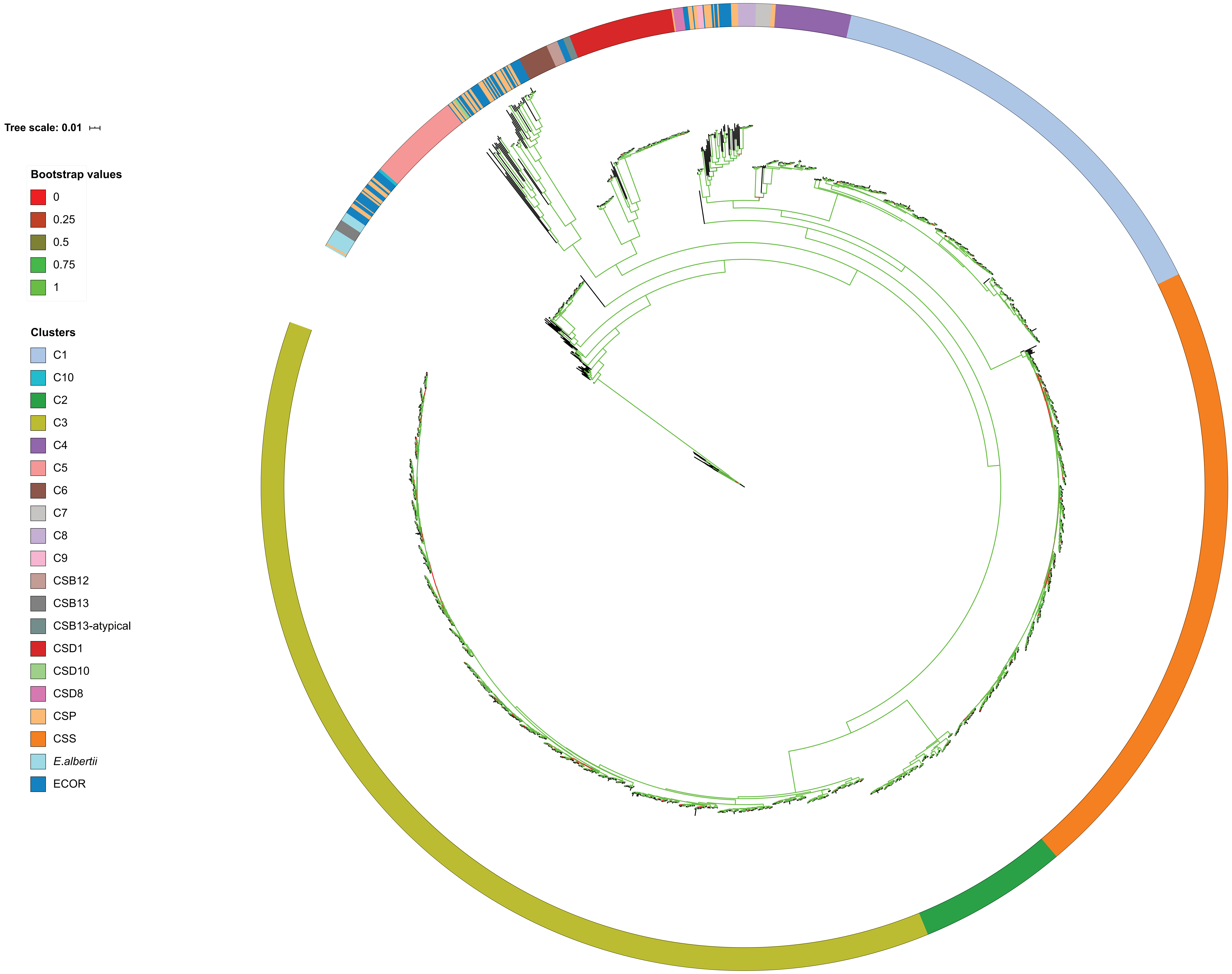

Figure S1: Identification phylogenetic tree
The identification phylogenetic tree was constructed using Quicktree v1.3 as Figure 1 was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli)* clusters were colored per cluster legend and shown as the ring. The internal branches are colored to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. Prefix CSP denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei,* S. *boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

Figure S2-A: Confirmation phylogenetic tree

The confirmation phylogenetic tree was constructed using Quicktree v1.3 based on 2,375 isolates and visualised using Grapetree's interactive mode. The tree shows the phylogenetic relationships between identified *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters in the identification dataset and non-enteroinvasive *E. coli* isolates. Branch lengths are log scale for clarity. The scalar bar represents 0.1 substitutions per site. Known *Shigella* and EIEC clusters from the identification dataset are coloured per colour legend. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. CSP prefix denotes sporadic EIEC lineages. ECOLI is *E. coli.* CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

0.1
Clusters
C1 [85]
C10 [2]
C2 [30]
C3 [167]
C4 [15]
C5 [21]
C6 [12]
C7 [7]
C8 [8]
C9 [4]
CSB12 [4]
CSB13 [7]
CSB13-atypical [3]
CSD1 [8]
CSD10 [2]
CSD8 [7]
CSP [58]
CSS [45]
ECOLI [1890]

