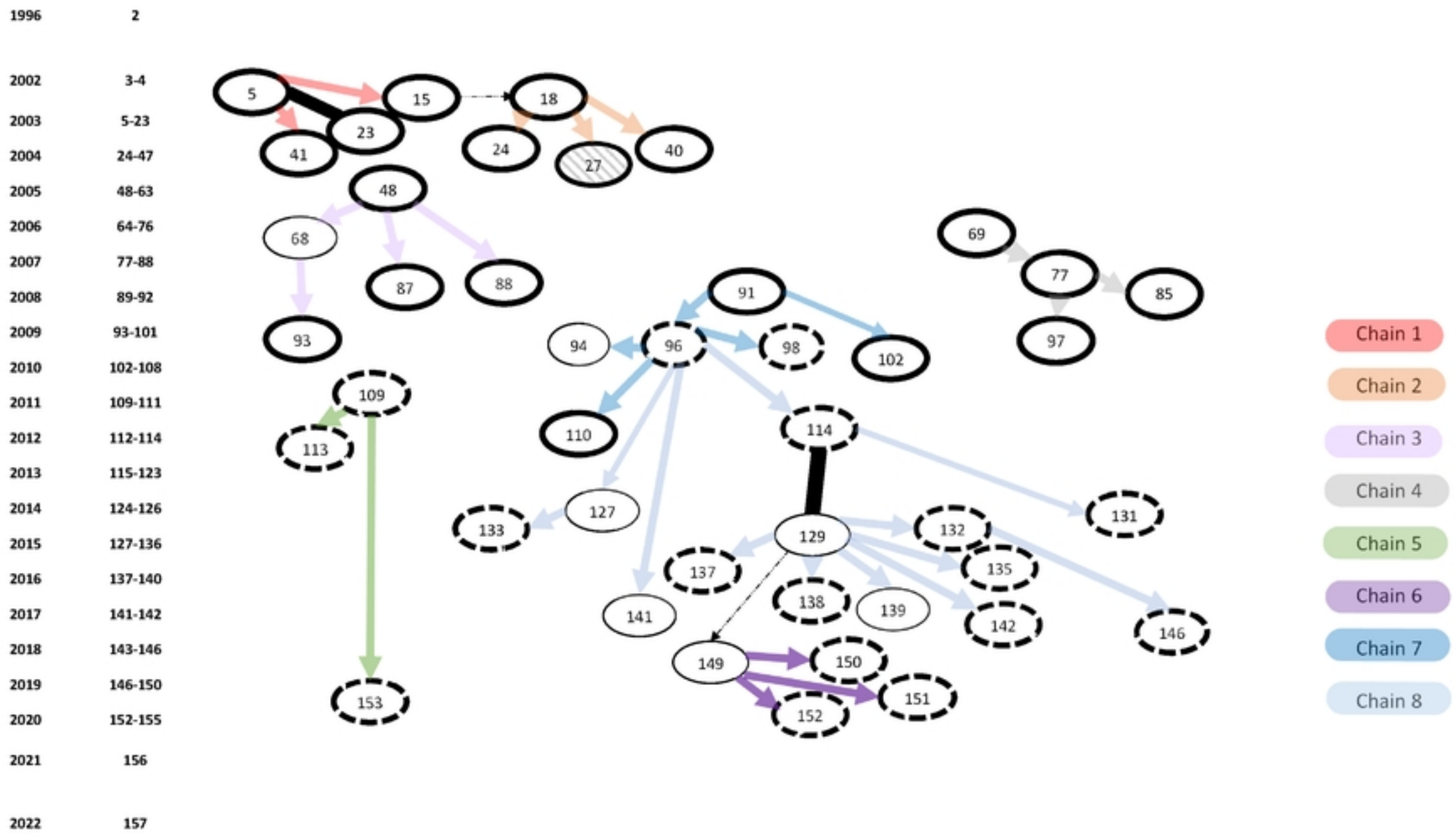

1996 2
2002 3-4
2003 5-23
2004 24-47
2005 48-63
2006 64-76
2007 77-88
2008 89-92
2009 93-101
2010 102-108
2011 109-111
2012 112-114
2013 115-123
2014 124-126
2015 127-136
2016 137-140
2017 141-142
2018 143-146
2019 146-150
2020 152-155
2021 156
2022 157
5
15
18
23
41
24
27
40
48
68
69
77
85
87
88
91
93
94
96
98
102
97
109
113
110
114
127
133
129
132
131
137
135
138
139
141
142
146
149
150
151
152
153
Chain 1
Chain 2
Chain 3
Chain 4
Chain 5
Chain 6
Chain 7
Chain 8

