Figure S1. cgMLST profile painting illustrates large recombinations

The 7198 total (upper right) and 138 hybrid (bottom) cgMLST profiles are represented. Loci are represented in their order along the reference genome NTUH-K2044. Each allele is colored according to its attribution to a phylogroup (see color key; white: unattributed). Gene loci are indicated at the bottom of the figure. Blocs with distinctive colors within some profiles correspond to large recombination events.

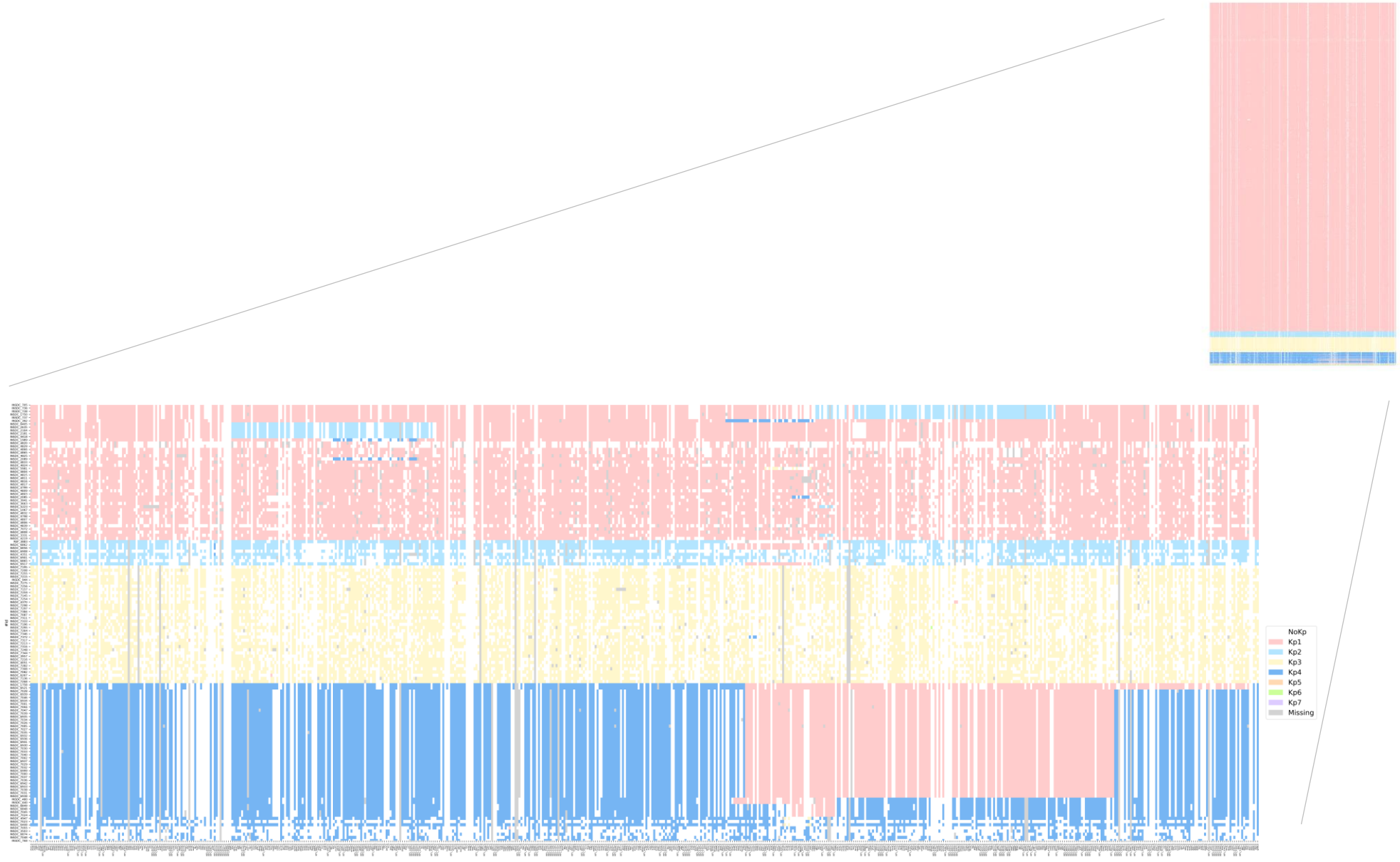


Figure S2. Genome inclusion flowchart

The flowchart summarizes the genomes inclusion process.

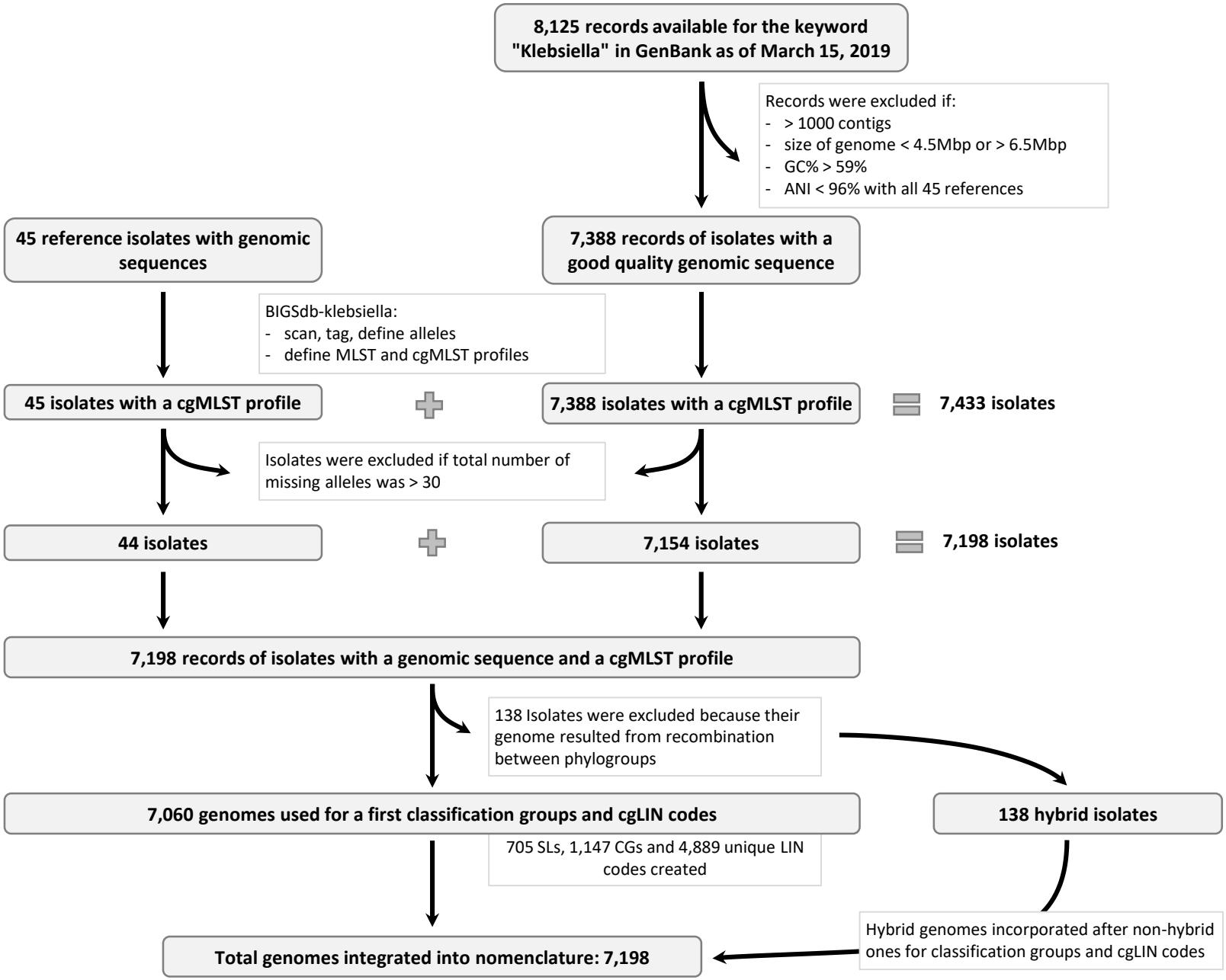
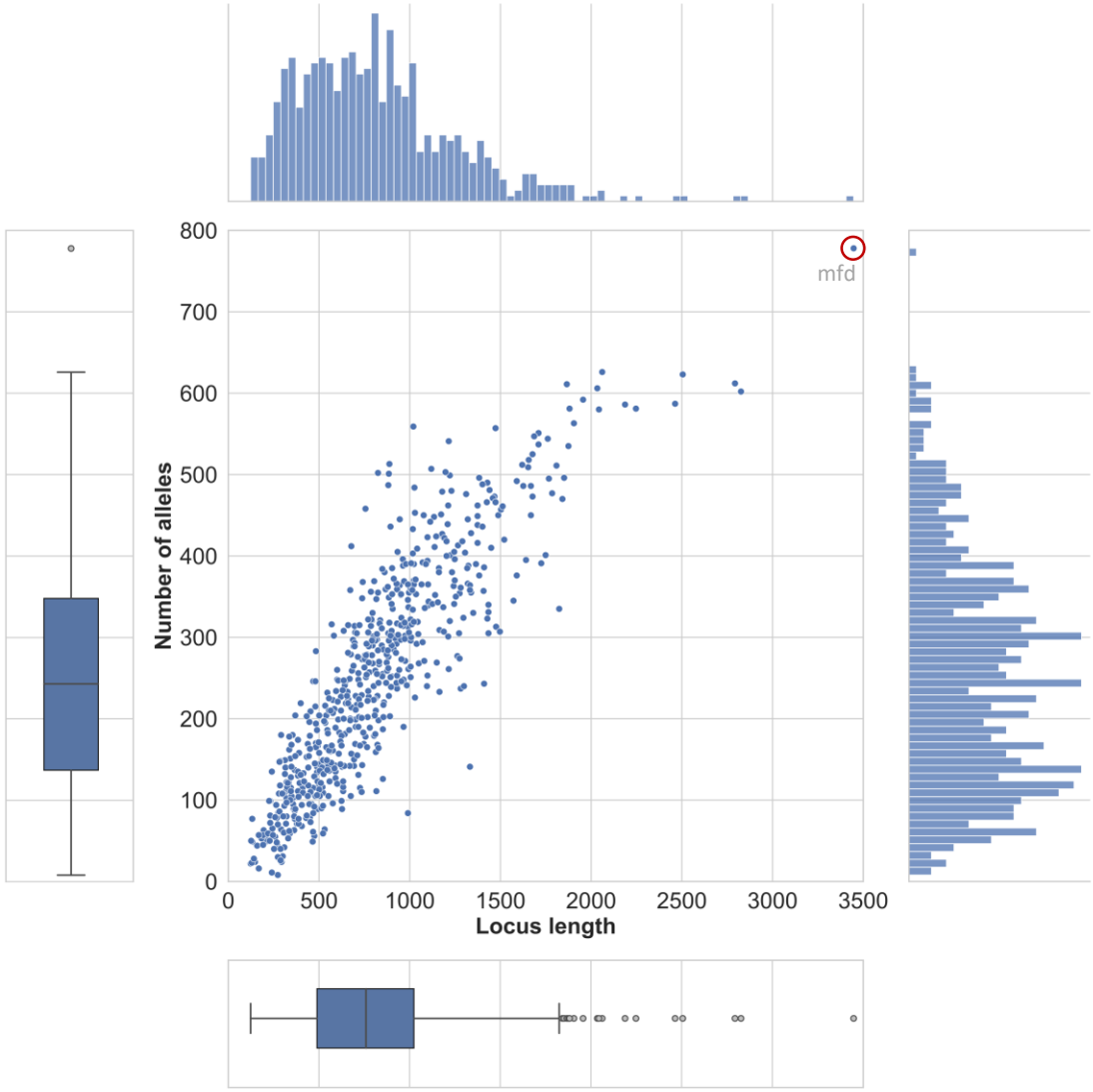


Figure S3. Characteristics of the 629 loci of the cgMLST scheme

(A) Effect of locus size on allele number. Central panel: each point represents a locus; the X-axis corresponds to the locus length and the Y-axis to the number of alleles; top/bottom and right/left panels: the distributions and the boxplots of each parameter are shown. (B) Effect of intra-gene recombination on allele diversity. Boxplots show the distribution of the number of alleles according to the different significance levels of the PHI statistic. n.a.: not applicable (polymorphism was too low); n.s.: loci for which no significant recombination was detected.

A. Locus length *versus* number of alleles



B. Effect of recombination on allele number per locus

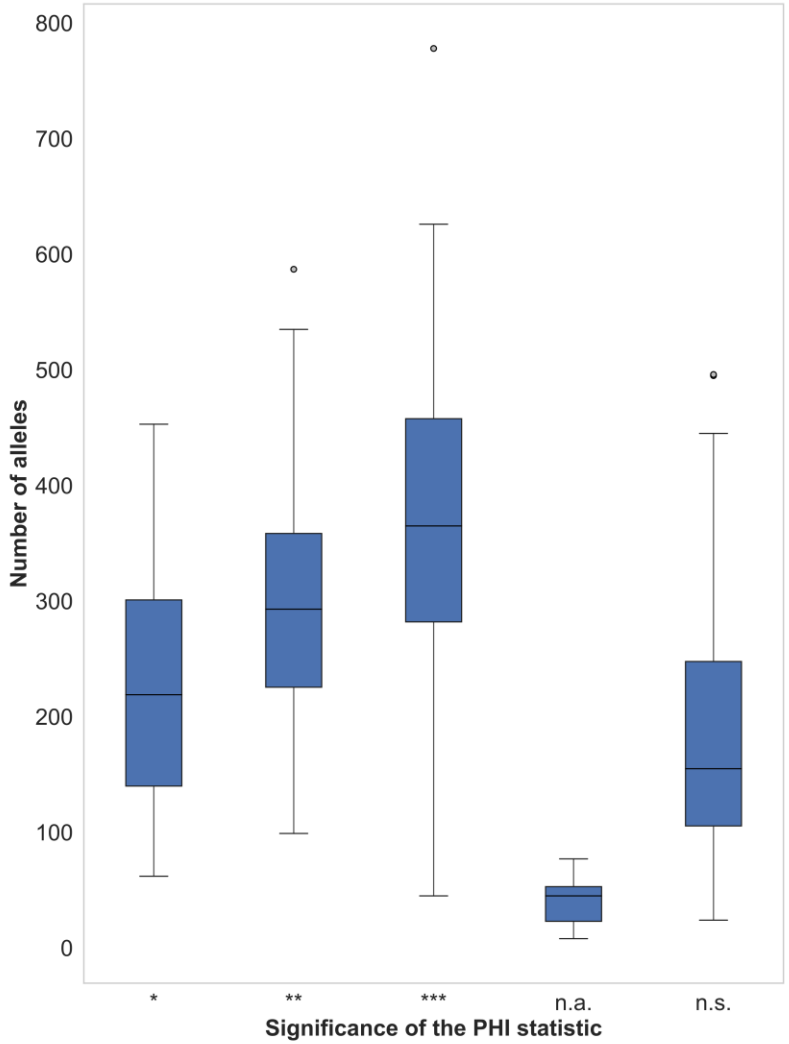


Figure S4. The distribution of pairwise distances based on Average Nucleotide Identity (ANI) and cgMLST

(A) cgMLST similarity versus ANI. In the central panel, each point represents a pair of strains; the X-axis corresponds to the cgMLST profile similarity whereas the Y-axis corresponds to the ANI calculated on the whole genomes (FastANI v1.1); the corresponding density distributions are shown on the outside of the graph. Colors correspond to inter-phylogroup (orange) and intra-phylogroup (blue) genome pairs. (B) Distribution of ANI values with box-plots of phylogroup comparisons. (C) Distribution of cgMLST similarity values with box-plots of phylogroup comparisons.