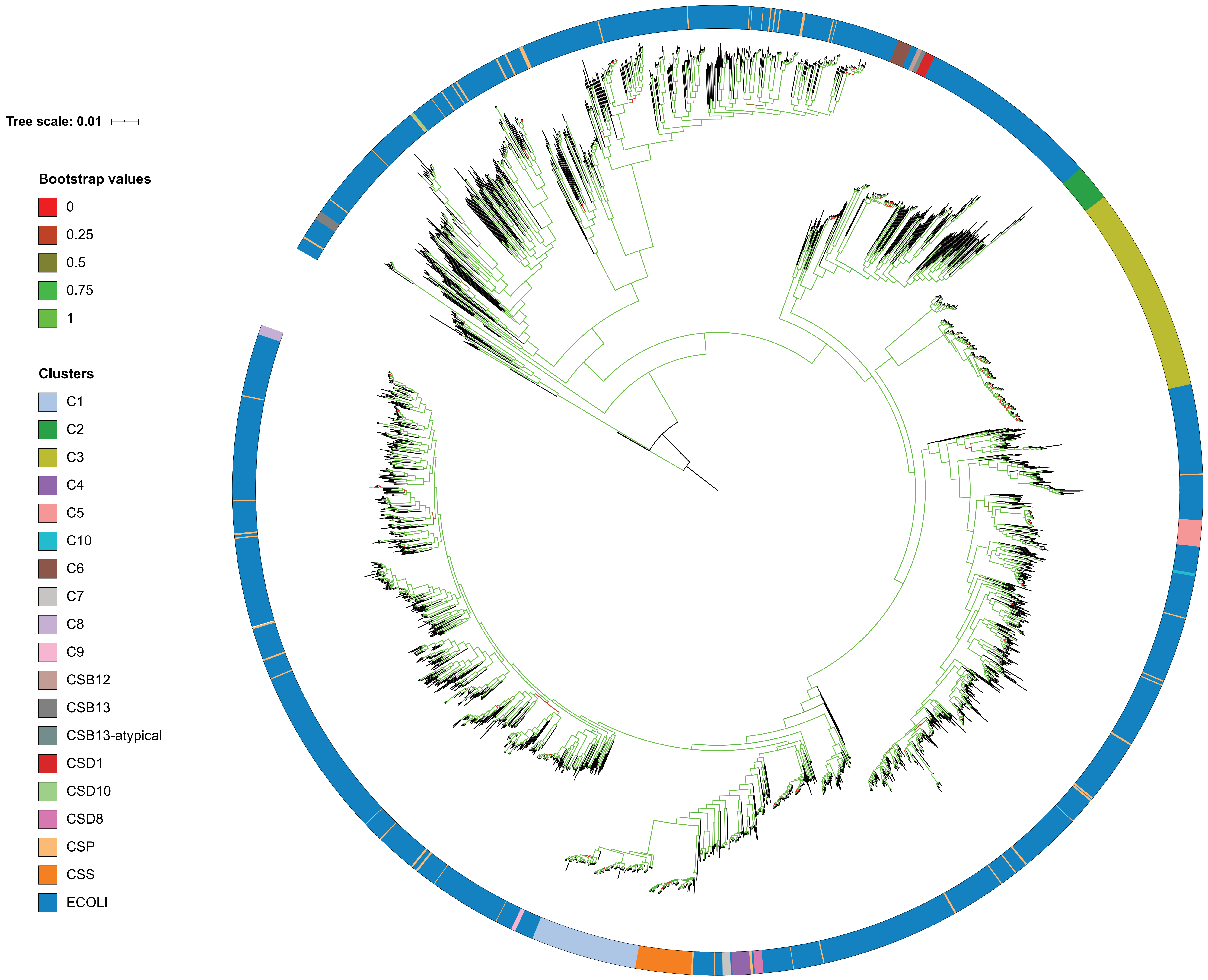

Figure S2-B: Confirmation phylogenetic tree

The confirmation phylogenetic tree constructed using Quicktree v1.3 as Figure S2-A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive E. coli) clusters were colored per cluster legend and shown as the ring. The internal branches are colored to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. Prefix CSP denotes sporadic EIEC lineages. ECOLI is *E. coli.* CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and S. *dysenteriae* serotype 10 respectively.

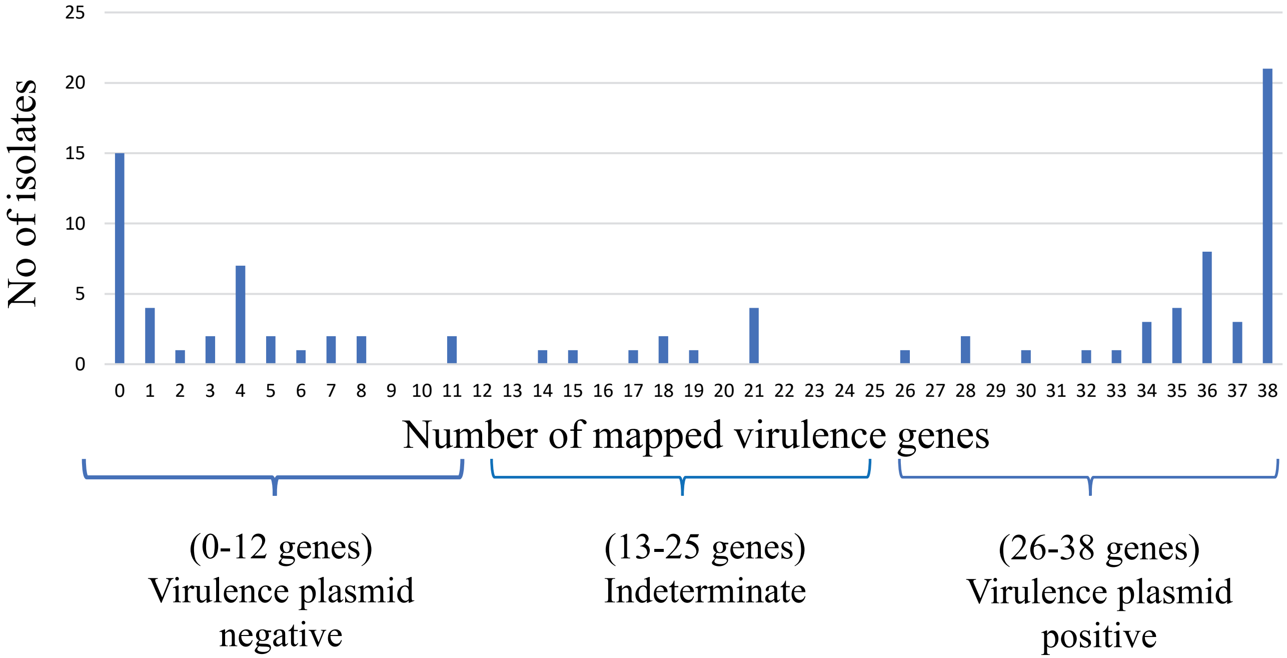

Figure S3: Distribution of the number of genes out of 38 virulence genes mapped in the 59 sporadic isolates
The presence of *Shigella* virulence plasmid pINV in the 59 sporadic isolates in identification dataset was determined by the mapping of reads of an isolate to the 38 virulence genes selected (The 38 virulence genes are listed in “Analysis of the 59 sporadic EIEC isolates” section in the main text). Details were described in Results “Investigation of *Shigella* virulence plasmid pINV in 59 sporadic isolates”. Three categories were defined based on the number of virulence genes mapped for an isolate: virulence plasmid positive: > 25 genes mapped to isolate; virulence plasmid indeterminate: 13 to 25 genes mapped to isolate; virulence plasmid negative: less than 13 genes mapped to isolate.

C3 [2]
CSS [3]
C1 [1,047]
S4-Tree 1A
● C1 [85]
● C10 [2]
● C2 [30]
● C3 [168]
● C4 [15]
● C5 [20]
● C6 [12]
● C7 [7]
● C8 [8]
● C9 [4]
● CSB12 [4]
● CSB13 [6]
● CSB13-atypical [3]
● CSD1 [8]
● CSD10 [2]
● CSD8 [7]
● CSP [59]
● CSS [45]
● *E. albertii* [18]
● ECOR [72]
○ valdb [1,052]
0.02

