1 M M Github

The whole-database tree was obtained using the SNPs only allignment.

The distance was computed from the SNP alignment using the command dna.dist in the ape package.

```
Dist<-dist.dna(all,model="raw",variance=FALSE,pairwise.deletion=FALSE,as.matrix=TRUE)
DistSNP<-Dist*2456
```

The fact that DistSNp are not integer numbers makes me think that somethis odd is going on, but maybe is just an approximation problem somewhere.

The members of each cluster where obtained using the hclust command

```
complete<-hclust(as.dist(Dist, diag=TRUE),method="complete")
groups<-cutree(complete,k=7)
```

Foreach group (except the first that contained the outgroups) the entire genome of its members was collected using the following awk command

```
awk 'BEGIN{c=0}FNR==NR{a[$1]=1;next}{if(match($1,">")){s=substr($1,2,length($1)-1);

if(a[s]==1){c=1} else {c=0}}; if(c==1)print $0}' groupN.list

 ST239_Thai_Imperial_Tong_Harris_Whole_Final.aln.fst  >groupN.fst
```

An outgroup was added to make it easier the tree rooting. The outgroups for groups 2, 3 and 4 was the referece genome TW20. For group 5 was $T099\_N07\_C02$, randomly choosen from group 6. For groups 6 and 7 was $T059\_N02\_C02$, randomly choosen from group 5. We did not use TW20 in these groups as it was too far apart and we wanted something closer.

The tree was built using the following RAxML command (that produces an ML tree but also 100 bootstraps)

```
./raxmlHPC-AVX -f a -m GTRCAT -p 34235 -x 34585 -# 100 -s group6OUT.fst

-n  group6.Boot.tre
```

Then we applied ClonalFrameML

```
ClonalFrameML RAxML_bestTree.group6.Boot.tre group6OUT.fst group6.out -kappa 5
```

Then we dated it it using an R script (on github is called xavierTime.R). To date we need a mutation rate per genome per year. We used 8.4 *add reference*. (Genome length 3043212 bp).

To sum up most of the epi data for the dataset we created two excel files.

finaltable.xlsx contains information on the sequences, their names, the sampling date and the group they belong to. It is color-coded like the colors in the group tree.

patientDemo.xlsx contains information on the patient. It is color coded like the group tree.

More epi information was used, extracted from the given files.

1.1 github

The fasta files ==zipped===

The results of the tree building (raxml+CF) ===zipped===

The script to date the treeis called xavierTime.R

2 Results

The dataset is composed by 1020 sequences. 173 of them have been previously described in the Tong et al paper and 20 of them have already been described in the Harris et al paper and 826 samples are new.

As for the coverage the average coverage of the samples included in the Tong et al study had an average coverage of 276, the samples in the Harris study had a coverage of 86 and the new samples had a coverage of 113, with an overall average coverage of 140 (minimum 48.5).

The dataset is composed by 277 samples from 76 patients. For each sample 1 to 27 colonies where sequenced. In 93 samples at least 7 colonies were sequence, in 88 8 colonies, in 45 nine colonies, in 6 ten colonies, In 5 samples 21 colonies and in 2 samples 29 colonies. For 182 samples we had just one colony sequenced. All the further colonies where sequenced in the last data release.