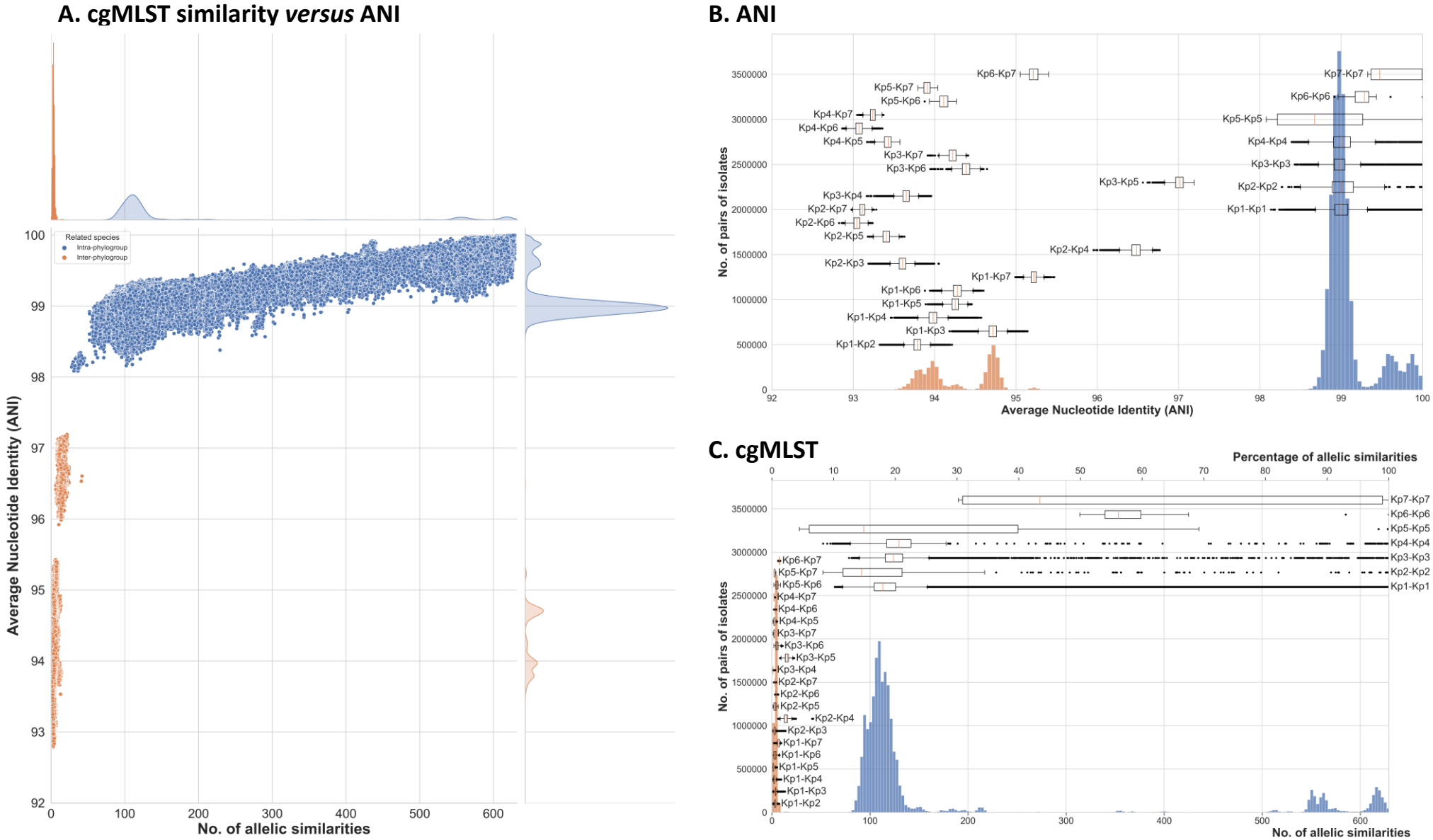
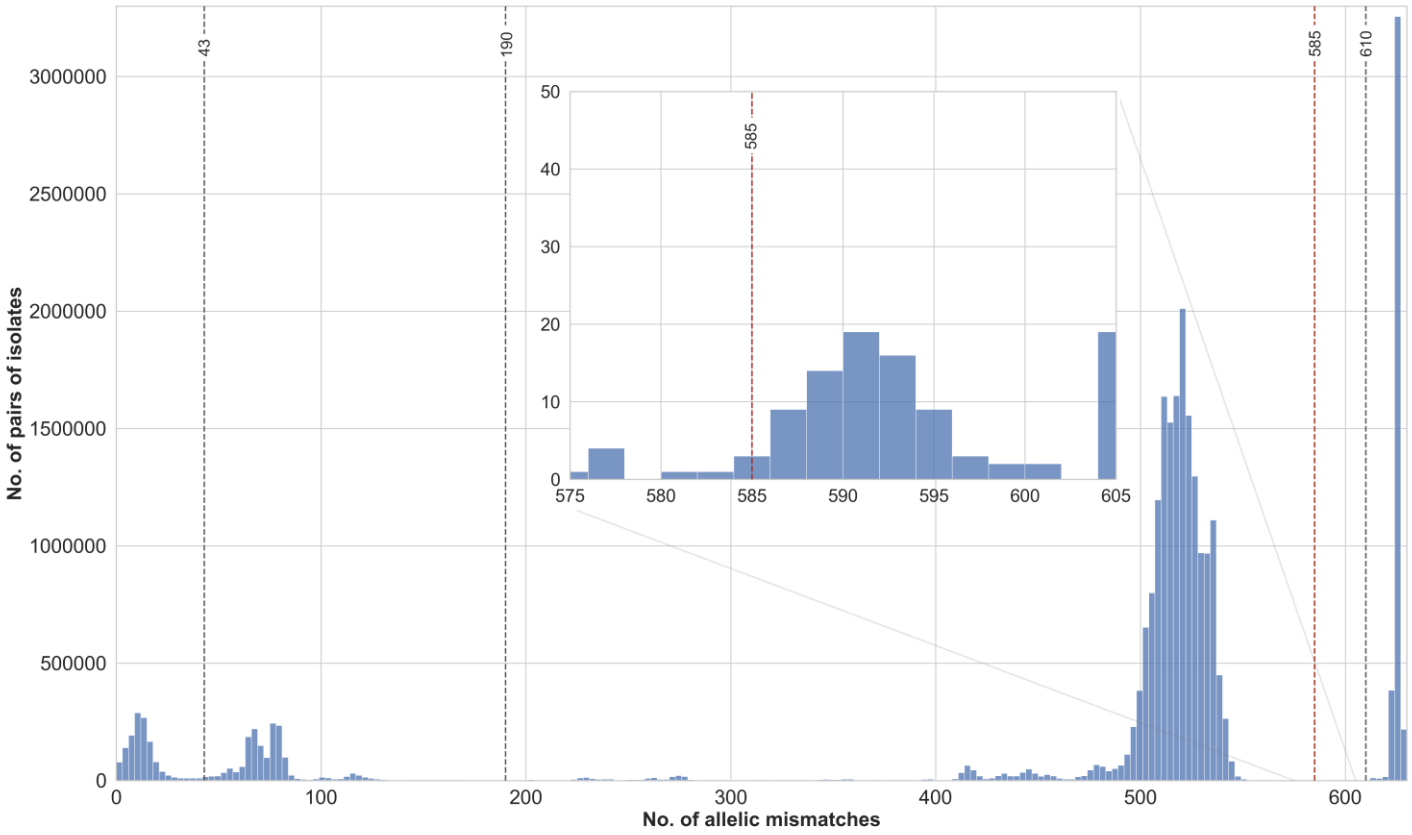


Figure S5. Details of cgMLST pairwise distances distributions

(A) cgMLST pairwise distance distribution, with a zoom on the range of 575-605 allelic differences (inter-subspecies comparisons). The chosen threshold of 585 for subspecies delineation is indicated by a red vertical line. (B) cgMLST pairwise distance distribution, with a zoom on the range of 40-100 allelic differences. Sectors of the bars are colored according to the 7-gene MLST sequence type (ST) of the compared genomes (see key). The 79-mismatch mode of the distribution is mostly composed of ST258-ST11 pairs. This observation is consistent with ST258 having evolved through a 1.1 MB large-scale recombination event of an ST11 ancestor with a ST442 donor (Chen et al., 2014): these two STs differ by 57 cgMLST alleles in this 1.1 MB region (computed for GCA_000445405.1 JM45 and SB4938_Kp13), and 21 additional allelic mismatches are observed on average between ST258 (GCA_000597905.1 NJST258_2) and the ST11 (GCA_000445405.1 JM45) ancestor outside the recombined region. Other important contributors to this third mode were comparisons between ST11-ST512, ST258-ST340, ST258-ST437.

Figure S5. Details of cgMLST pairwise distances distributions

A. cgMLST pairwise distance distribution, with a zoom in range 575-605 allelic differences



B. cgMLST pairwise distance distribution, with a zoom in range 40-100 allelic differences

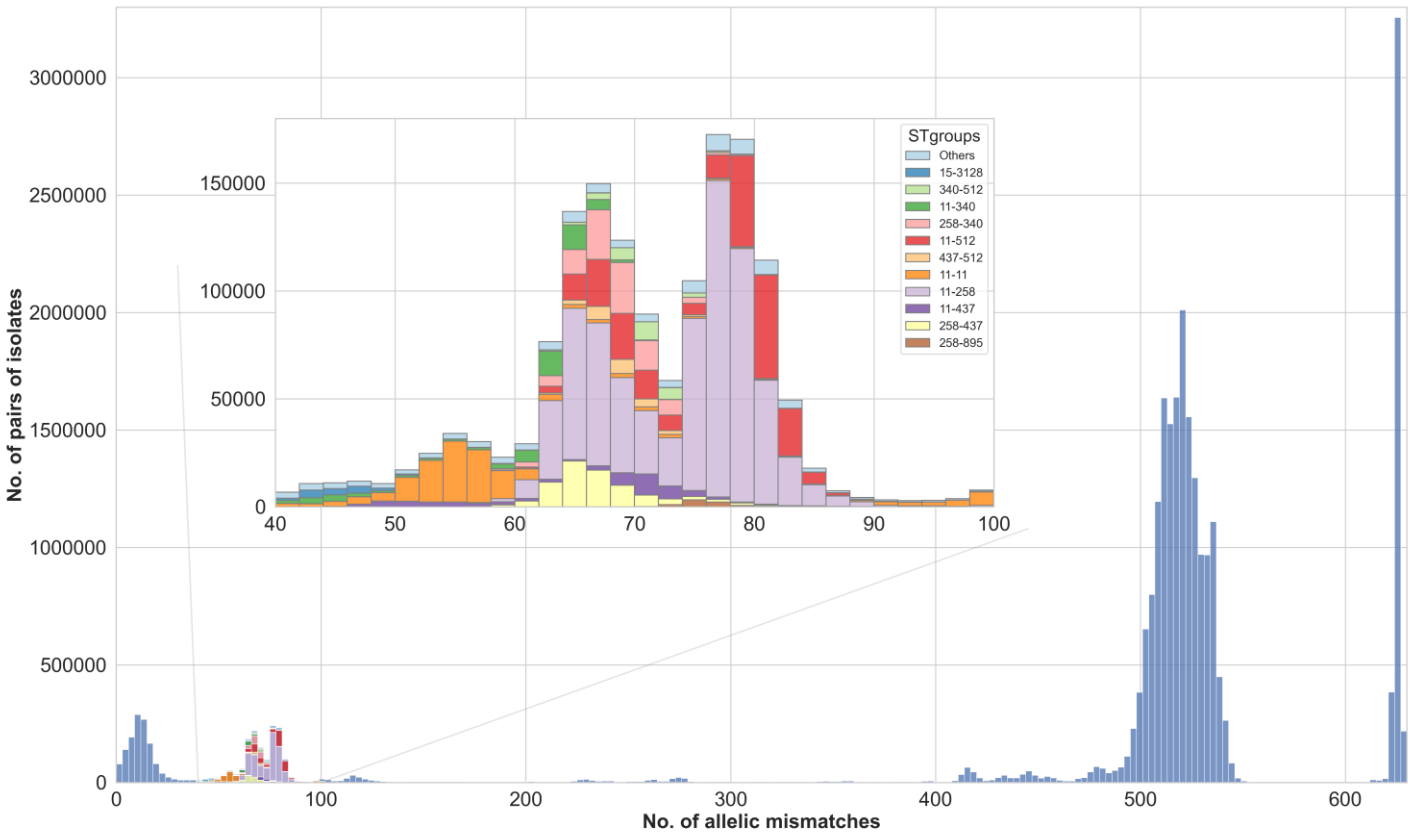


Figure S6. Correspondence of ST, sublineage and clonal group classifications for 9 major *K. pneumoniae* sublineages

ST: 7-gene MLST sequence type. The genomes without ST are denoted NA. The identifiers of the sublineages and clonal groups classifications are those inherited from MLST using our mapping algorithm.