Figure S4-Tree 1A: Validation phylogenetic tree

The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

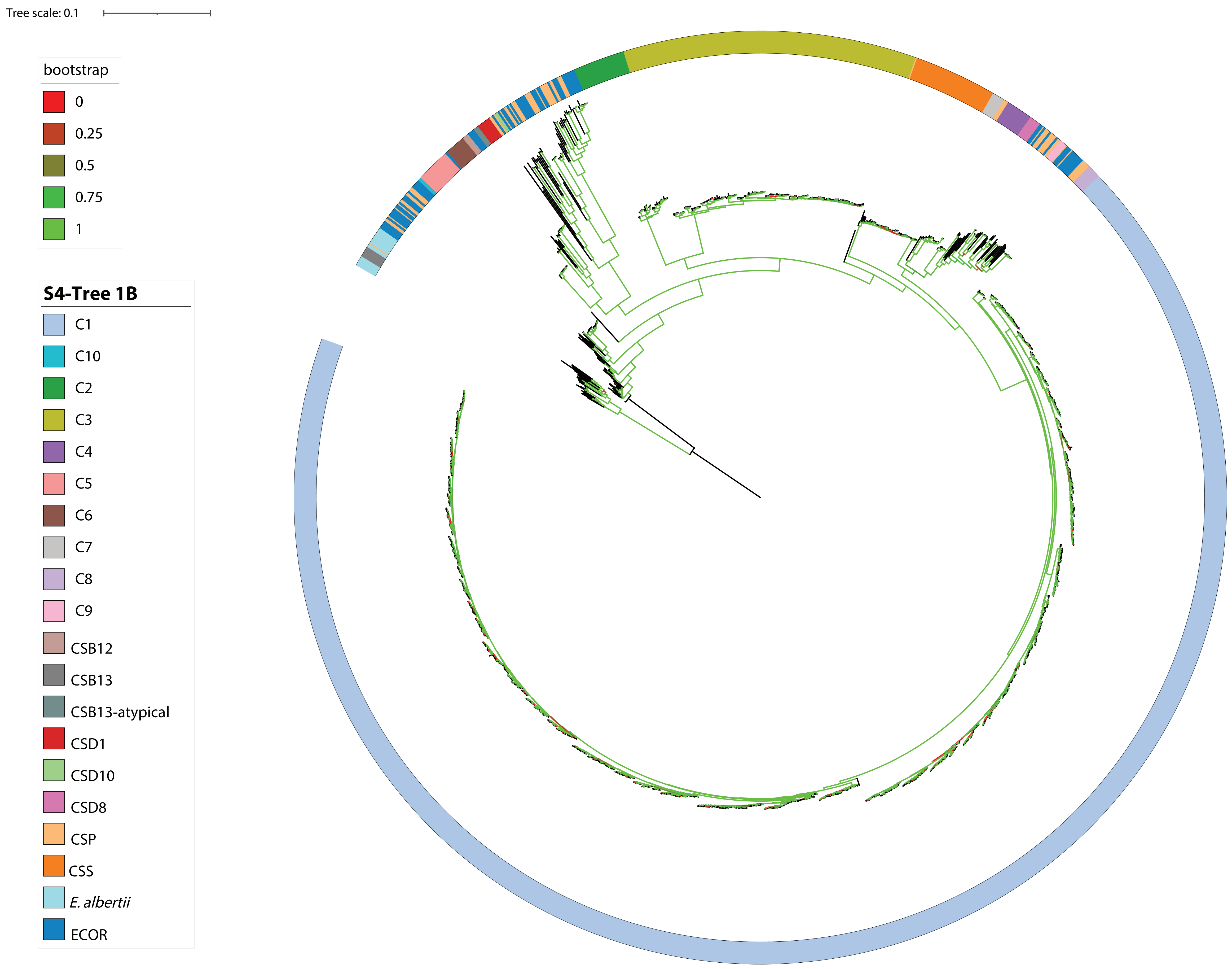

Figure S4-Tree 1B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 1A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

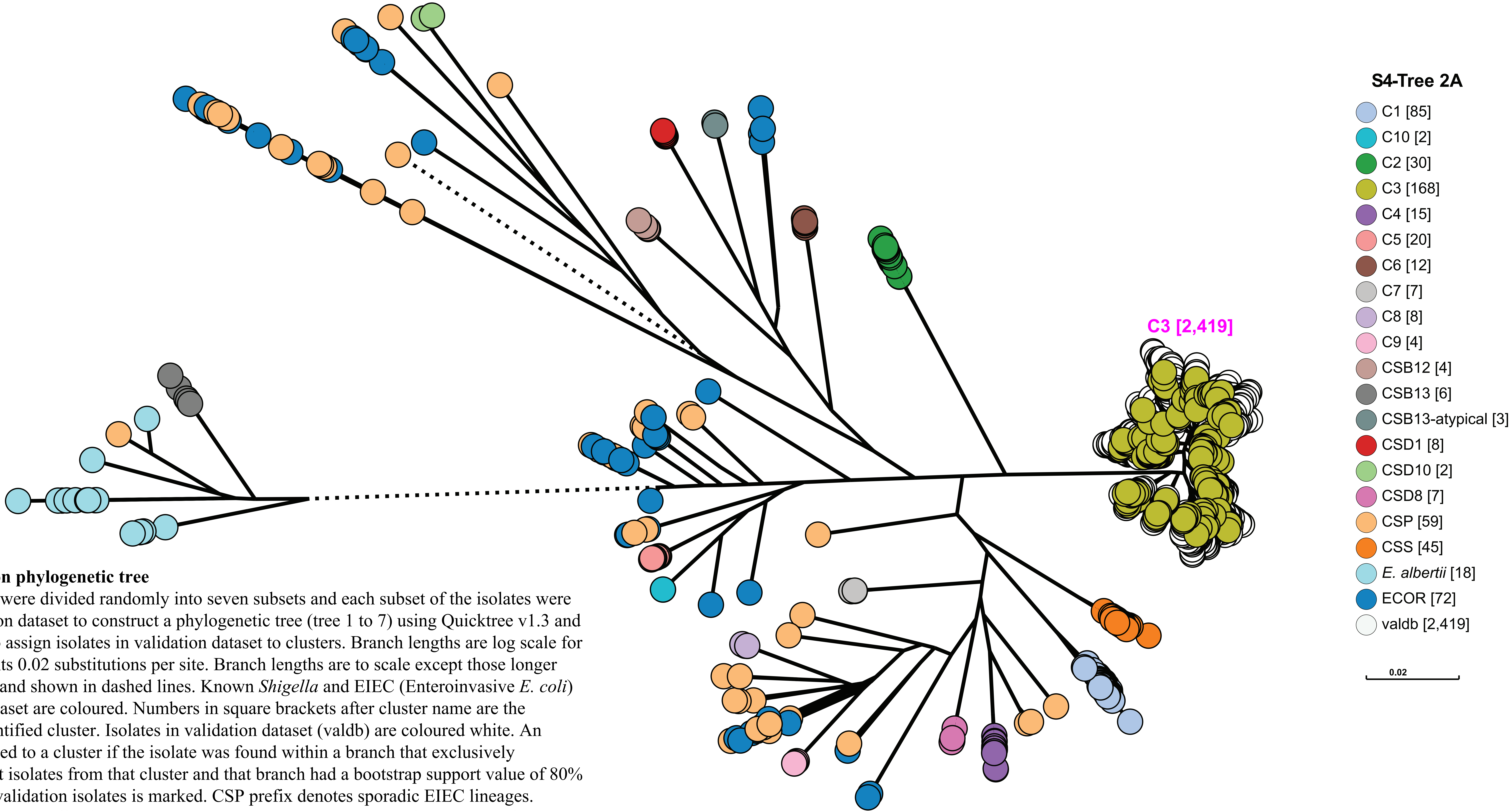

Figure S4-Tree 2A: Validation phylogenetic tree
The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