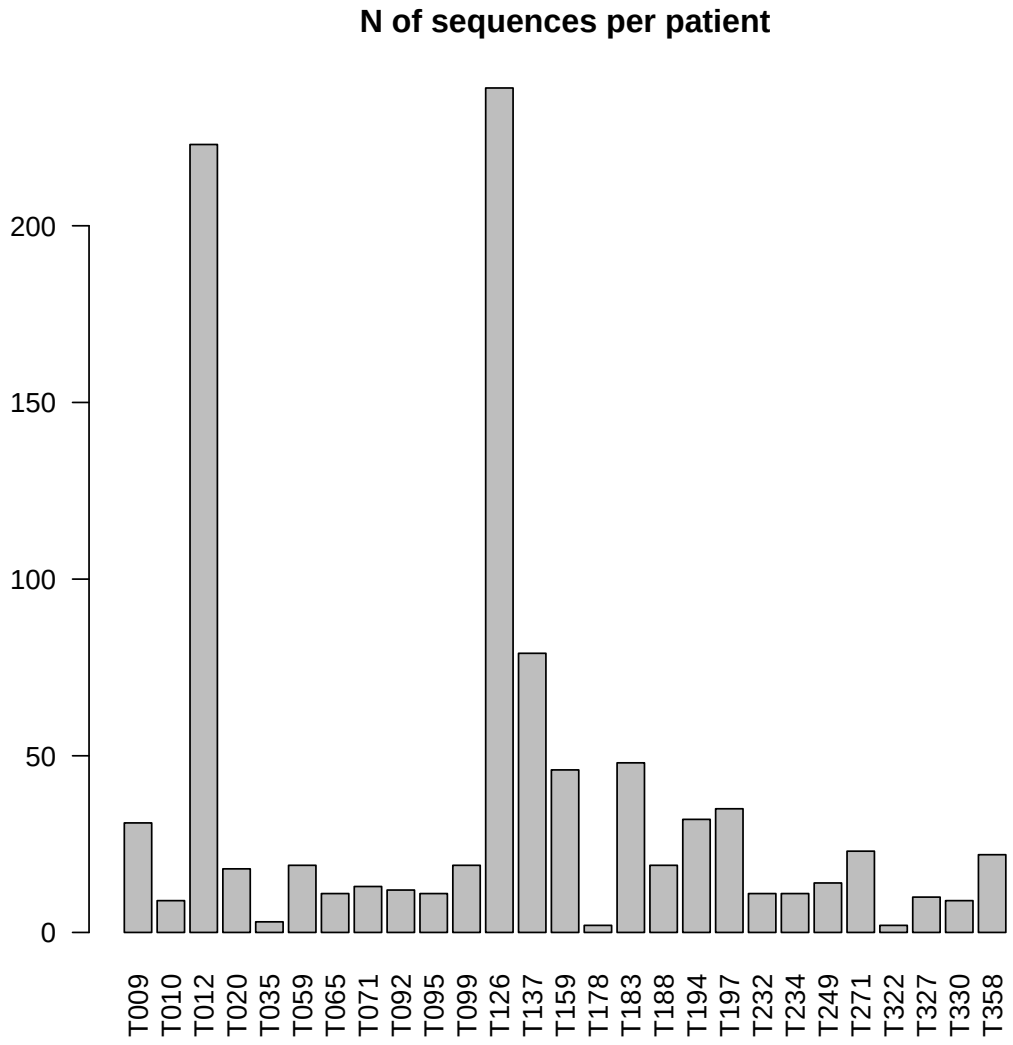

Figure 1: Number of sequences per patients, for the patients for which we had more than one. Probably I should color the patient ID so that we know whether is adult or child.

As for the bodyparts, the vast majority of the sequences (863) come from the nose. That's because all the samples with multiple sequenced colonies are nasal samples. We also have 27 axilla samples (A),36 Tracheal samples (C), 62 throat samples(T), 11 wound samples (W),2

urine samples (U) and 1 H sample (I do not know what it is, but whatever it is, it's taken from a nurse). Interestingly, patients whose samples are H and U only have that sample, so no comparison is possible. This happens also for some other bodyparts. I.e. wound for person T156.

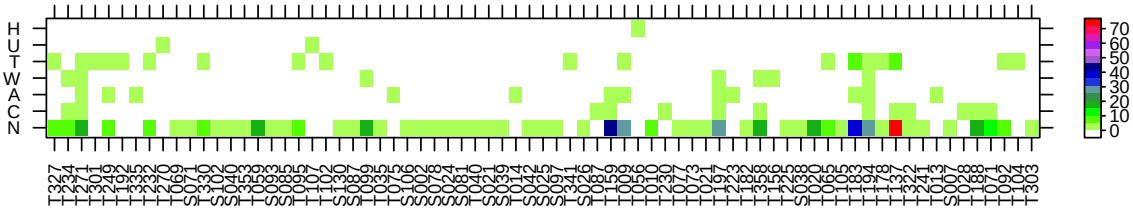

Figure 2: The body parts for all the people for which we had more than one sample. They can be all from the same bodypart.

2.1 The groups

Given the distances among samples in the tree (realized justfrom the SNPs data) we divided the dataset into 7 groups and explored each group more in detail to seek answers to our questions. The composition of the groups where highly nonhomogeneous as for number of samples, number of patients and bodyparts.