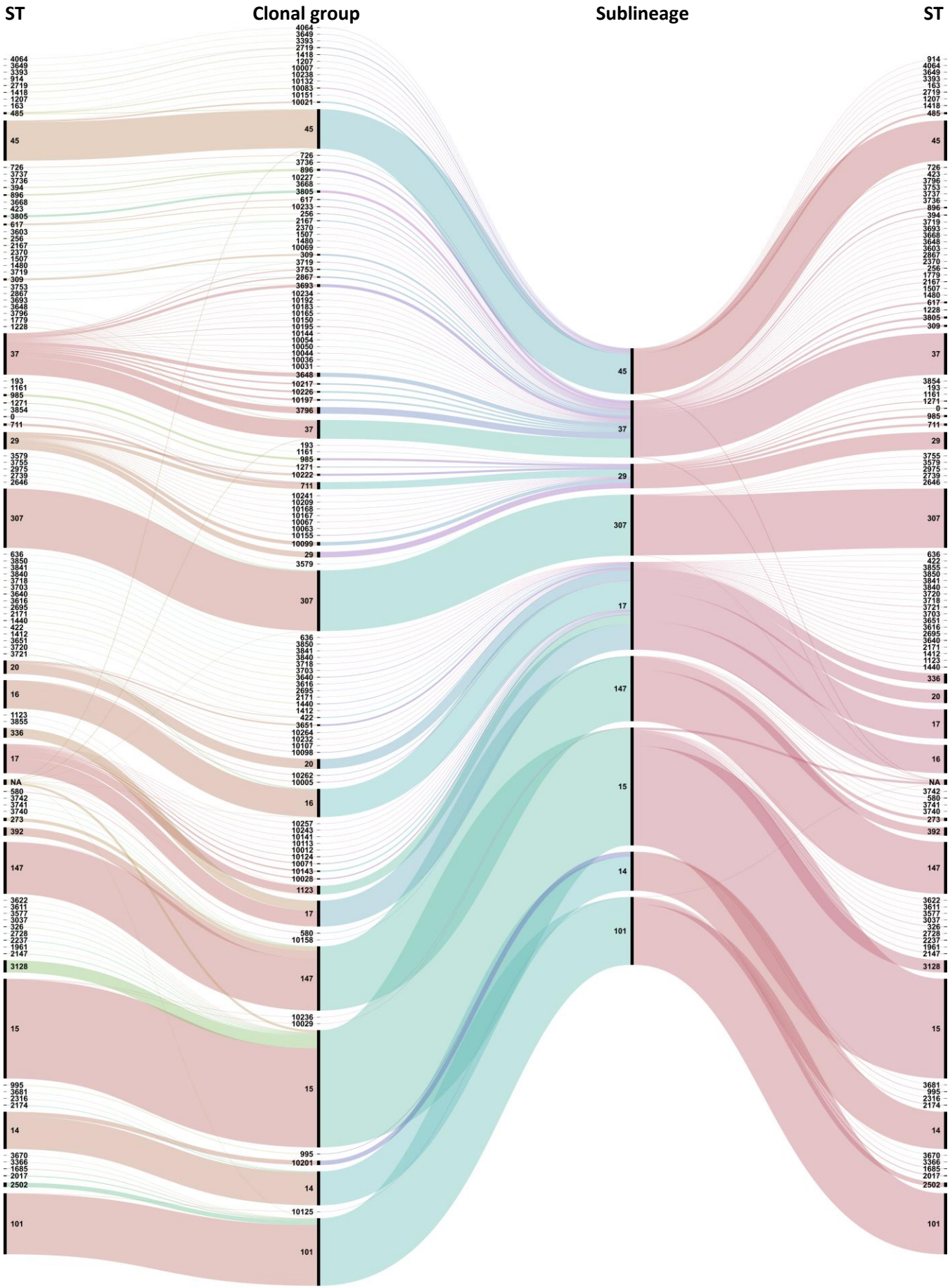
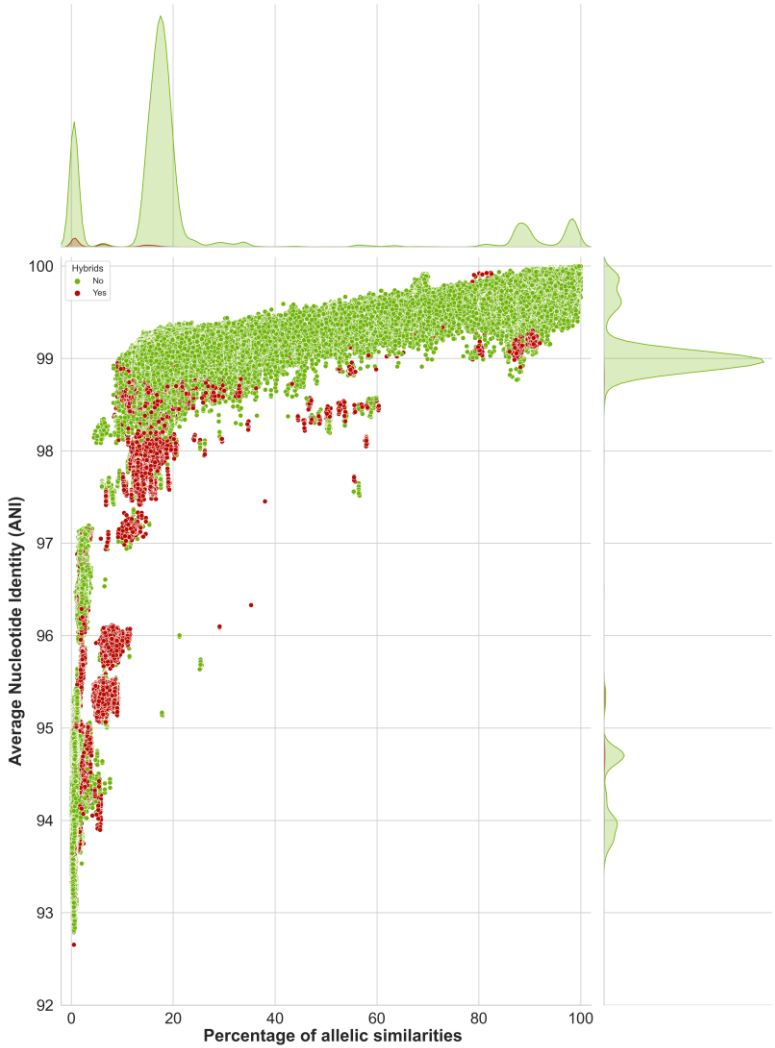


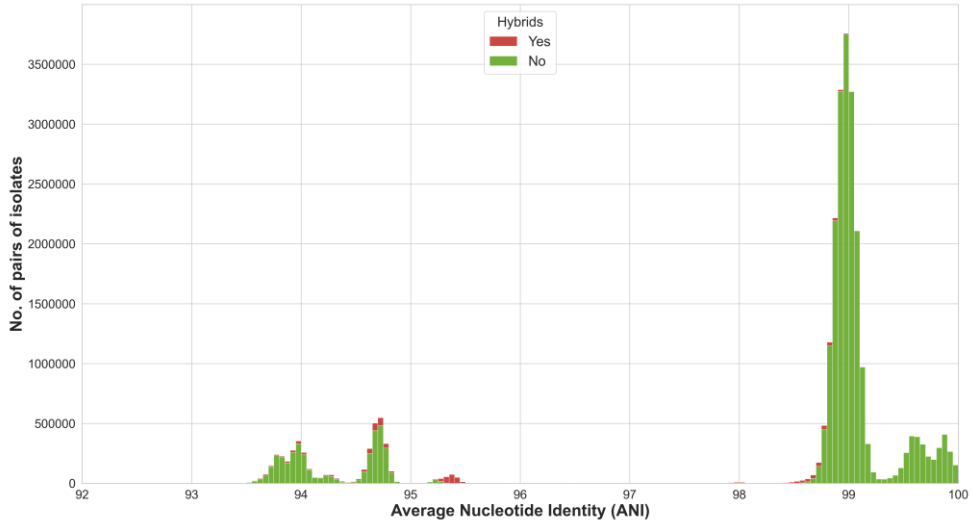
Figure S7. The distribution of pairwise distances based on Average Nucleotide Identity (ANI) and cgMLST, with hybrid genomes

(A) cgMLST similarity versus ANI. In the central panel, each point represents a pair of strains; the X-axis corresponds to the cgMLST profile similarity whereas the Y-axis corresponds to the ANI calculated on the whole genomes (FastANI v1.1); the corresponding density distributions are shown on the outside of the graph. Colors correspond to genome pairs involving only non-hybrid genomes (green) or at least one hybrid genome (red). (B) Distribution of ANI values. (C) Distribution of cgMLST similarity values.

A. cgMLST similarity *versus* ANI



B. ANI



C. cgMLST

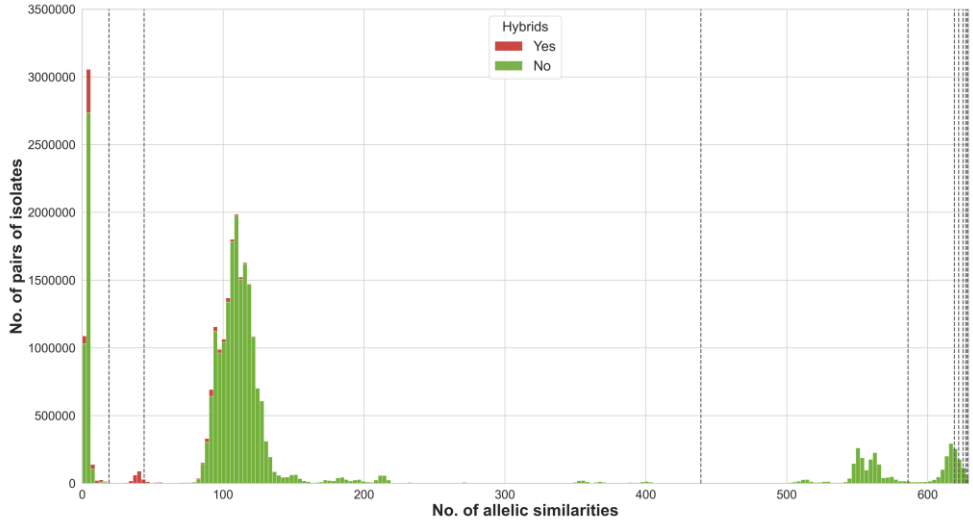
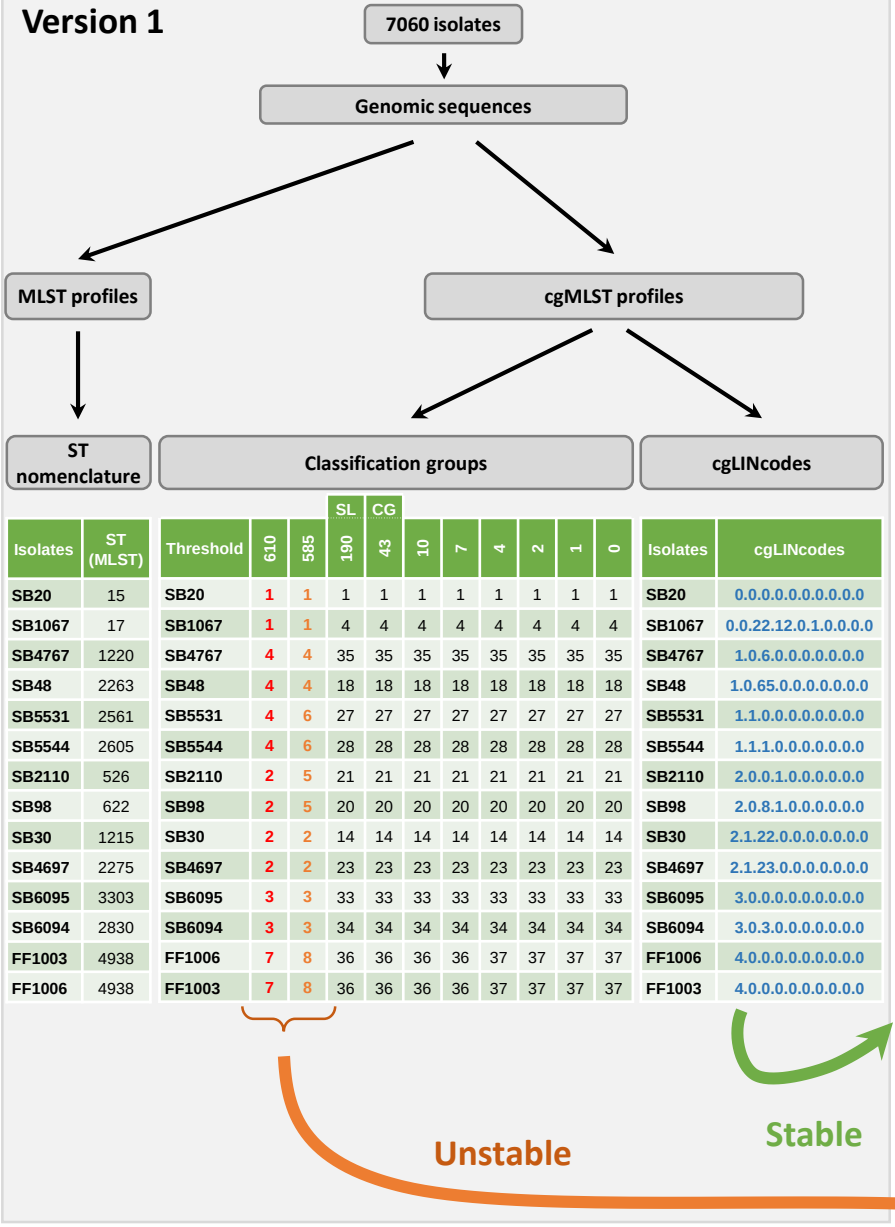


Figure S8. Impact of inter-phylogroup hybrid genomes on cgMLST classification groups

The effect of incorporating the hybrid genomes on MLSL and cgLIN codes is illustrated by comparing example genome codes before (left) and after (right) hybrid were included into the nomenclature. Version 1 of nomenclature was created from the 7060 non-hybrid genomes; version 2 was obtained after the 138 hybrid genomes were included. Note that the example MLSL genome codes are unstable, as the 610 and 585 threshold levels were affected by the incorporation of hybrids; in contrast, cgLIN codes were unaffected, as expected by design.

Figure S8. Impact of inter-phylogroup hybrid genomes on cgMLST classification groups

Version 1



Isolates	ST (MLST)
SB20	15
SB1067	17
SB4767	1220
SB48	2263
SB5531	2561
SB5544	2605
SB2110	526
SB98	622
SB30	1215
SB4697	2275
SB6095	3303
SB6094	2830
FF1003	4938
FF1006	4938