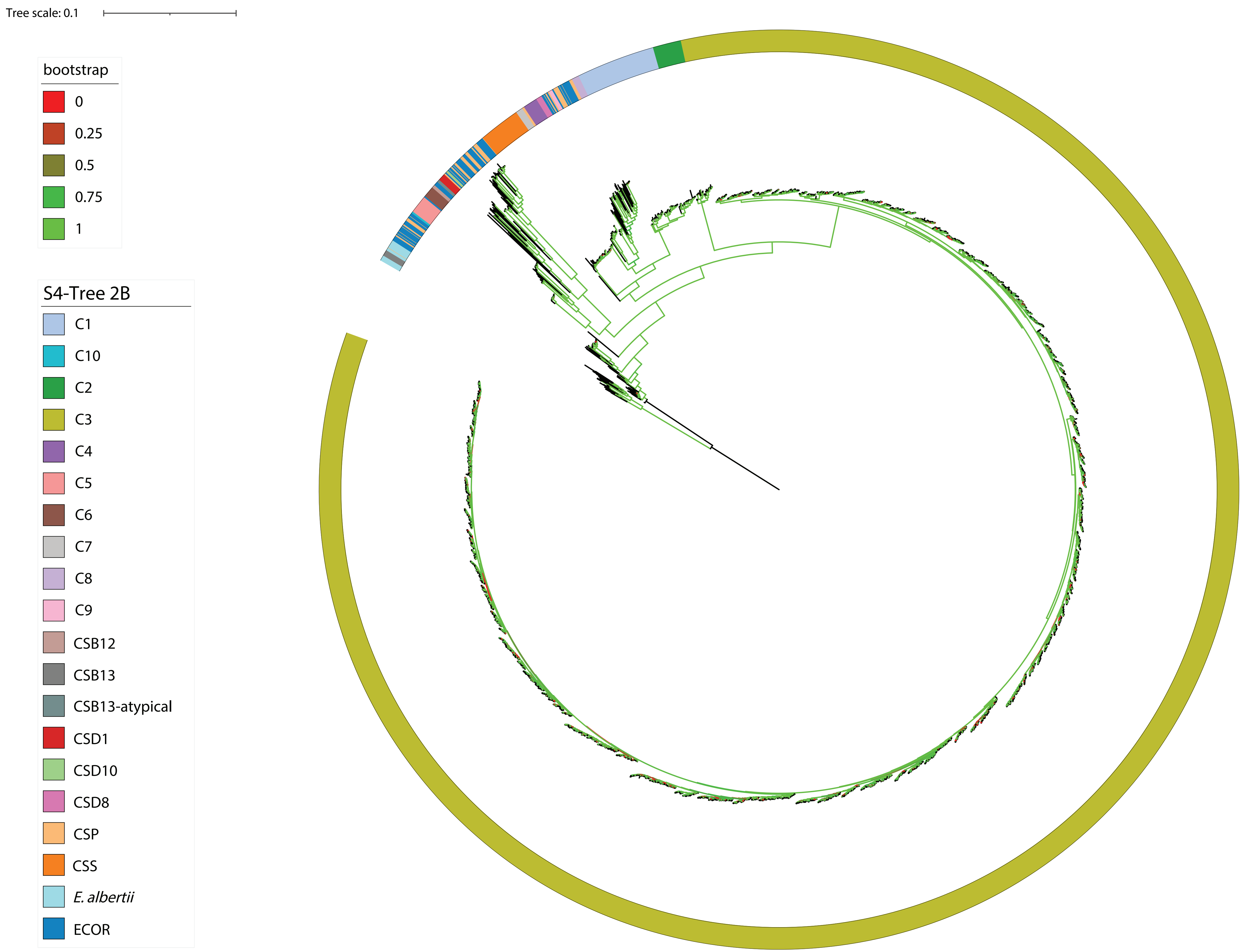

Figure S4-Tree 2B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 2A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