Within and between groups distances are showed in the plots. I am not including the heatmap of the distances among all the sequences just because it is very heavy as an image. If you want to see it you find it in the report folder in the zipped file on github

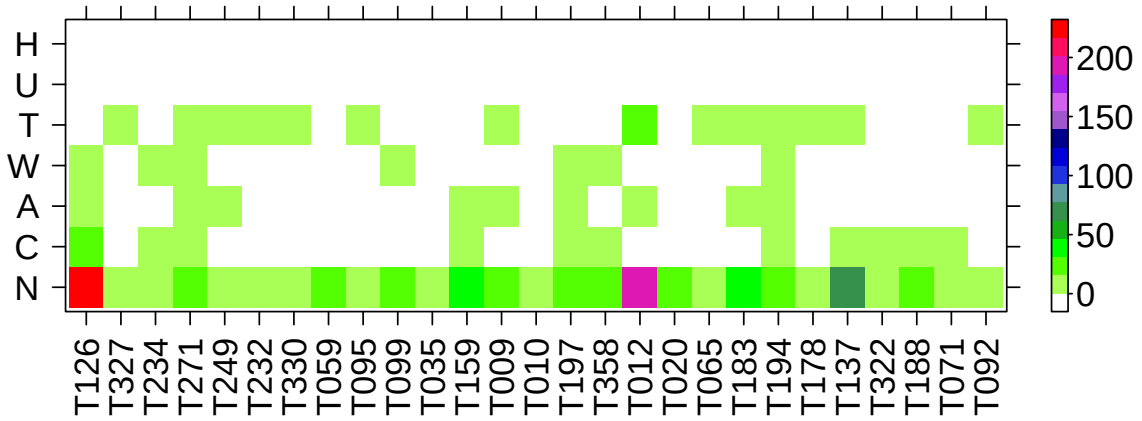

Figure 3: Only patients for which we had samples from more than one bodypart.

2.2 Group1

Just the outgroup TW20.

2.3 Group2

65 sequences from 25 individuals. Mainly infants a part from 3 adults and 2 nurses. All the possible bodyparts except urine.

There is essentially one main clade, all the outliers are too far to infer any transmission so we focused on the main clade.

In the main clade there are two mainly monoclonal clades. One is T009 and the other T010. They're both infants. T009 was negative on firts admission but positive on readmission. T010 was negative on admission.

The H sample from nurse T056 clusters well in the middle of T009. Is it a transmission?

T159 has samples in this group and in group 3.

All the colonies belonging to T010 cluster very closely together and separated from the rest of the tree.

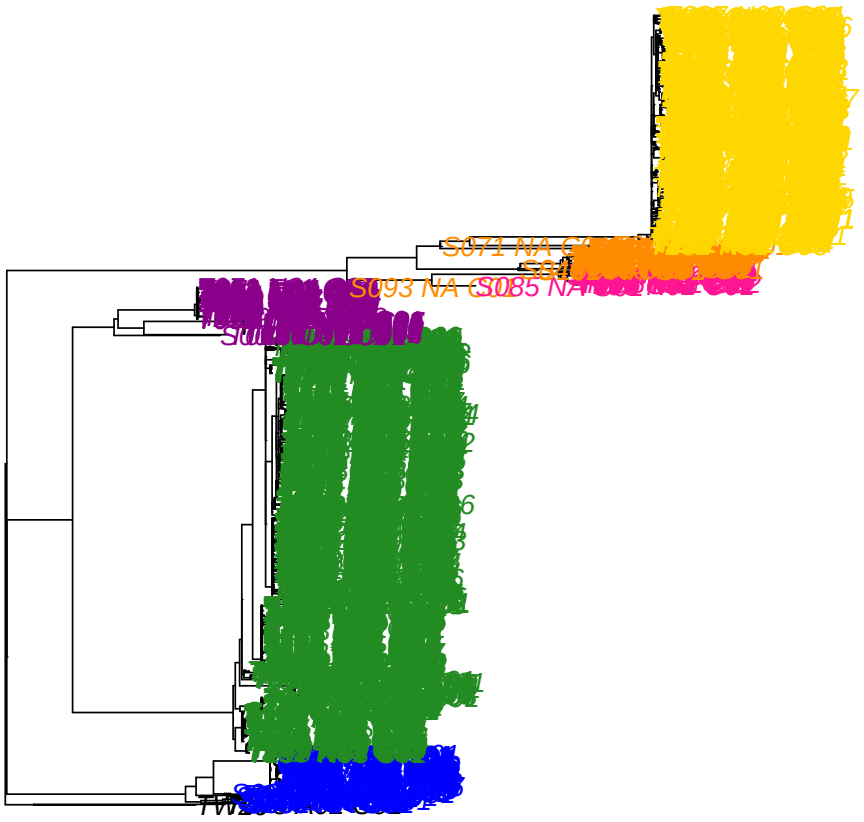

Figure 4: The groups on the total tree. The color code will be kept throughout the report.