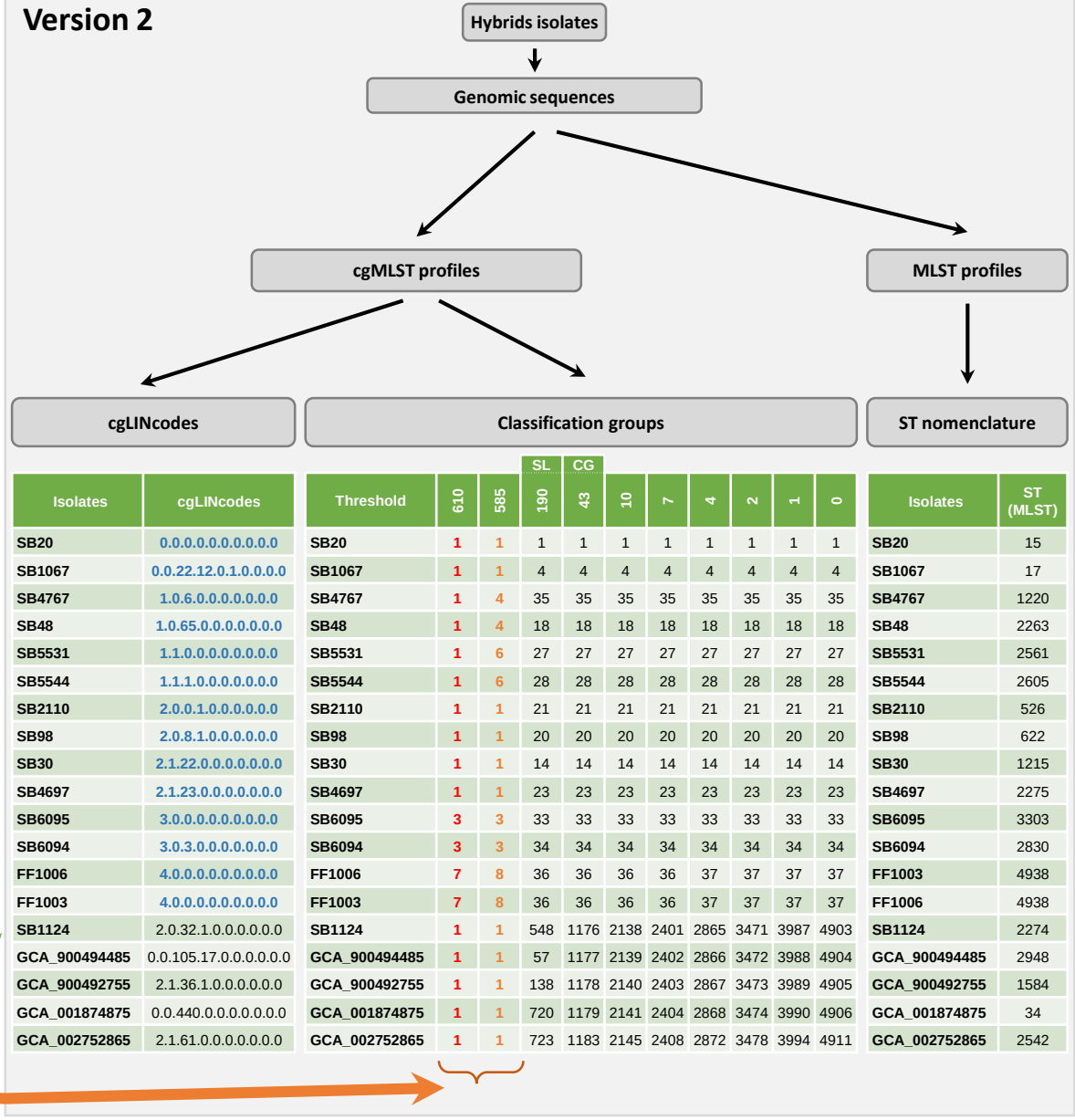
Threshold	610	585	SL 190	CG 43	10	7	4	2	1	0
SB20	1	1	1	1	1	1	1	1	1	1
SB1067	1	1	4	4	4	4	4	4	4	4
SB4767	4	4	35	35	35	35	35	35	35	35
SB48	4	4	18	18	18	18	18	18	18	18
SB5531	4	6	27	27	27	27	27	27	27	27
SB5544	4	6	28	28	28	28	28	28	28	28
SB2110	2	5	21	21	21	21	21	21	21	21
SB98	2	5	20	20	20	20	20	20	20	20
SB30	2	2	14	14	14	14	14	14	14	14
SB4697	2	2	23	23	23	23	23	23	23	23
SB6095	3	3	33	33	33	33	33	33	33	33
SB6094	3	3	34	34	34	34	34	34	34	34
FF1006	7	8	36	36	36	36	37	37	37	37
FF1003	7	8	36	36	36	36	37	37	37	37

Isolates	cgLINcodes
SB20	0.0.0.0.0.0.0.0.0.0
SB1067	0.0.22.12.0.1.0.0.0.0
SB4767	1.0.6.0.0.0.0.0.0.0
SB48	1.0.65.0.0.0.0.0.0.0
SB5531	1.1.0.0.0.0.0.0.0.0
SB5544	1.1.1.0.0.0.0.0.0.0
SB2110	2.0.0.1.0.0.0.0.0.0
SB98	2.0.8.1.0.0.0.0.0.0
SB30	2.1.22.0.0.0.0.0.0.0
SB4697	2.1.23.0.0.0.0.0.0.0
SB6095	3.0.0.0.0.0.0.0.0.0
SB6094	3.0.3.0.0.0.0.0.0.0
FF1006	4.0.0.0.0.0.0.0.0.0
FF1003	4.0.0.0.0.0.0.0.0.0

Unstable

Stable

Version 2



Isolates	cgLINcodes
SB20	0.0.0.0.0.0.0.0.0.0
SB1067	0.0.22.12.0.1.0.0.0.0
SB4767	1.0.6.0.0.0.0.0.0.0
SB48	1.0.65.0.0.0.0.0.0.0
SB5531	1.1.0.0.0.0.0.0.0.0
SB5544	1.1.1.0.0.0.0.0.0.0
SB2110	2.0.0.1.0.0.0.0.0.0
SB98	2.0.8.1.0.0.0.0.0.0
SB30	2.1.22.0.0.0.0.0.0.0
SB4697	2.1.23.0.0.0.0.0.0.0
SB6095	3.0.0.0.0.0.0.0.0.0
SB6094	3.0.3.0.0.0.0.0.0.0
FF1006	4.0.0.0.0.0.0.0.0.0
FF1003	4.0.0.0.0.0.0.0.0.0
SB1124	2.0.32.1.0.0.0.0.0.0
GCA_900494485	0.0.105.17.0.0.0.0.0.0
GCA_900492755	2.1.36.1.0.0.0.0.0.0
GCA_001874875	0.0.440.0.0.0.0.0.0.0
GCA_002752865	2.1.61.0.0.0.0.0.0.0

Threshold	610	585	SL 190	CG 43	10	7	4	2	1	0
SB20	1	1	1	1	1	1	1	1	1	1
SB1067	1	1	4	4	4	4	4	4	4	4
SB4767	1	4	35	35	35	35	35	35	35	35
SB48	1	4	18	18	18	18	18	18	18	18
SB5531	1	6	27	27	27	27	27	27	27	27
SB5544	1	6	28	28	28	28	28	28	28	28
SB2110	1	1	21	21	21	21	21	21	21	21
SB98	1	1	20	20	20	20	20	20	20	20
SB30	1	1	14	14	14	14	14	14	14	14
SB4697	1	1	23	23	23	23	23	23	23	23
SB6095	3	3	33	33	33	33	33	33	33	33
SB6094	3	3	34	34	34	34	34	34	34	34
FF1006	7	8	36	36	36	36	37	37	37	37
FF1003	7	8	36	36	36	36	37	37	37	37
SB1124	1	1	548	1176	2138	2401	2865	3471	3987	4903
GCA_900494485	1	1	57	1177	2139	2402	2866	3472	3988	4904
GCA_900492755	1	1	138	1178	2140	2403	2867	3473	3989	4905
GCA_001874875	1	1	720	1179	2141	2404	2868	3474	3990	4906
GCA_002752865	1	1	723	1183	2145	2408	2872	3478	3994	4911

Isolates	ST (MLST)
SB20	15
SB1067	17
SB4767	1220
SB48	2263
SB5531	2561
SB5544	2605
SB2110	526
SB98	622
SB30	1215
SB4697	2275
SB6095	3303
SB6094	2830
FF1003	4938
FF1006	4938
SB1124	2274
GCA_900494485	2948
GCA_900492755	1584
GCA_001874875	34
GCA_002752865	2542

Figure S9. Virulence and resistance scores of major sublineages

Left: Heatmap of the percentage of strains classified by virulence and resistance scores, broken down by sublineage; middle panel: Median of the virulence and resistance scores (the scale for the virulence score is negativized). Right panel: number of strains present in each sublineage

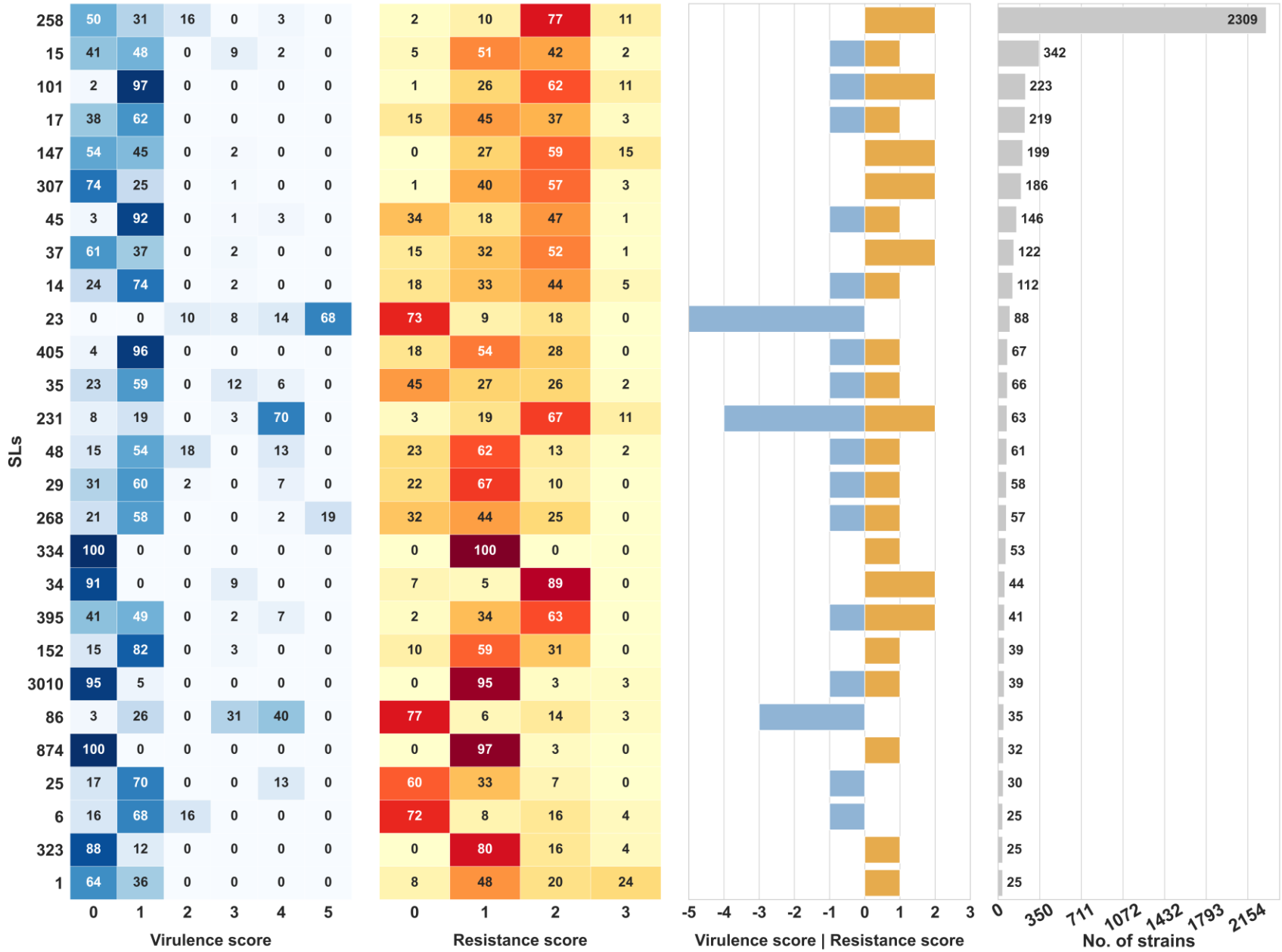


Figure S10. Impact of input order on the number of partitions in the resulting LIN codes

The variation in the number of partitions at a given threshold, as defined by the number of distinct prefixes, was quantified. For each threshold ranging from 1% to 99%, a cgLIN encoding was defined, and the 7,060 high-quality non-hybrid cgMLST profiles were encoded 500 times with random input orders. Blue: number of partitions created; red: the variance of this number.

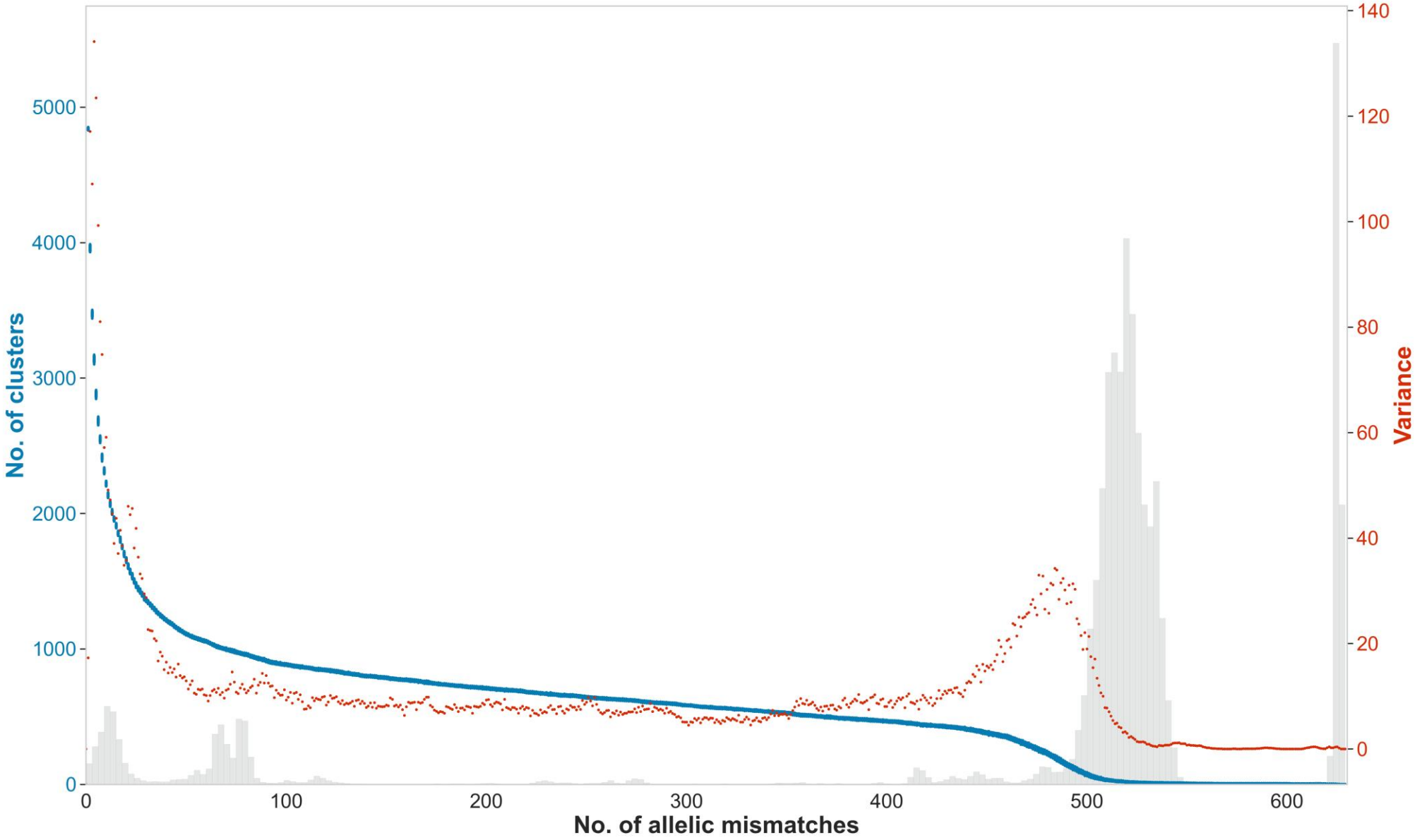
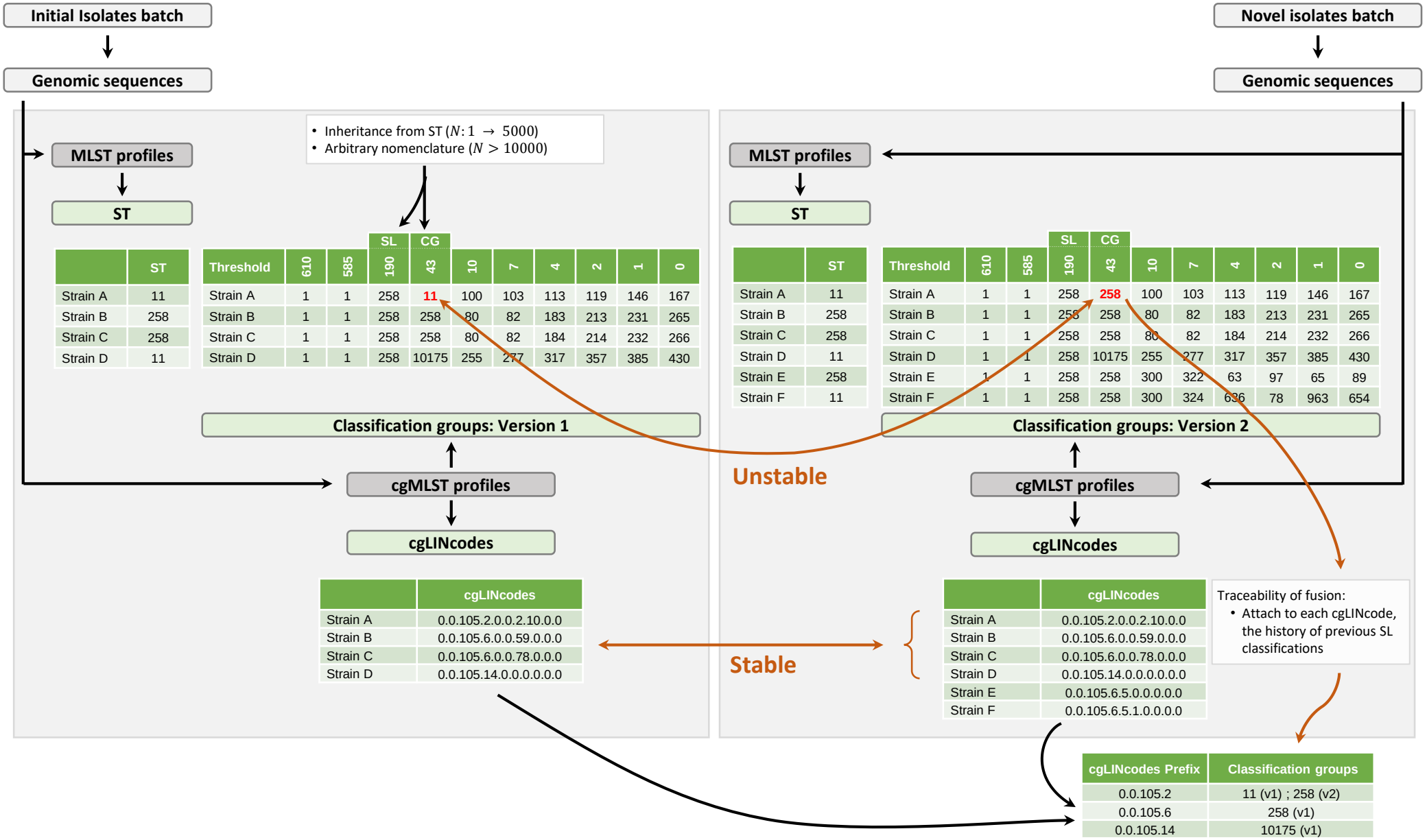