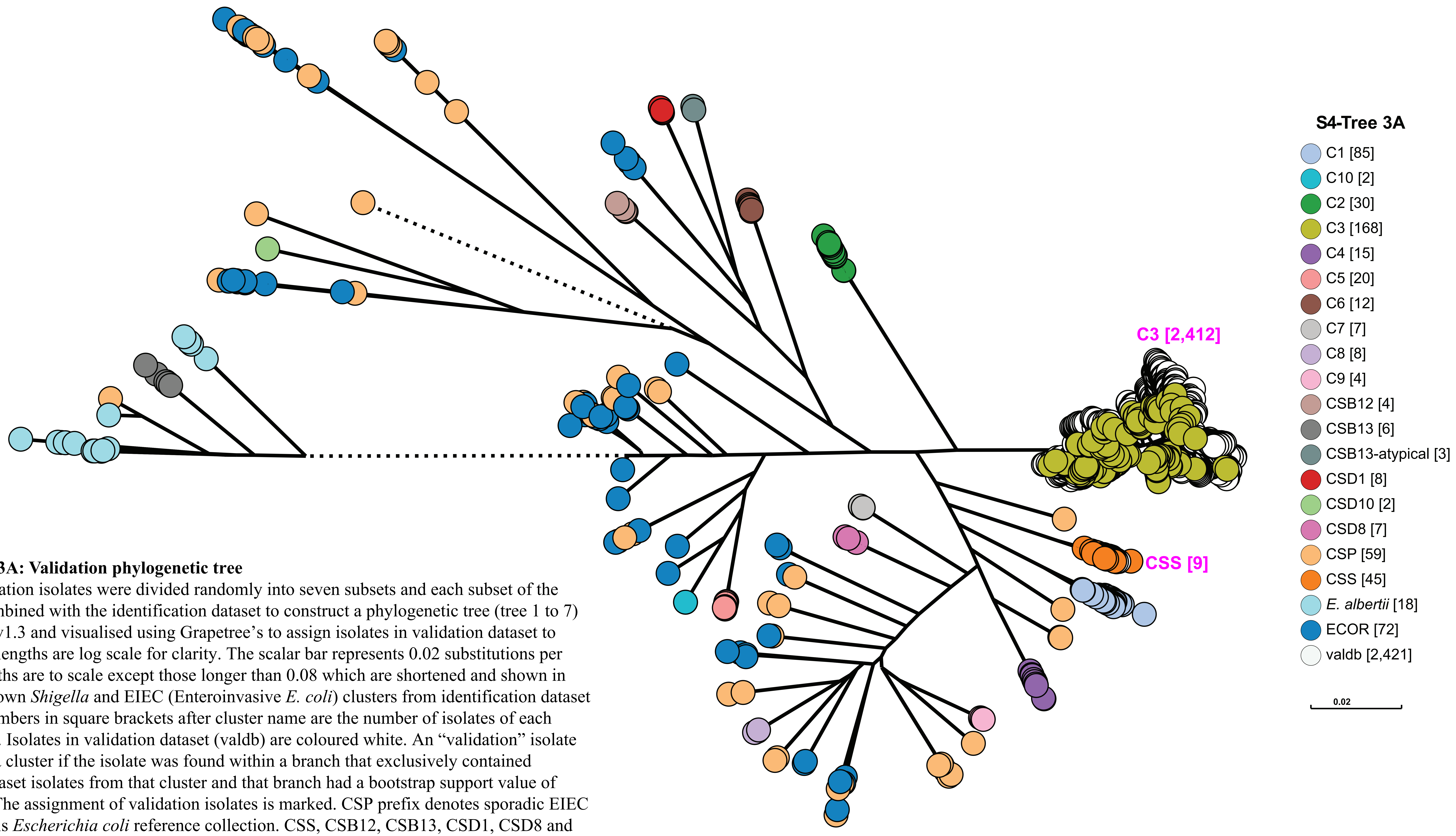

Figure S4-Tree 3A: Validation phylogenetic tree

The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

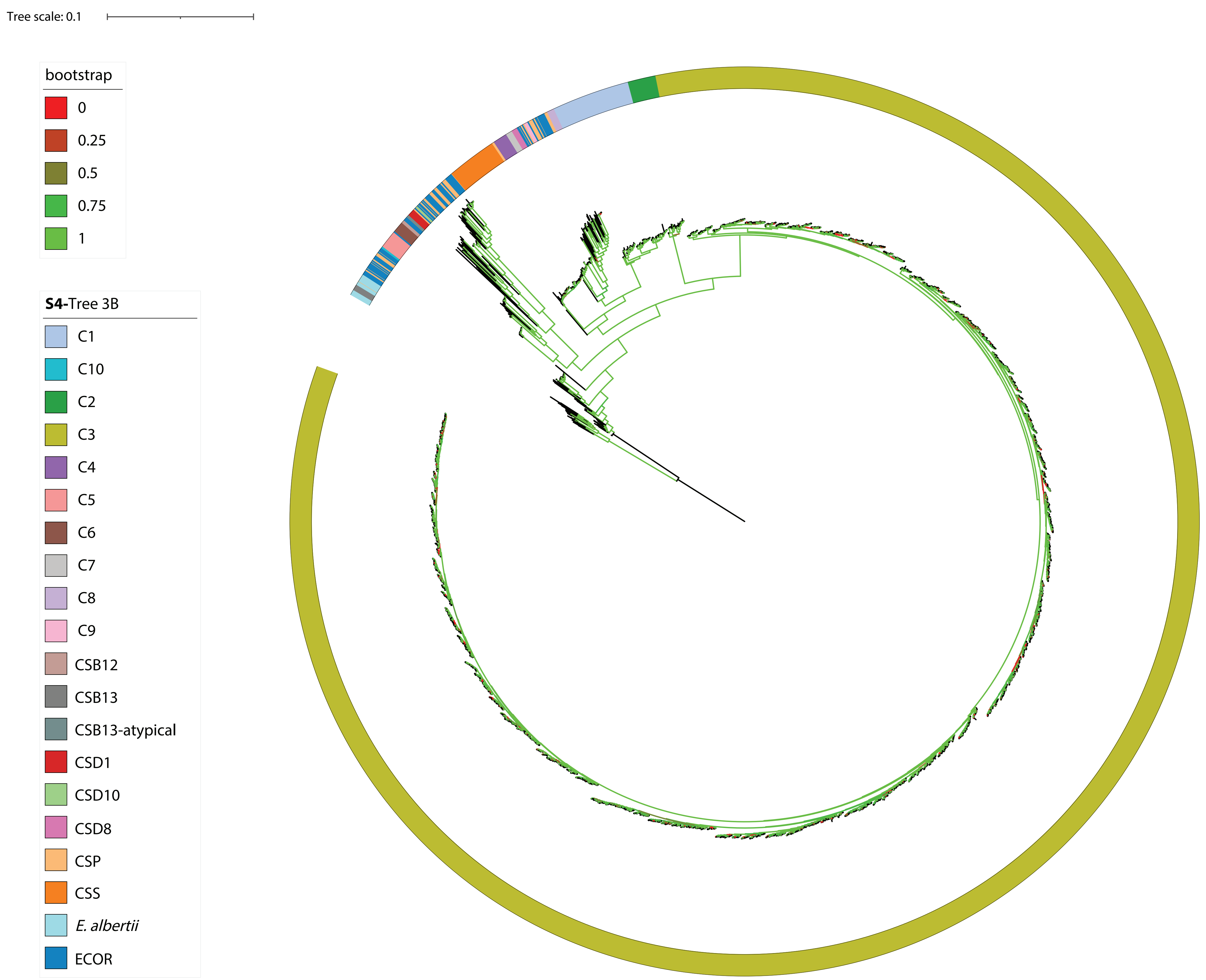

Figure S4-Tree 3B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 3A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

CSP4 [1]
CSB13-atypical [2]
NA
NA
CSD1 [328]
CSP51 [1]
NA
CSB12 [6]
C2 [139]
C3 [2]
CSB13 [1]
CSD10 [1]
NA
CSP19 [1]
CSS [1,010]
CSD8 [3]
C5 [51]
C9 [10]
C8 [21]
CSP41 [1]
C4 [21]
C1 [6]
S4-Tree 4A
C1 [85]
C10 [2]
C2 [30]
C3 [168]
C4 [15]
C5 [20]
C6 [12]
C7 [7]
C8 [8]
C9 [4]
CSB12 [4]
CSB13 [6]
CSB13-atypical [3]
CSD1 [8]
CSD10 [2]
CSD8 [7]
CSP [59]
CSS [45]
E. albertii [18]
ECOR [72]
valdb [1,609]
0.02

Figure S4-Tree 4A: Validation phylogenetic tree

The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

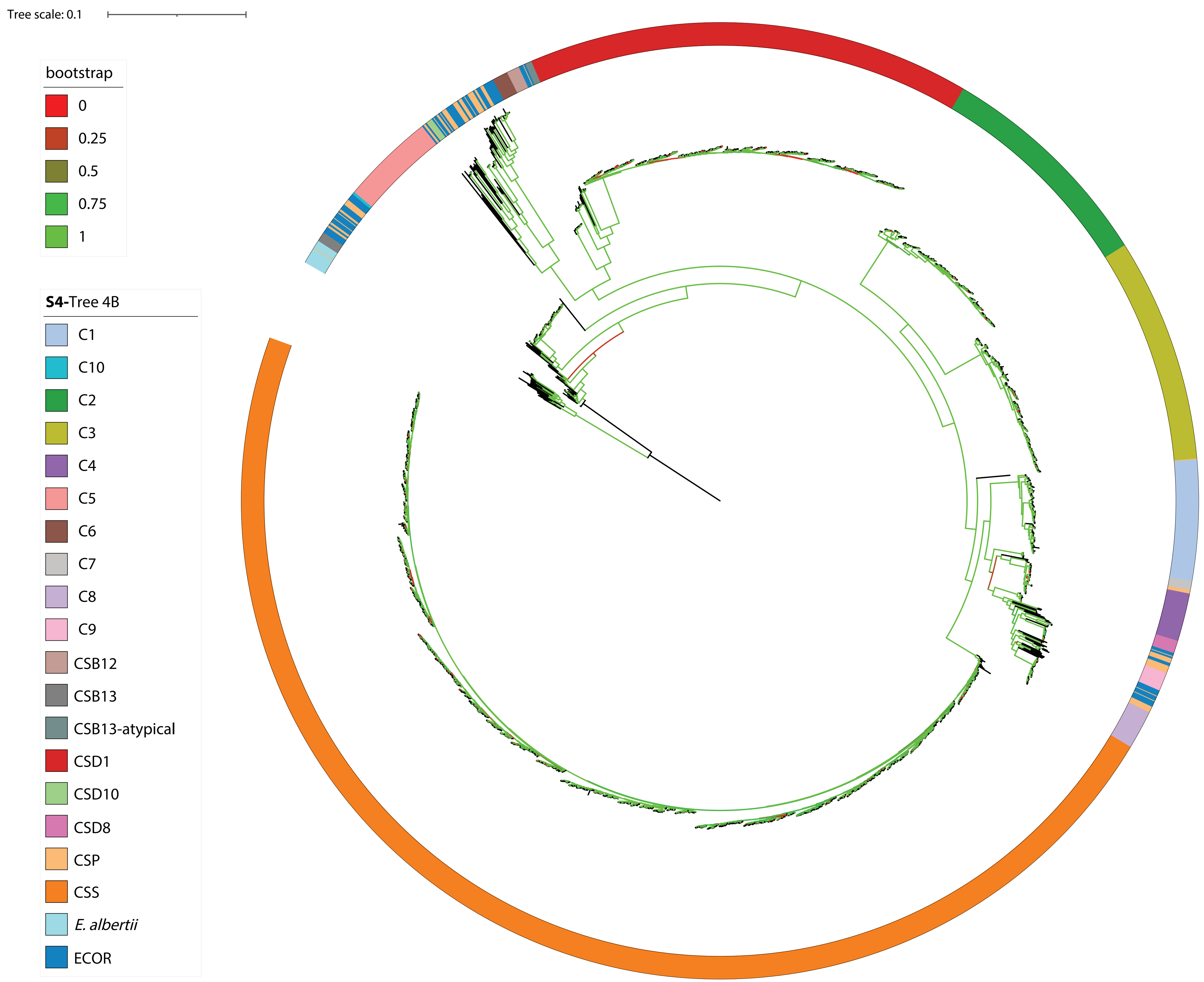

Figure S4-Tree 4B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 4A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