There might have been a transmission between T073 (infant) and T077 (adult), but we have just one sample so it is hard to determine. T073 was admitted a couple of days before T077 and was positive on admission while T073 was negative. T021 is genetically not far from these two, is an infant and negative on admission. Was admitted before T073, and discharged after.

2.4 Group3

539 sequences from 23 people. Only five bodyparts: nose, axilla, throat, trachea and wound.

No tree yet

2.5 Group4

Group 4 has 63 sequences from 7 patients. 4 patients are adults and 2 are children. Patient T099 is also in group 6, so he is clearly infected with at least two different strains. This group has some tragic outcomes as one infant (T104) died and the other infant (T188) and one adult (T028) were *"disc mori"* (discharged to die at home?).

Both the infants were positive on admission while none of the adults was. While most of

Group	N of sequences	N of people	body parts	Average distance	color
2	65	25	A,C,H,N,T,W		blue
3	539	23	A,C,N,T,W		green
4	63	7	C,N,T		purple
5	34	9	N,T		orange
6	20	8	N,T,U,W		pink
7	297	11	A,C,N,T,U,W		yellow

Table 1: Sum up of group features

Group	min	mean	max
2	2006.5 (2008.158)	2008.004 (2008.236)	2008.391
3	2006.5 (2008.158)	2008.311 (2008.314)	2008.415
4	2006.5 (2008.183)	2008.238 (2008.266)	2008.41
5	2006.5 (2008.183)	2008.224 (2008.353)	2008.393
6	2007.5 (2008.191)	2008.114 (2008.223)	2008.262
7	2008.24	2008.24	2008.415

Table 2: Sampling times. In parenthesis the same quantities without the S samples

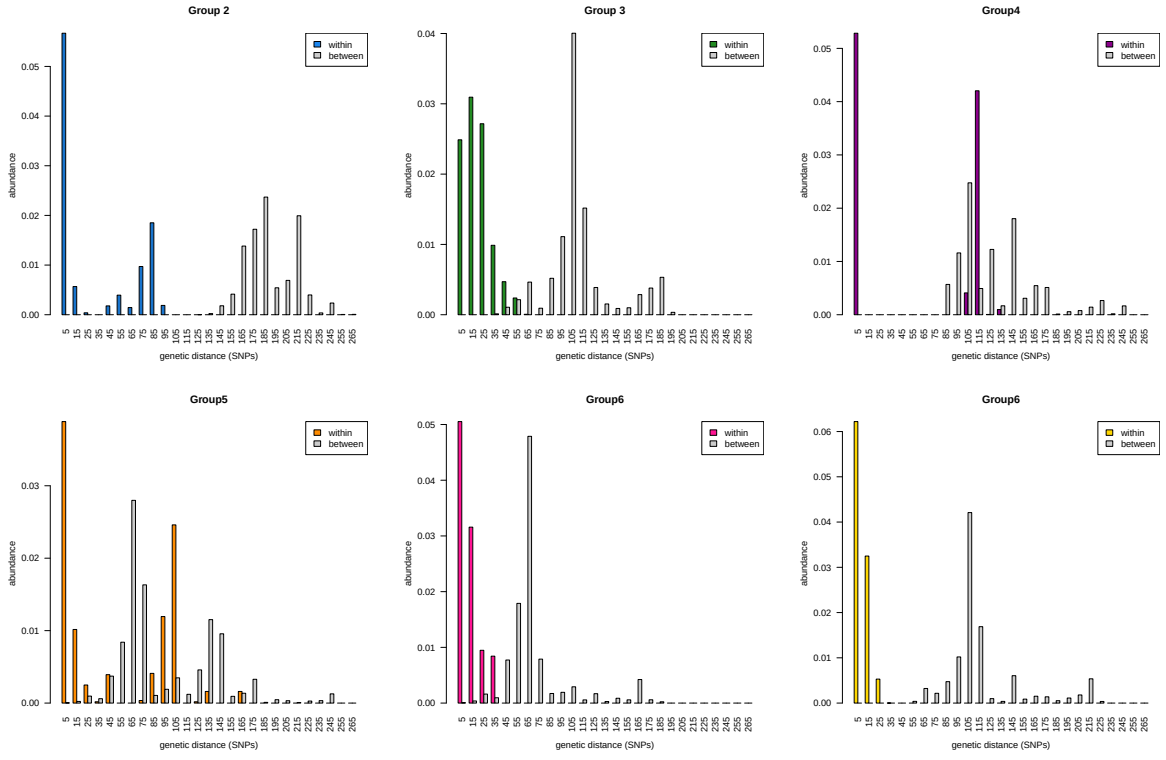

Figure 5: The distances for each group. In color the distances between any two sequences within the group, in grey the distances between a sequence than belonged to the group and a sequence that did not belong to the group.