Figure S11. Relationships between ST, cgMLST and cgLIN codes, and their behavior upon novel genomes inclusion

From the cgMLST profiles, MLSL classification groups and cgLIN codes are generated. On the right, a second batch was submitted, leading to MLST classification groups to evolve, unlike the cgLIN codes, which are stable. A table of correspondence between MLSL classification and cgLIN codes can be used to follow MLSL classifications attached to each cgLIN code over time.



	ST
Strain A	11
Strain B	258
Strain C	258
Strain D	11

Threshold	610	585	SL 190	CG 43	10	7	4	2	1	0
Strain A	1	1	258	11	100	103	113	119	146	167
Strain B	1	1	258	258	80	82	183	213	231	265
Strain C	1	1	258	258	80	82	184	214	232	266
Strain D	1	1	258	10175	255	277	317	357	385	430

	cgLINcodes
Strain A	0.0.105.2.0.0.2.10.0.0
Strain B	0.0.105.6.0.0.59.0.0.0
Strain C	0.0.105.6.0.0.78.0.0.0
Strain D	0.0.105.14.0.0.0.0.0.0

	ST
Strain A	11
Strain B	258
Strain C	258
Strain D	11
Strain E	258
Strain F	11

Threshold	610	585	SL 190	CG 43	10	7	4	2	1	0
Strain A	1	1	258	258	100	103	113	119	146	167
Strain B	1	1	258	258	80	82	183	213	231	265
Strain C	1	1	258	258	80	82	184	214	232	266
Strain D	1	1	258	10175	255	277	317	357	385	430
Strain E	1	1	258	258	300	322	63	97	65	89
Strain F	1	1	258	258	300	324	636	78	963	654

	cgLINcodes
Strain A	0.0.105.2.0.0.2.10.0.0
Strain B	0.0.105.6.0.0.59.0.0.0
Strain C	0.0.105.6.0.0.78.0.0.0
Strain D	0.0.105.14.0.0.0.0.0.0
Strain E	0.0.105.6.5.0.0.0.0.0
Strain F	0.0.105.6.5.1.0.0.0.0

cgLINcodes Prefix	Classification groups
0.0.105.2	11 (v1) ; 258 (v2)
0.0.105.6	258 (v1)
0.0.105.14	10175 (v1)

Figure S12. Distribution of the phylogroup homogeneity index

Each graph is based on the population of strains belonging to the indicated phylogoup. The Y-axis correspond to the number of strains (genomes) ; the X-axis corresponds to the percentage of alleles per cgMLST profile attributed to the corresponding phylogroup (see Methods). An additional panel provides details for phylogroup Kp1. Red vertical lines indicate the threshold used to define genomes as 'hybrids' (none were defined in phylogroups Kp5, Kp6 and Kp7).

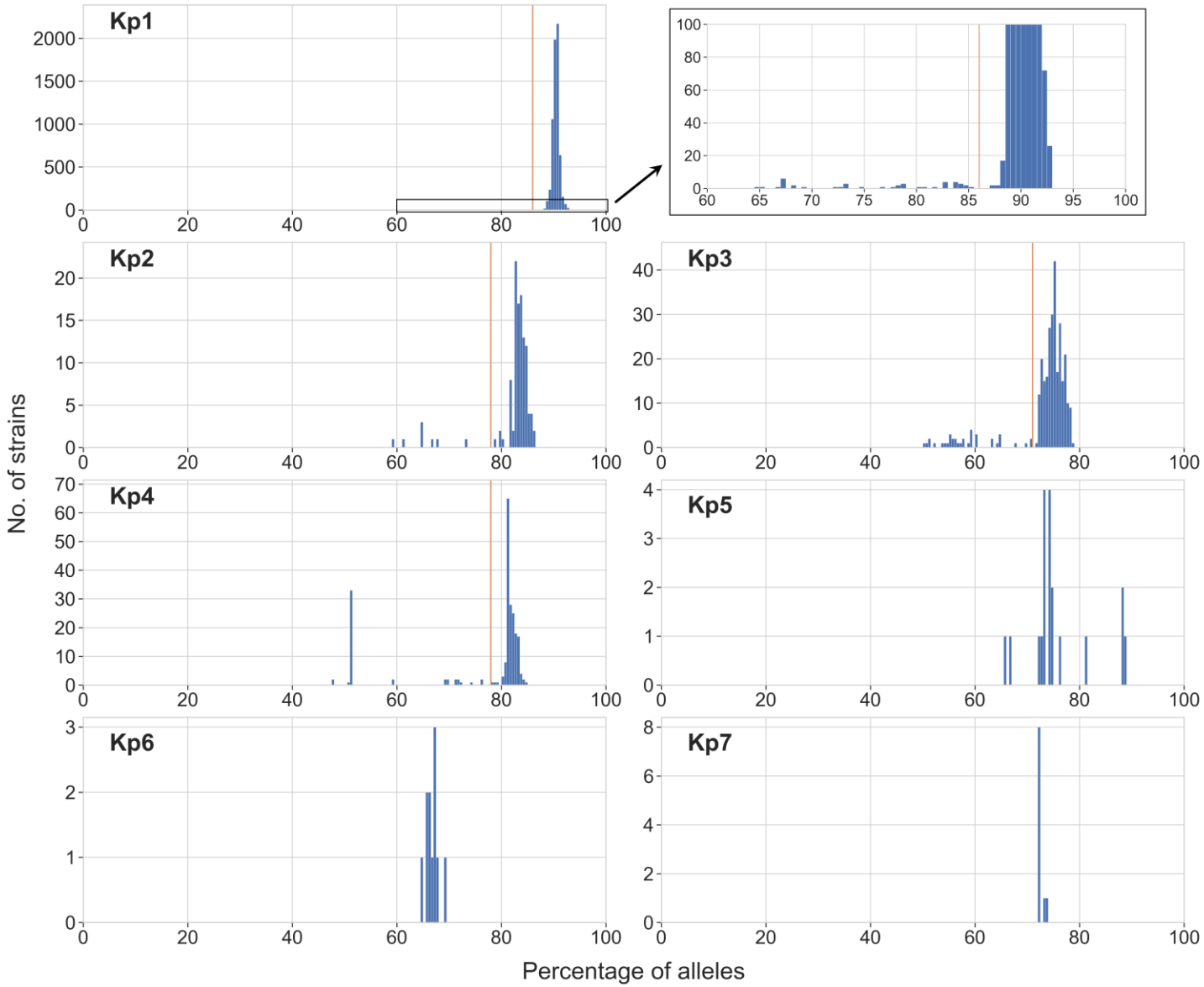


Figure S13. Principle of cgLIN code implementation

The principle is derived from Marakeby et al. (Marakeby et al., 2014), using cgMLST distances instead of ANI. Step 1: the code is initialized with the first genome being assigned the value "0" at all positions. Step 2: the encoding for an incoming genome is based on the closest genome already encoded In the chosen example, when the genome of strain B is submitted, the only genome in the database is that of strain A. Therefore, the cgLIN code of strain B is assigned based on the cgLIN code of strain A. By the time strain C is submitted, strains A and B are in the database. Based on the distance between the cgMLST profile of strain C and those present in the database, strain B is found to be the most similar. Therefore, the cgLIN code of strain C is assigned based on the cgLIN code of strain B and a value corresponding to the value in the cgLIN code of strain B, plus 1 (here, 0+1=1) is attributed to the bin corresponding to the lowest bin with a higher identity threshold that the similarity between B and C. Downstream bins are attributed “0” in the cgLIN code of strain C. Upstream bins are attributed the same value as for strain B. When D is introduced, it is coded according to strain C, its closest genome already encoded.

Figure S13. Principle of cgLIN code implementation

Step 1: initialization

Database

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
Number of match threshold	19	44	439	586	619	622	625	627	628	629
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain A	0	0	0	0	0	0	0	0	0	0

Step 2: incrementation

Lowest bin with higher identity threshold

Assignment of the cgLINcode of next strain

Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain B	Strain A	8	621	$\frac{\Omega - \beta}{\Omega} = 98.711$

Database

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain A	0	0	0	0	0	0	0	0	0	0
Strain B	0	0	0	0	0	1	0	0	0	0

Assignment of the cgLINcode of next strain

Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain C	Strain B	2	625	$\frac{\Omega - \beta}{\Omega} = 99.680$

Database

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain A	0	0	0	0	0	0	0	0	0	0
Strain B	0	0	0	0	0	1	0	0	0	0
Strain C	0	0	0	0	0	1	0	1	0	0

Assignment of the cgLINcode of next strain

Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain D	Strain C	4	627	$\frac{\Omega - \beta}{\Omega} = 99.362$

Database

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain A	0	0	0	0	0	0	0	0	0	0
Strain B	0	0	0	0	0	1	0	0	0	0
Strain C	0	0	0	0	0	1	0	1	0	0
Strain D	0	0	0	0	0	1	1	0	0	0

Figure S14. cgLIN codes implementation for nearly-identical cgMLST profiles

This figure illustrates the principle of cgLIN code implementation, when the genomes are very closely related to each other, as well as when there are missing data in the cgMLST profiles. The encoding process is similar to the general case (Figure S13). In case of complete identity and no missing data, cgLIN codes are exactly identical (see strains X and W). In cases where there is identity at all called loci, but with some loci being uncalled, the similarity would still be 100% resulting in identical cgLIN codes (not shown). In case of a single mismatch and no missing data, the cgLIN codes differ only at their last position (strains Y and X). In case of a single or several mismatches, the effect of missing data at some other loci will be to decrease the similarity ratio compared to the case where no data would be missing at other loci (strain Z therefore differs from Y already at the penultimate bin, even with a single allelic mismatch).

Figure S14. cgLIN codes implementation for nearly-identical cgMLST profiles

Sequential assignment of cgLIN codes of strains

Database

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
Number of match threshold	19	44	439	586	619	622	625	627	628	629
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain W	0	0	396	0	0	0	0	10	0	0

Strain W

Strain X

Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain X	Strain W	0	629	$\frac{\Omega - \beta}{\Omega} = 100$

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain W	0	0	396	0	0	0	0	10	0	0
Strain X	0	0	396	0	0	0	0	10	0	0

Strain Y

Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain Y	Strain X	1	629	$\frac{\Omega - \beta}{\Omega} = 99.8410$

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain W	0	0	396	0	0	0	0	10	0	0
Strain X	0	0	396	0	0	0	0	10	0	0
Strain Y	0	0	396	0	0	0	0	10	0	1

Strain Z

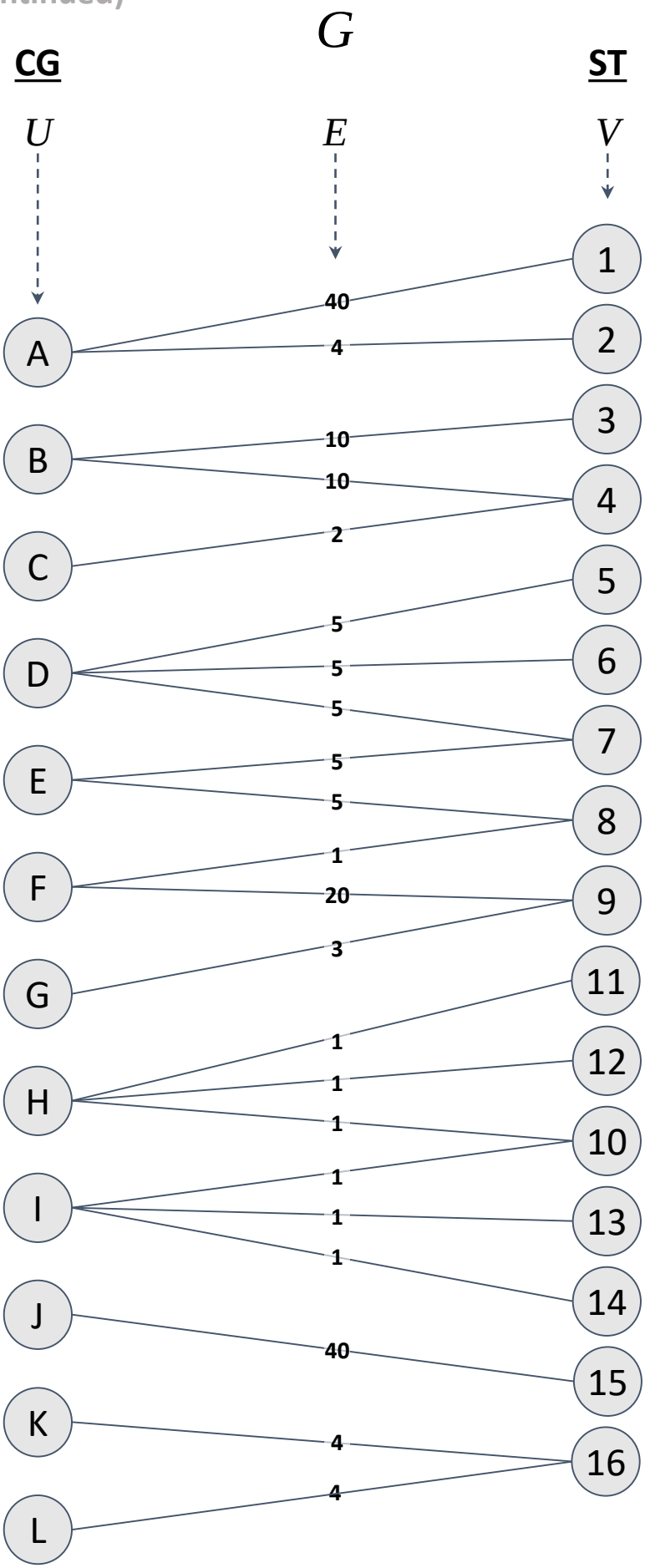
Strain	Closest genome	Number of mismatch (β)	Number of loci called in both strains (Ω)	% identity
Strain Z	Strain Y	1	$600 \leq \Omega \leq 628$	$99{,}83 \leq \frac{\Omega - \beta}{\Omega} \leq 99.840$