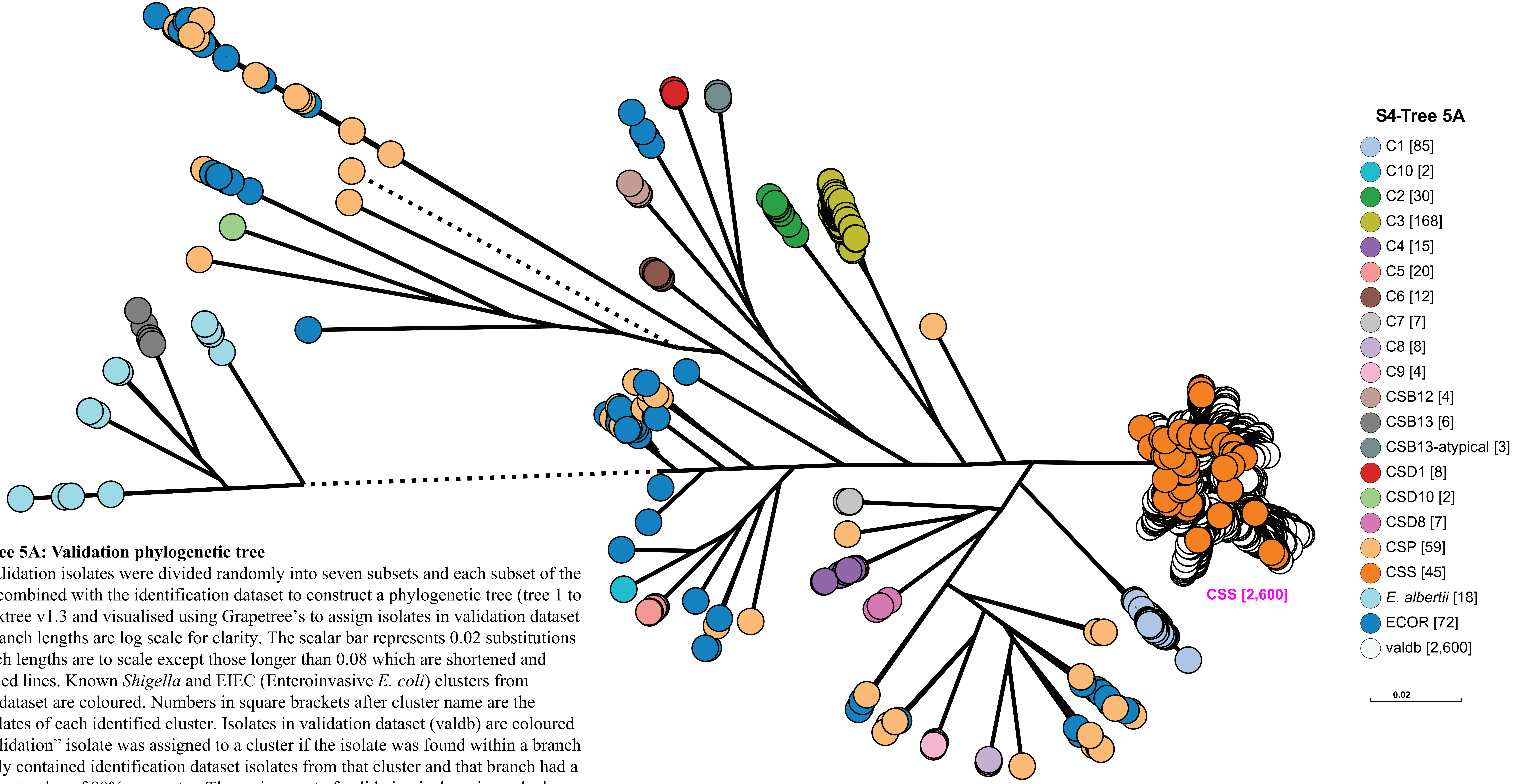

Figure S4-Tree 5A: Validation phylogenetic tree

The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

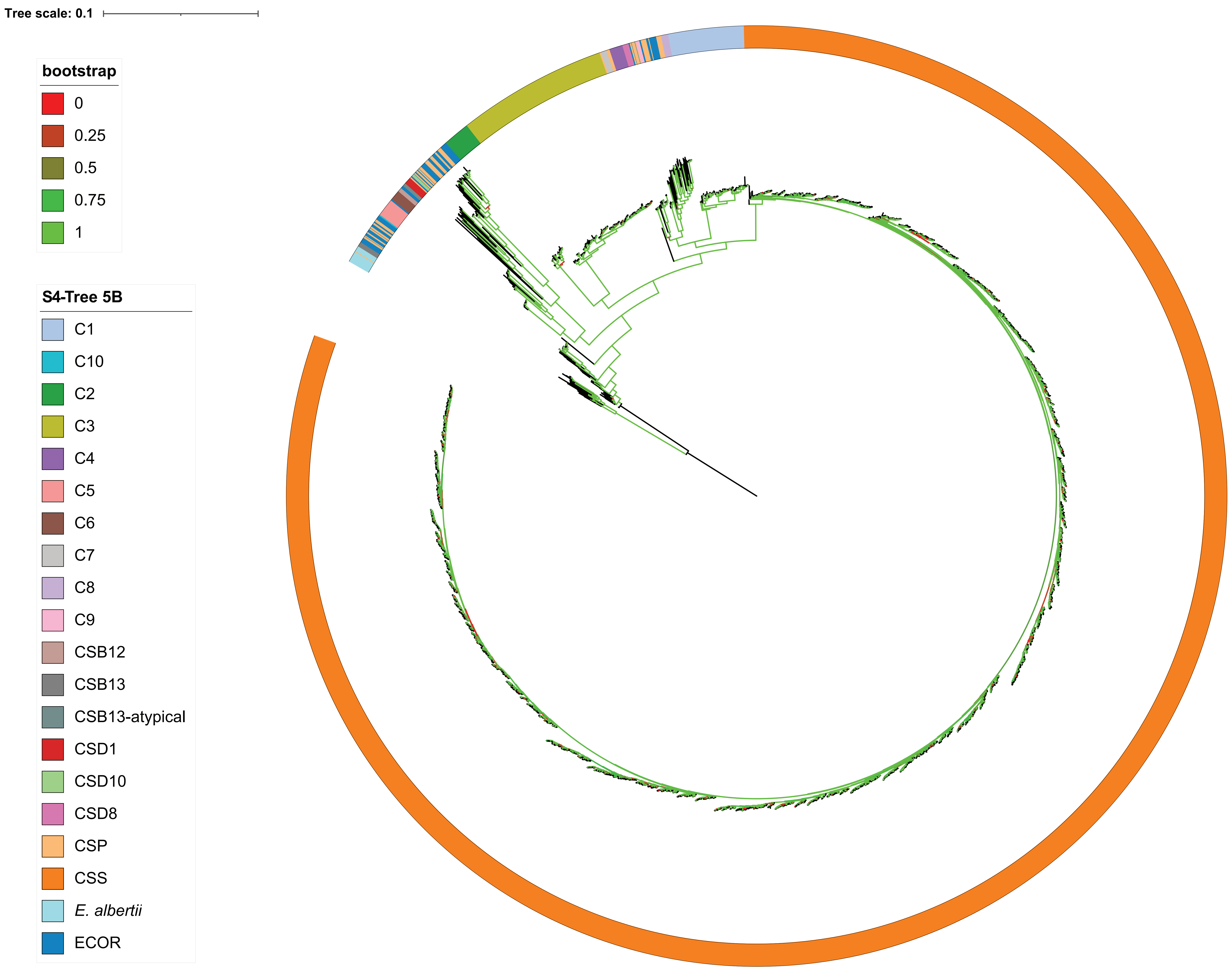

Figure S4-Tree 5B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 5A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