Figure1

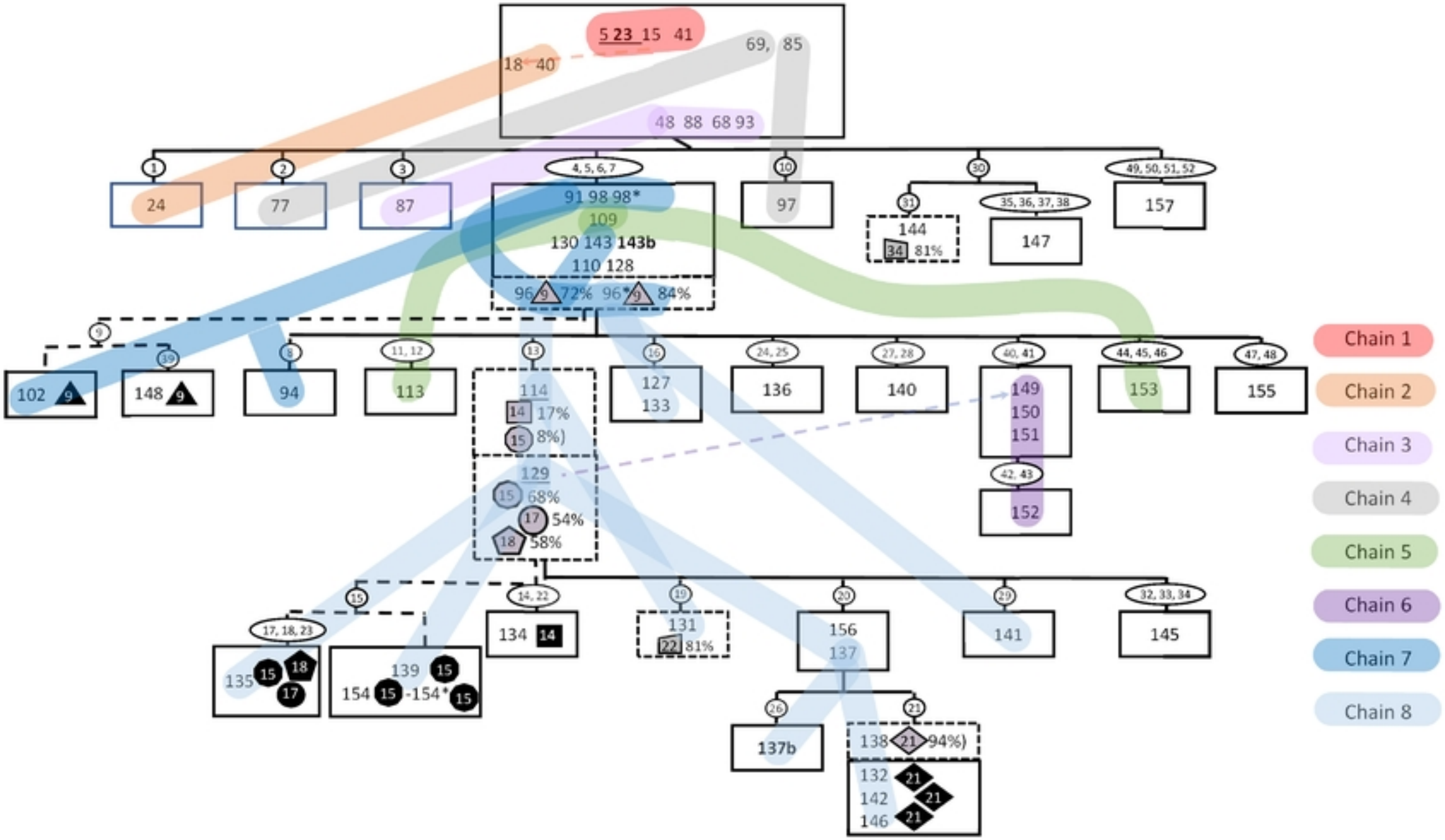

5 23 15 41
18 40
69, 85
48 88 68 93
24
77
87
91 98 98*
109
130 143 143b
110 128
96 72% 96* 84%
97
144
81%
147
157
102
148
94
113
114
17%
8%)
127
133
136
140
149
150
151
152
153
155
129
68%
54%
58%
134
131
81%
156
137
141
145
135
139
154 -154*
137b
138 94%)
132
142
146
Chain 1
Chain 2
Chain 3
Chain 4
Chain 5
Chain 6
Chain 7
Chain 8

Figure2