Number of mismatch threshold	610	585	190	43	10	7	4	2	1	0
% identity threshold	3.0207	6.9952	69.7933	93.1638	98.4102	98.8871	99.3641	99.6820	99.8410	100.000
Strain W	0	0	396	0	0	0	0	10	0	0
Strain X	0	0	396	0	0	0	0	10	0	0
Strain Y	0	0	396	0	0	0	0	10	0	1
Strain Z	0	0	396	0	0	0	0	10	1	0

Figure S15. Step-by-step illustration of the taxonomic inheritance algorithm

Given 12 hypothetical clonal group (CG, named A-L) and 16 sequence types (ST, labelled from 1 to 16), a weighted bipartite graph *G* is shown, as well as the use of *G* to label the CG based on their relation to their related ST. For eacj step illustration, the corresponding lines of the algorithm pseudo-code (see SupMat) are specified (bottom of each sub-figure).

Figure S15. (continued)

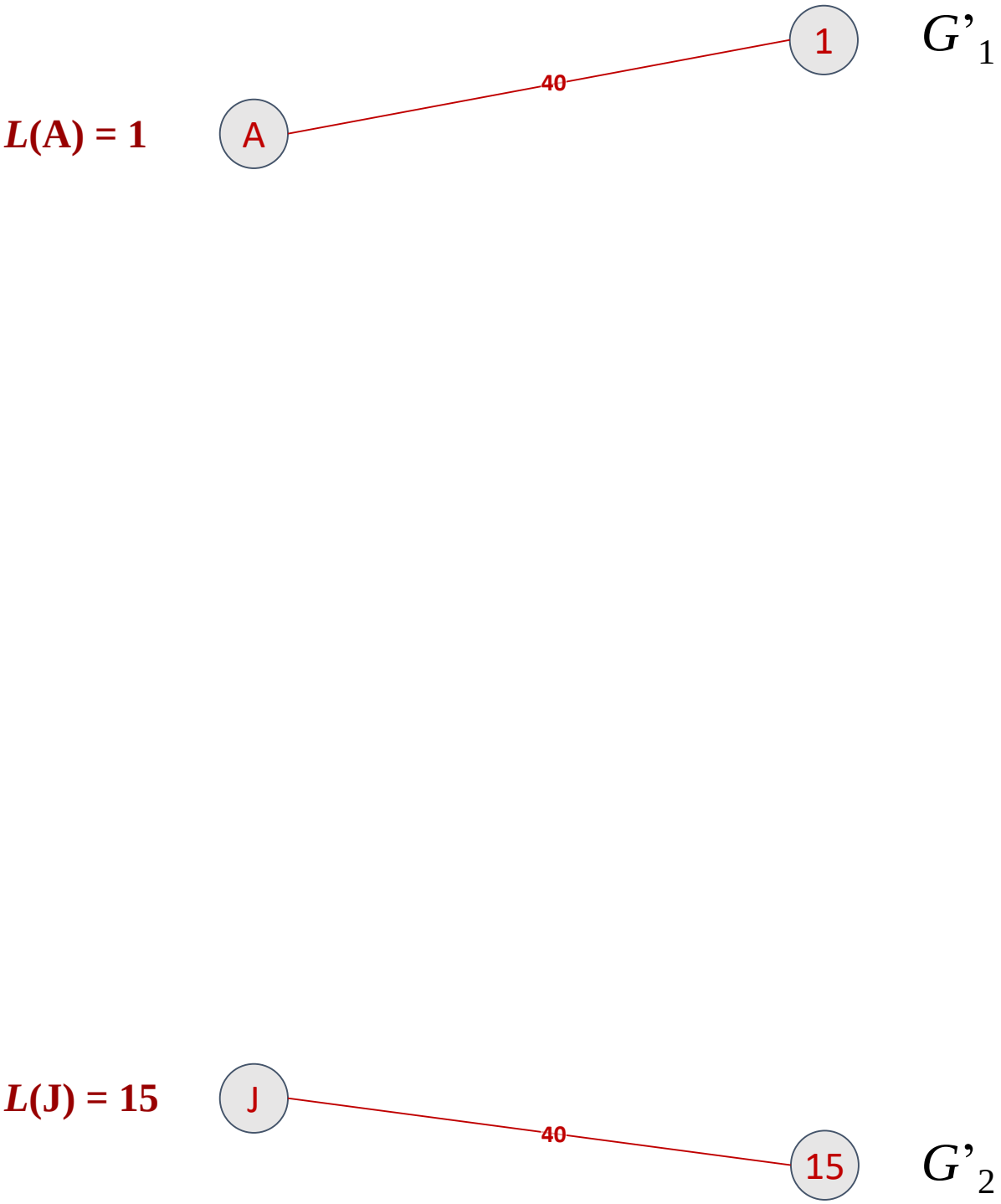


$\lambda = 16$

(a)

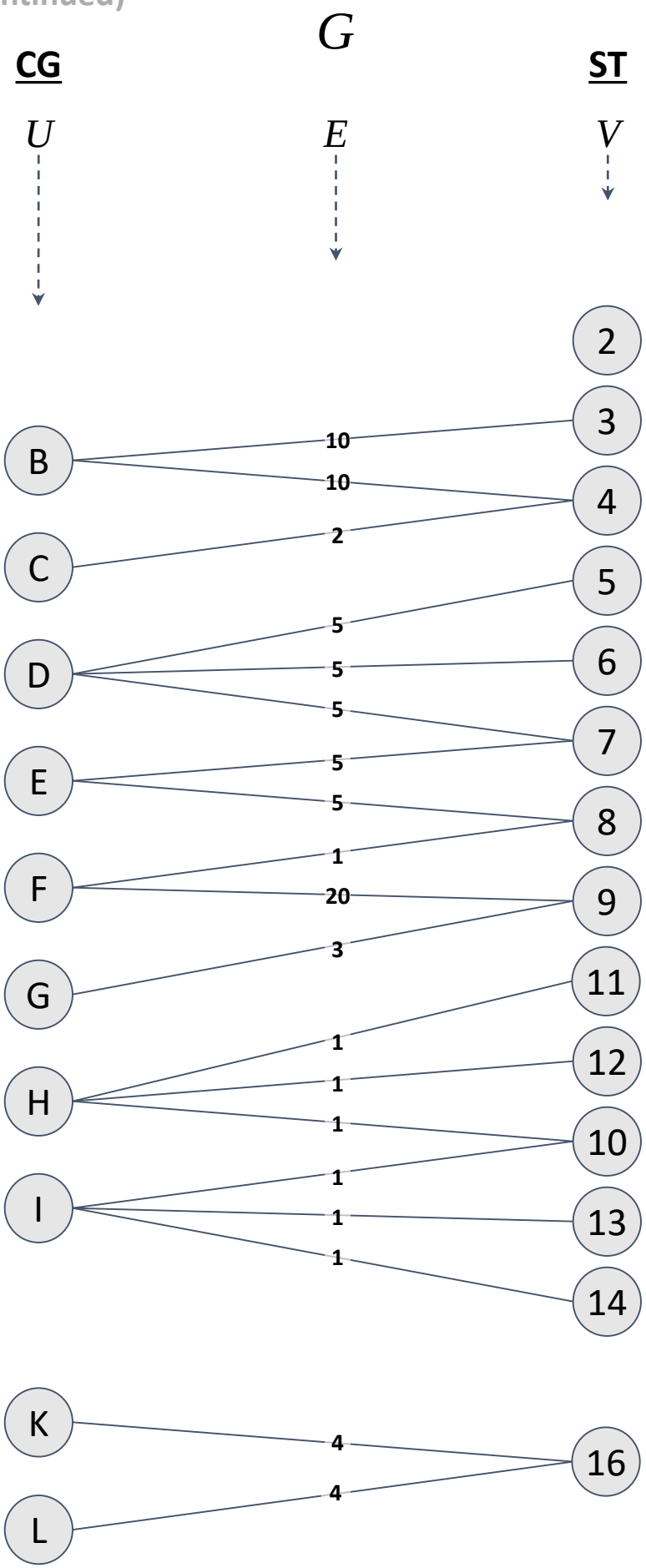
Figure S15. (continued)

$\Gamma(G)$



(c), (d), (e) and (j)

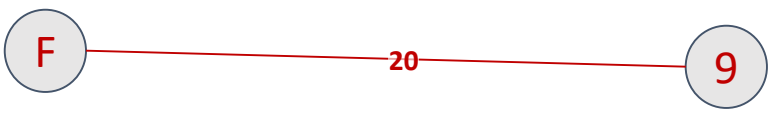
Figure S15. (continued)



(i)

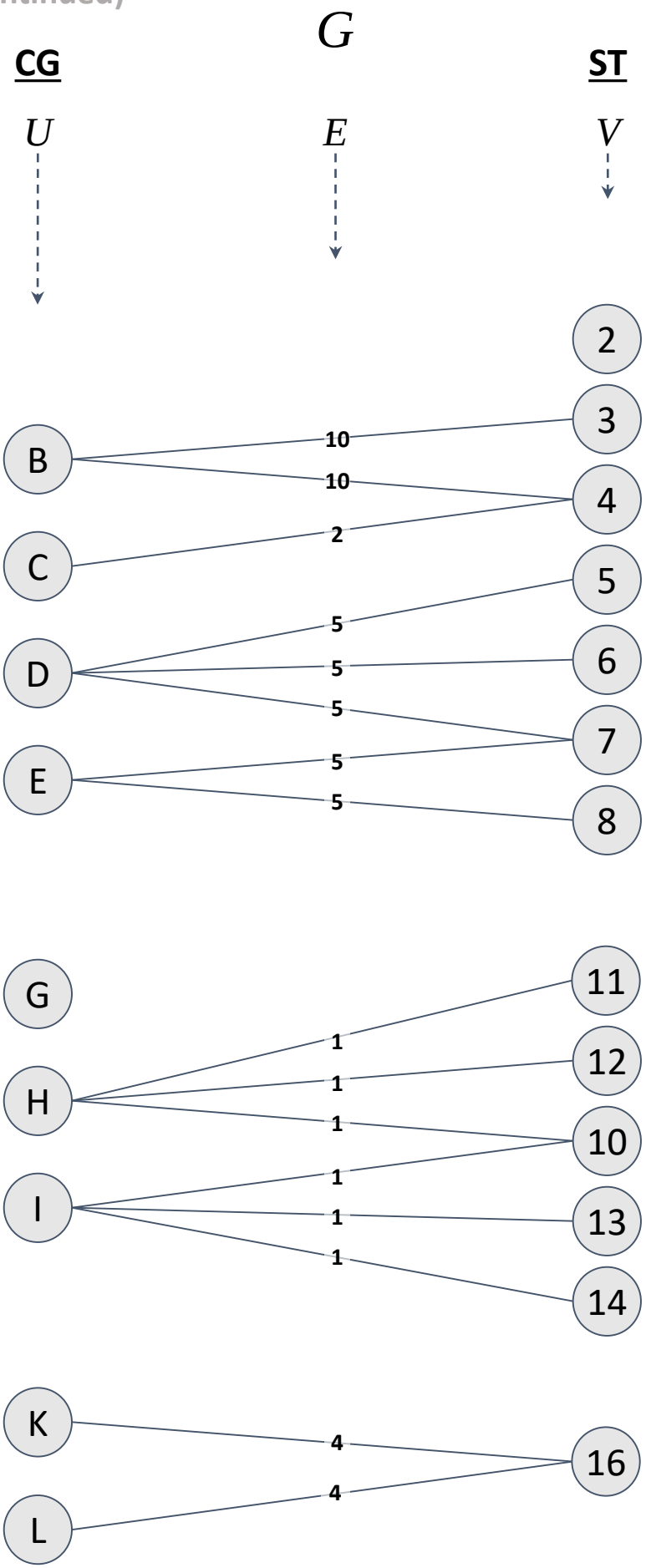
$\Gamma(G)$

$L(F) = 9$



(c), (d), (e) and (j)

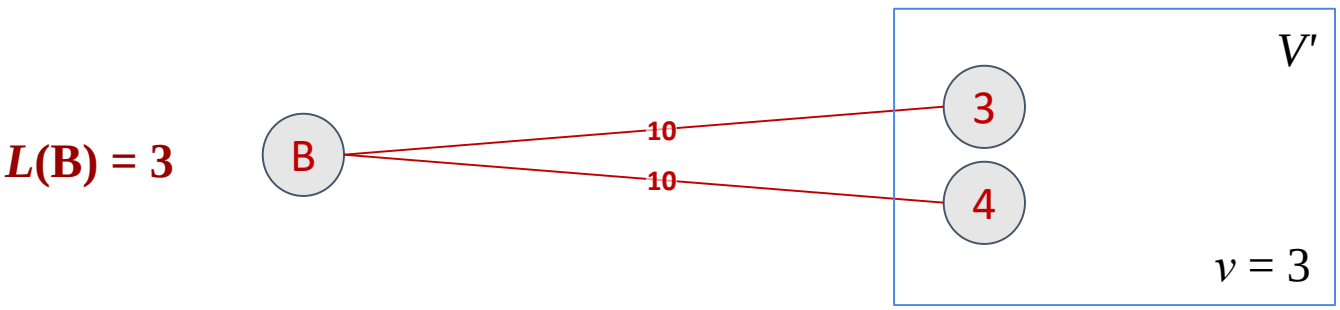
Figure S15. (continued)



(i)