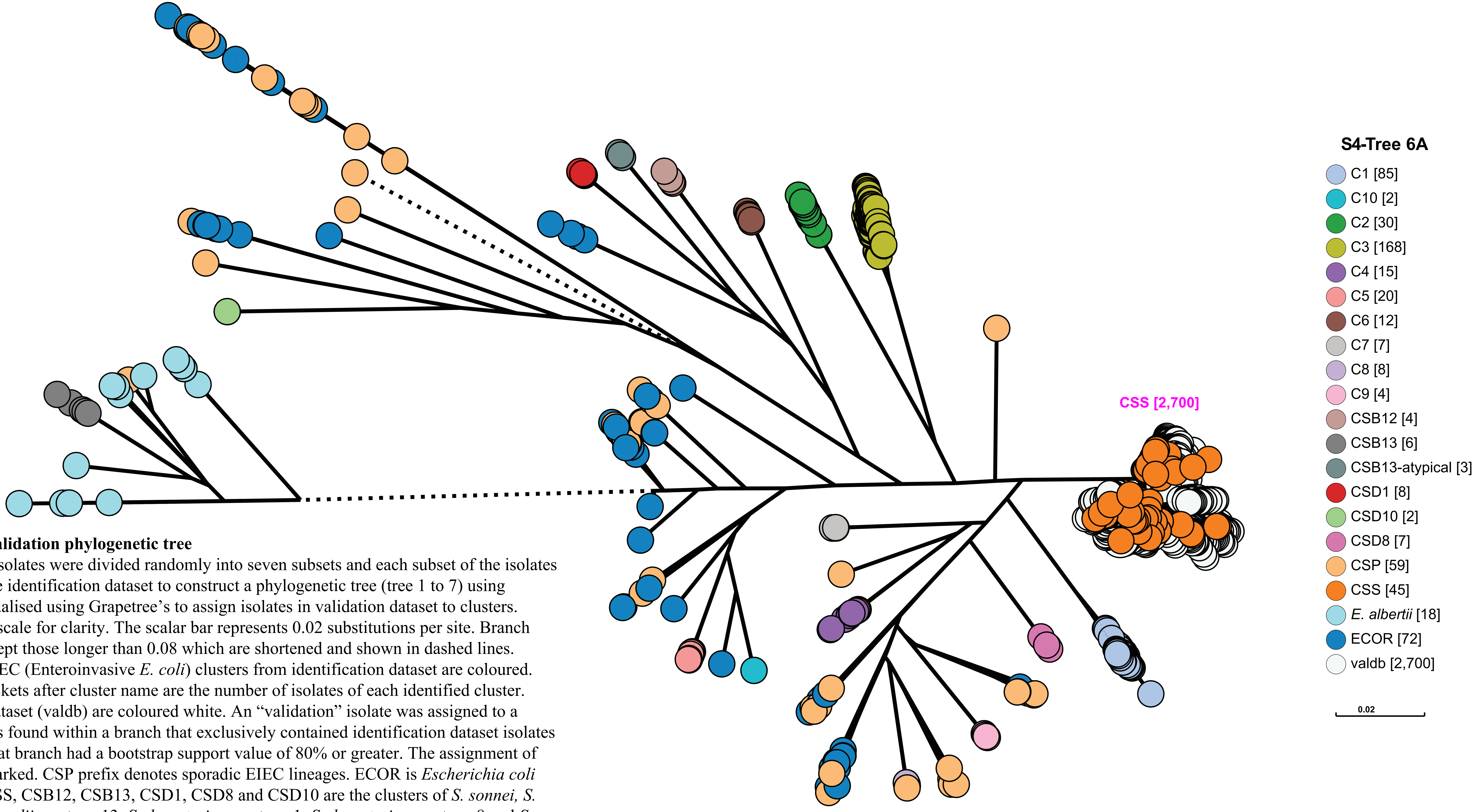

Figure S4-Tree 6A: Validation phylogenetic tree

The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

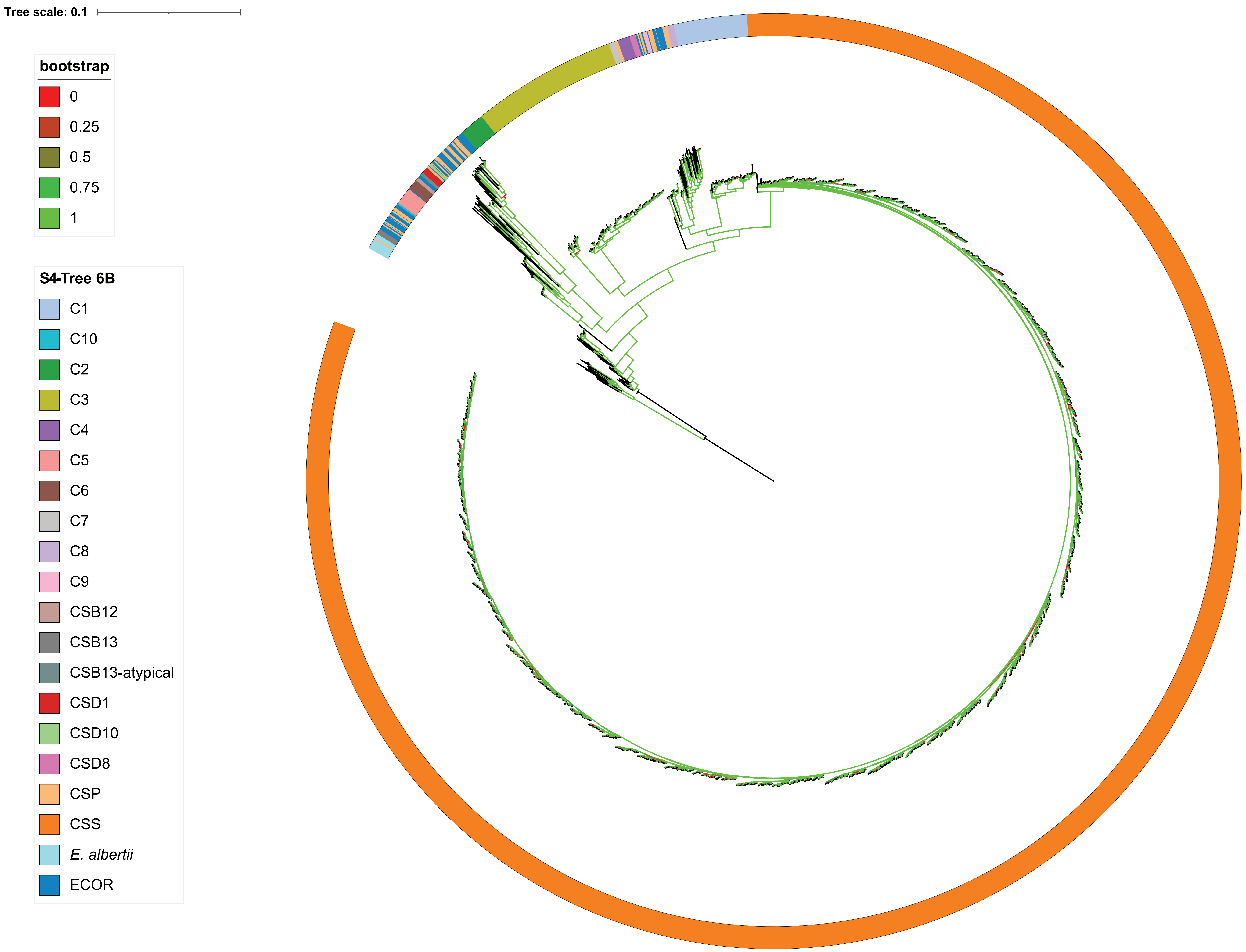

Figure S4-Tree 6B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 6A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