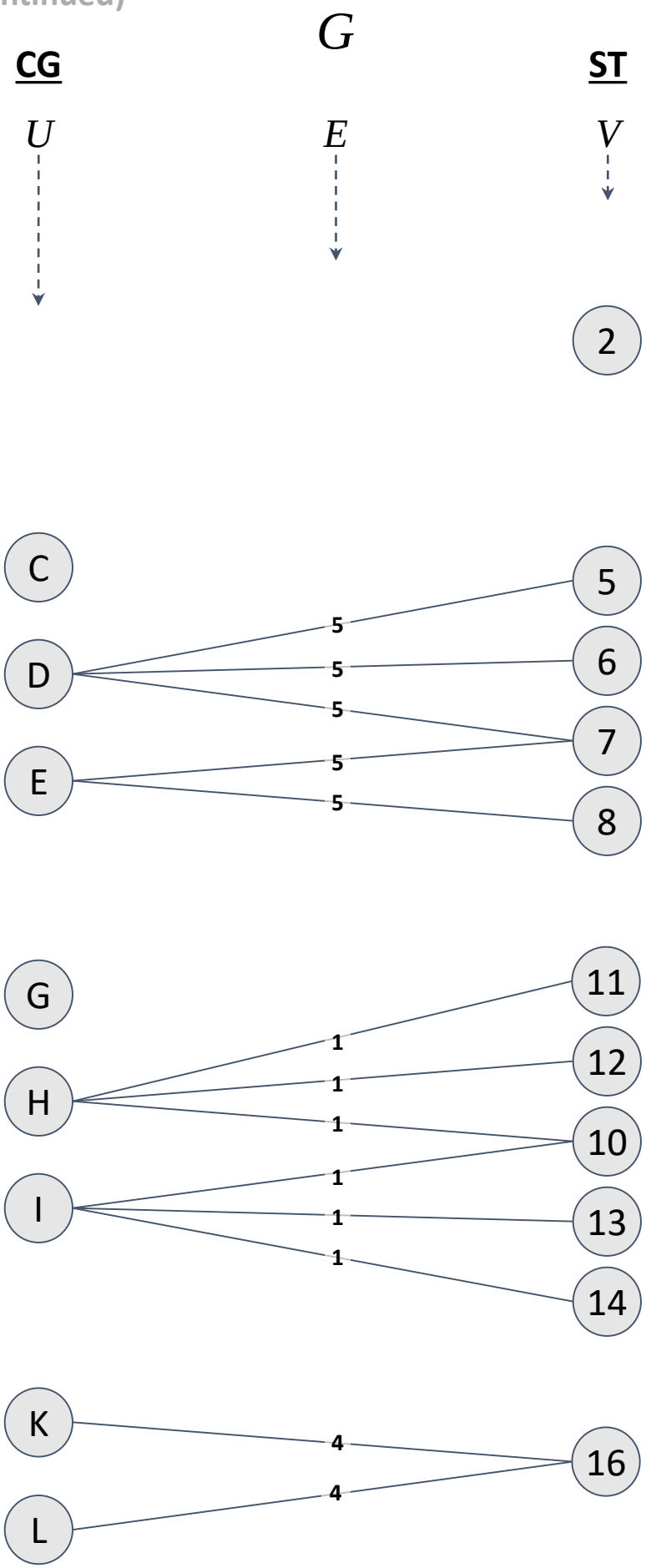
Figure S15. (continued)

$\Gamma(G)$



(c), (d), (e) and (j)

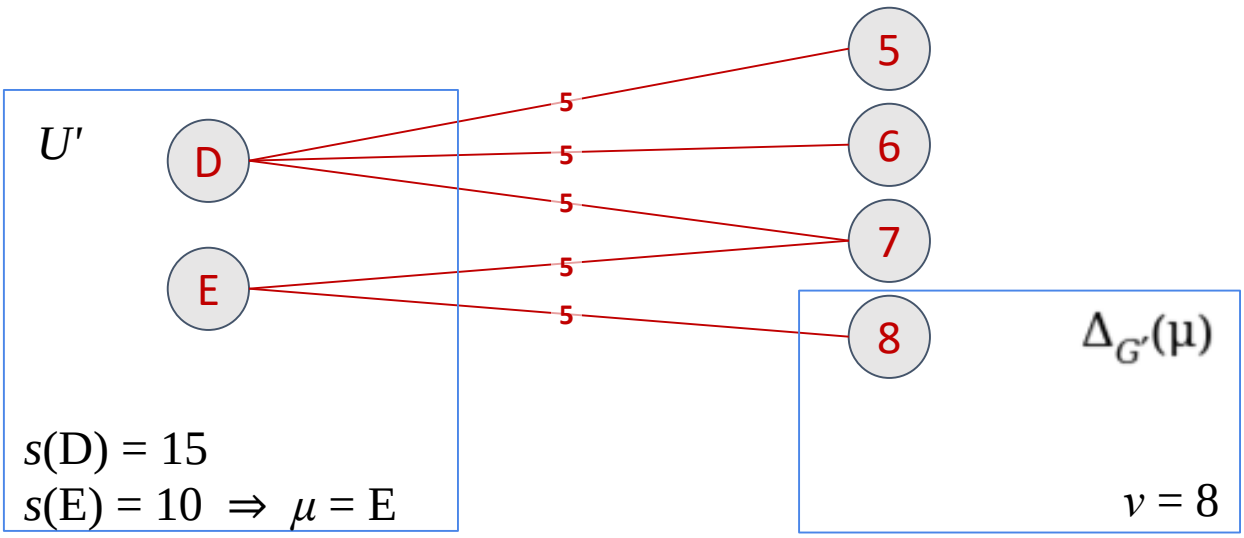
Figure S15. (continued)



(i)

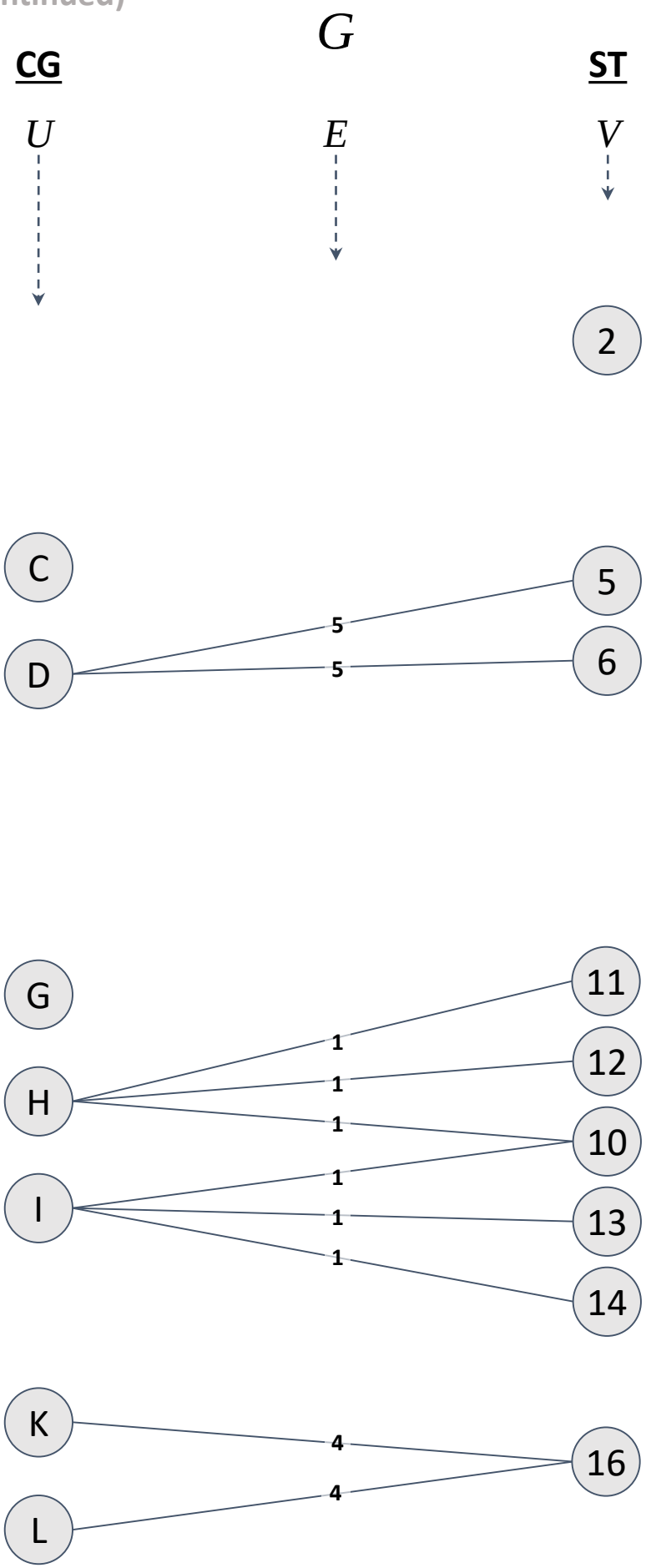
Figure S15. (continued)

$\Gamma(G)$



L(E) = 8

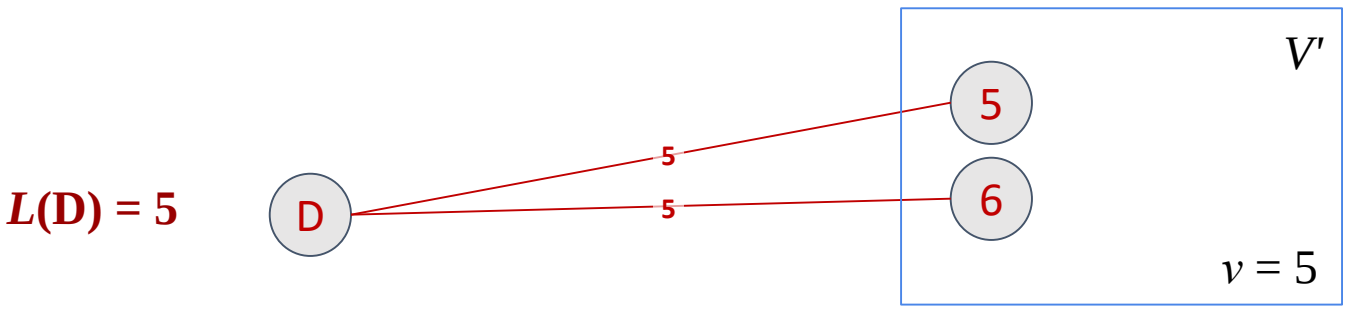
(c), (f), (i) and (j)



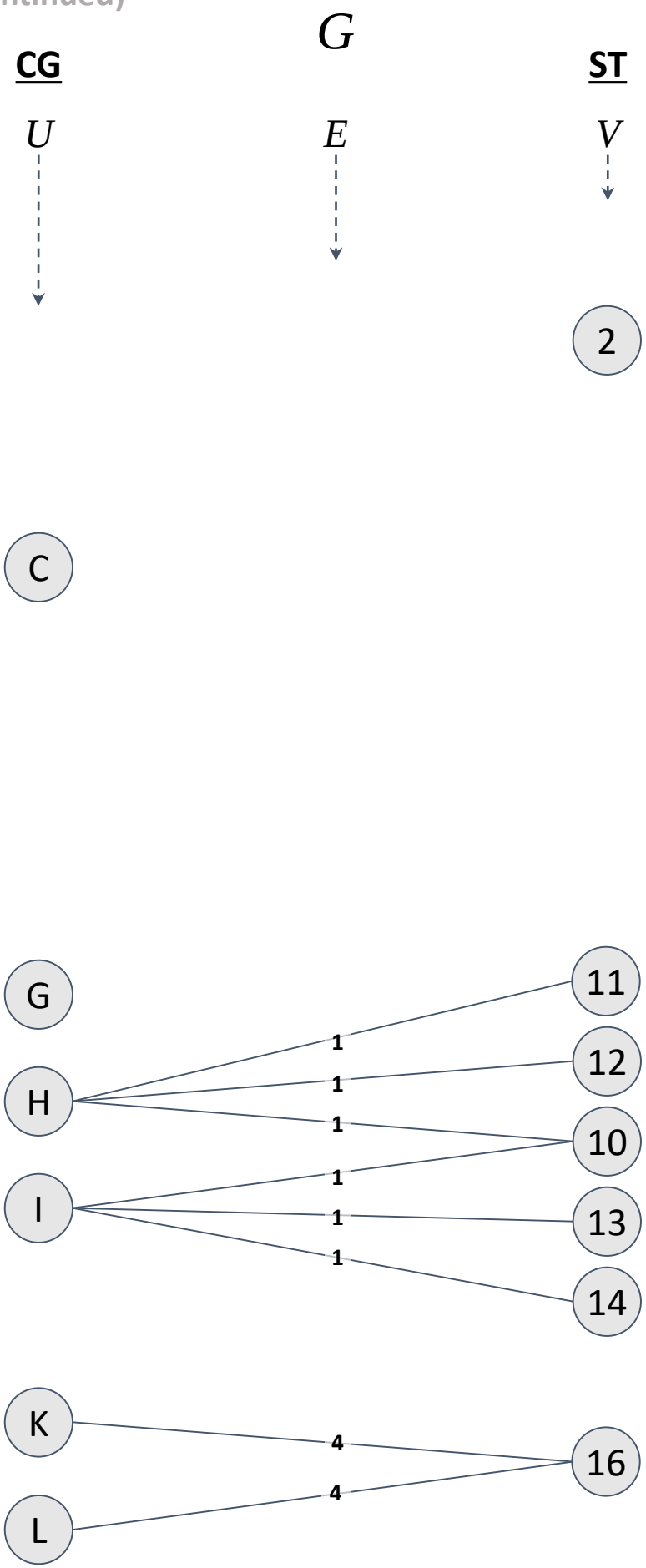
(i)

Figure S15. (continued)

$\Gamma(G)$



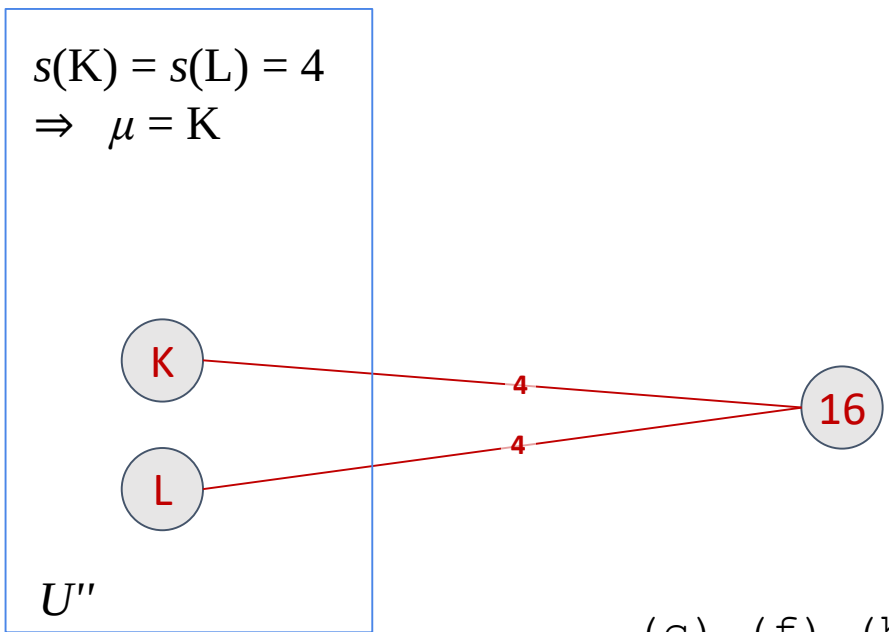
(c), (d), (e) and (j)