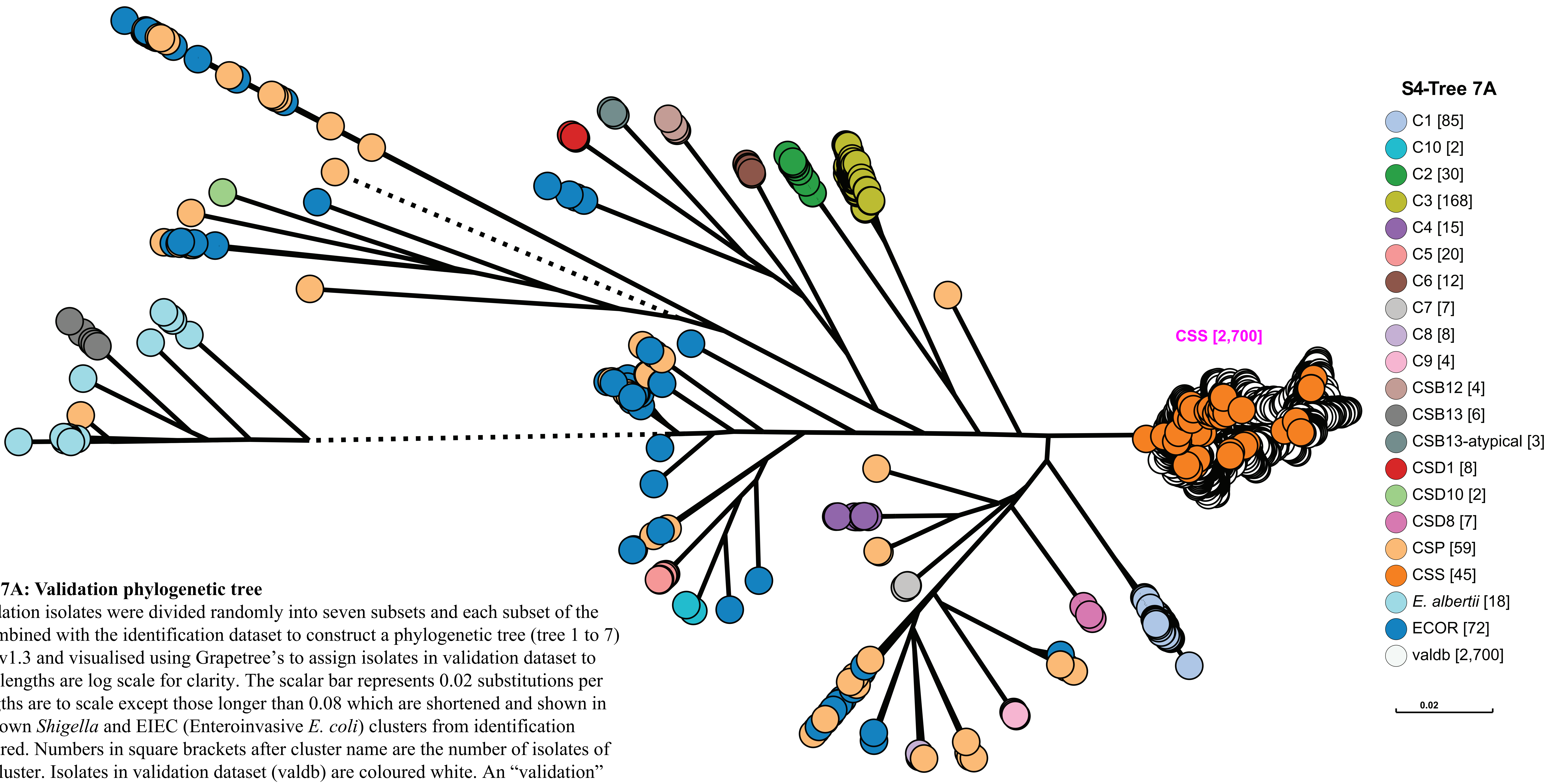

Figure S4-Tree 7A: Validation phylogenetic tree
The 15,501 validation isolates were divided randomly into seven subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 7) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The scalar bar represents 0.02 substitutions per site. Branch lengths are to scale except those longer than 0.08 which are shortened and shown in dashed lines. Known *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. The assignment of validation isolates is marked. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei, S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.

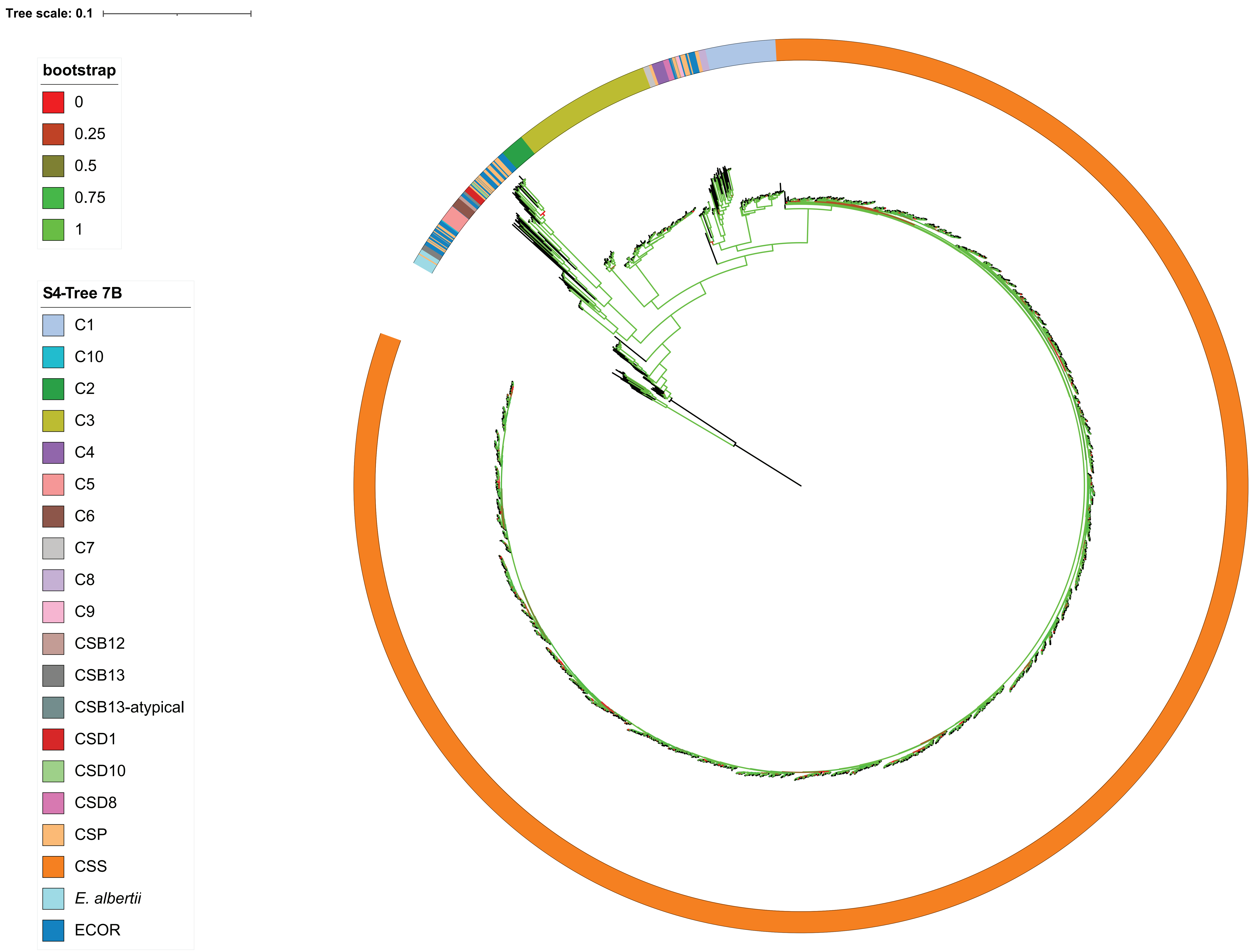

Figure S4-Tree 7B: Validation phylogenetic tree The same phylogenetic tree as Figure S4-Tree 7A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. *Shigella* and EIEC (Enteroinvasive *E. coli*) clusters were coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. CSP prefix denotes sporadic EIEC lineages. ECOR is *Escherichia coli* reference collection. CSS, CSB12, CSB13, CSD1, CSD8 and CSD10 are the clusters of *S. sonnei*, *S. boydii* serotype 12, *S. boydii* serotype 13, *S. dysenteriae* serotype 1, *S. dysenteriae* serotype 8 and *S. dysenteriae* serotype 10 respectively.