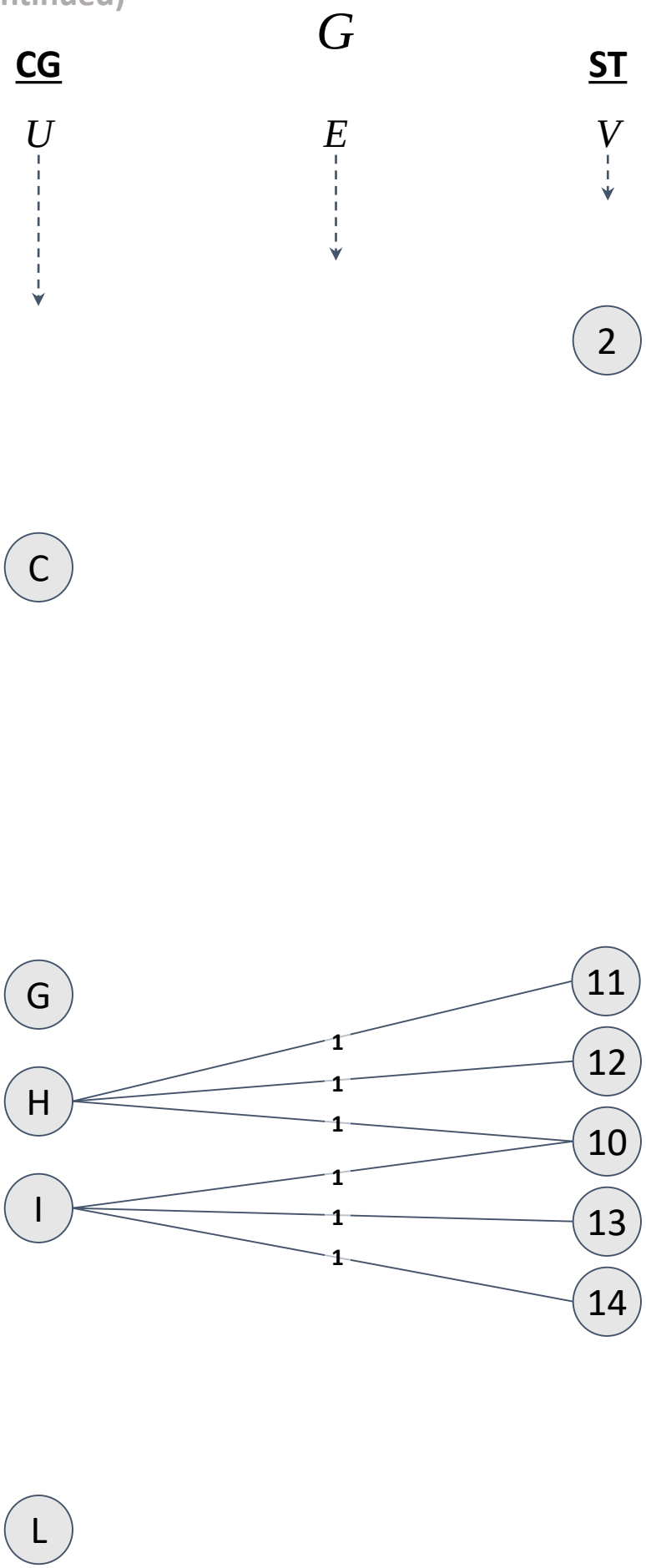
(i)

$\Gamma(G)$

L(K) = 16

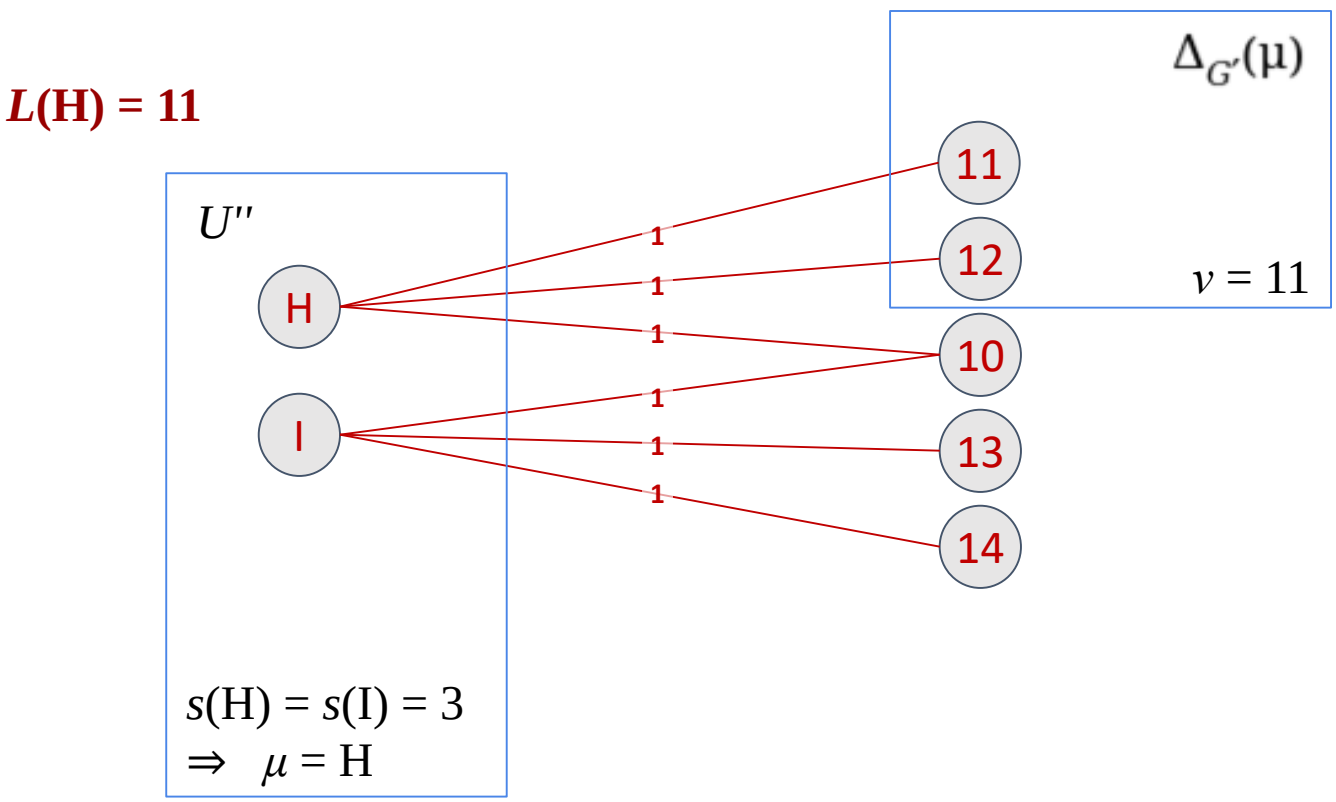


(c), (f), (h), (i) and (j)



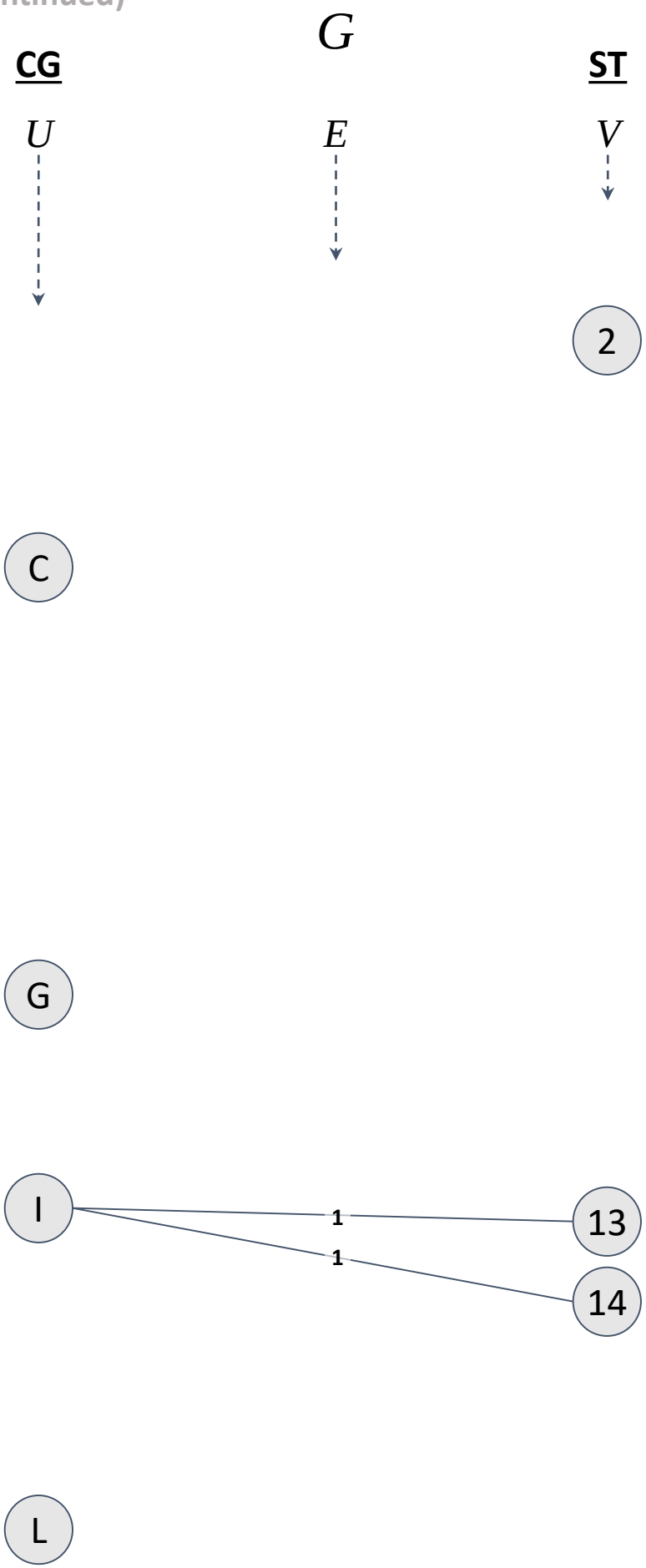
(i)

$\Gamma(G)$



(c), (f), (h), (i) and (j)

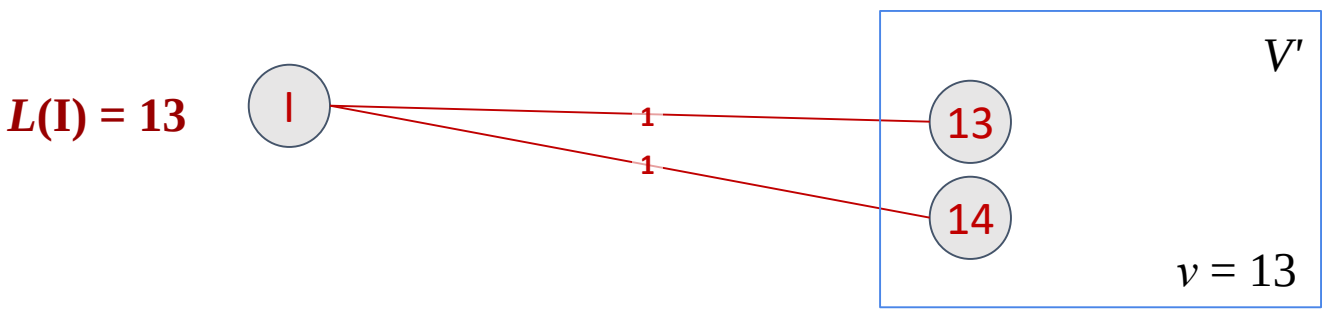
Figure S15. (continued)



(i)

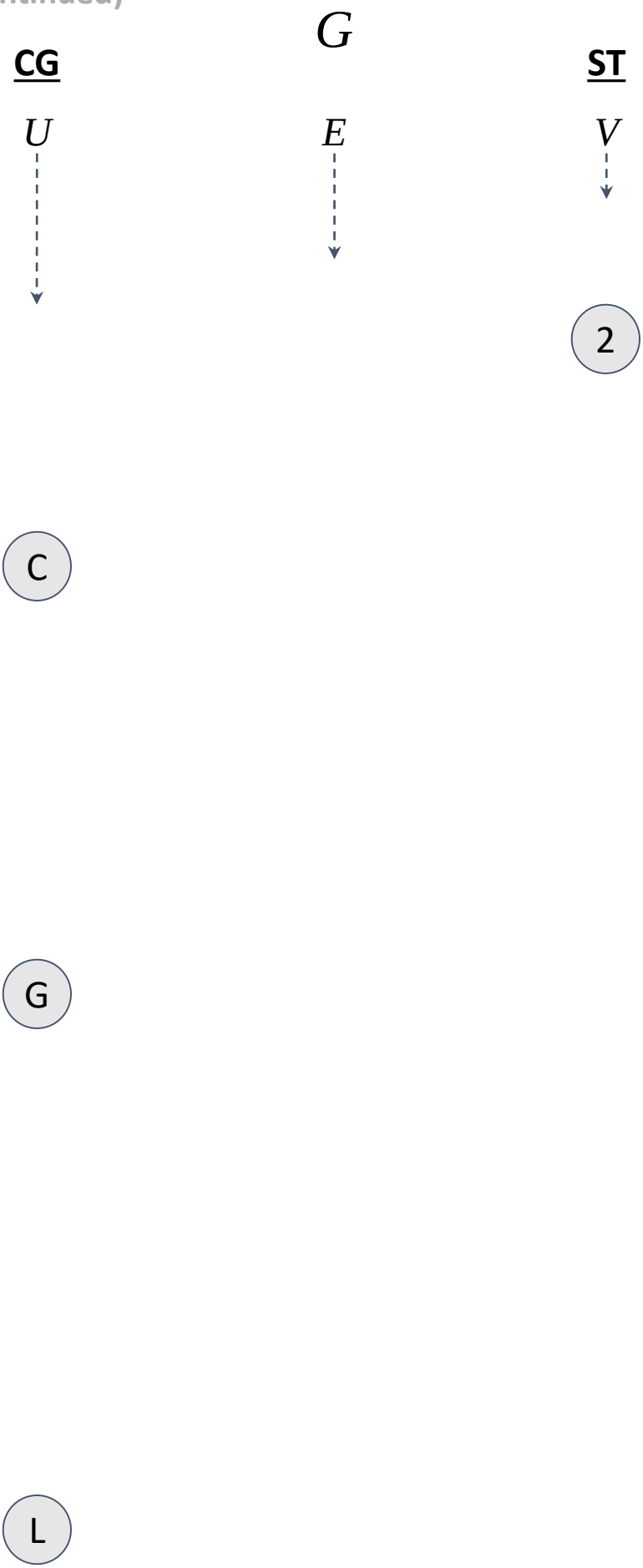
Figure S15. (continued)

$\Gamma(G)$



(c), (d), (e) and (j)

Figure S15. (continued)



(i)

$\lambda = 16$

$L(C) = 17$ C

$L(G) = 18$ G

$L(L) = 19$ L

(a), (l), (m) and (n)