them were positive already on the second day at least in one bodypart, T099 stayed negative for a while, becoming positive only after one week.

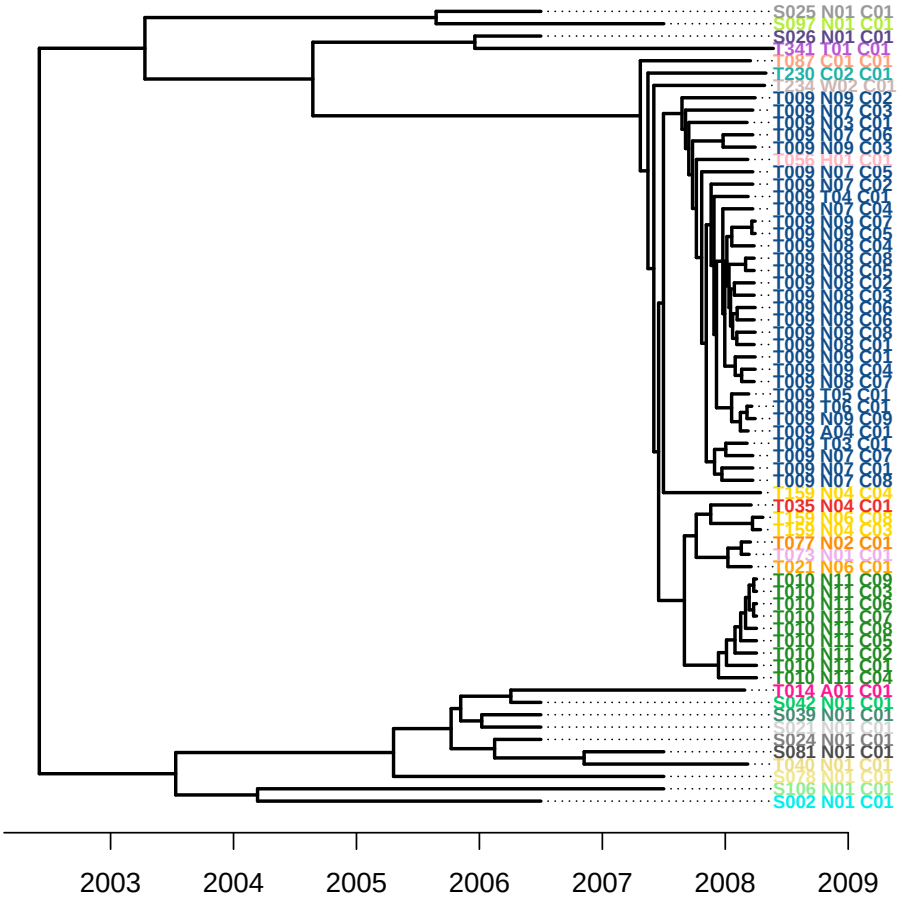

Figure 6: Group2

The group is clearly divided in two main clades plus a couple of outliers. S007 is clearly an outlier.

T188 clearly forms a clade on its own and T028 is clearly an outlier to him. This is interesting because, although they weren't in the same ward nor in the hospital at the same time they were both *"disc mori"*. T188 was admitted twice to the ICU. One on April 10 to 15 (in the bed logs looks more 12 to 14) and the second on May 29. But from her sequences we see that she was infected just once and that the bacteria sequenced in the second admission are the same as in her first. In this case sequencing many colonies (on her last day of first admission and on the day of her second admission) does not add much information.

The other clade is mainly located in the adult ICU (just one sample from the infant ward). While the samples from T071 and T092 are completely intermingled, most of the samples from T099 cluster in one clade, and only 3 of the samples are scattered around. Is it a problem with the bottleneck, like the amount/type of bacteria transmitted to T099 is different from the amount/type transmitted to T073 and T092? Or is the clade expanding in T099 more adapted to him?

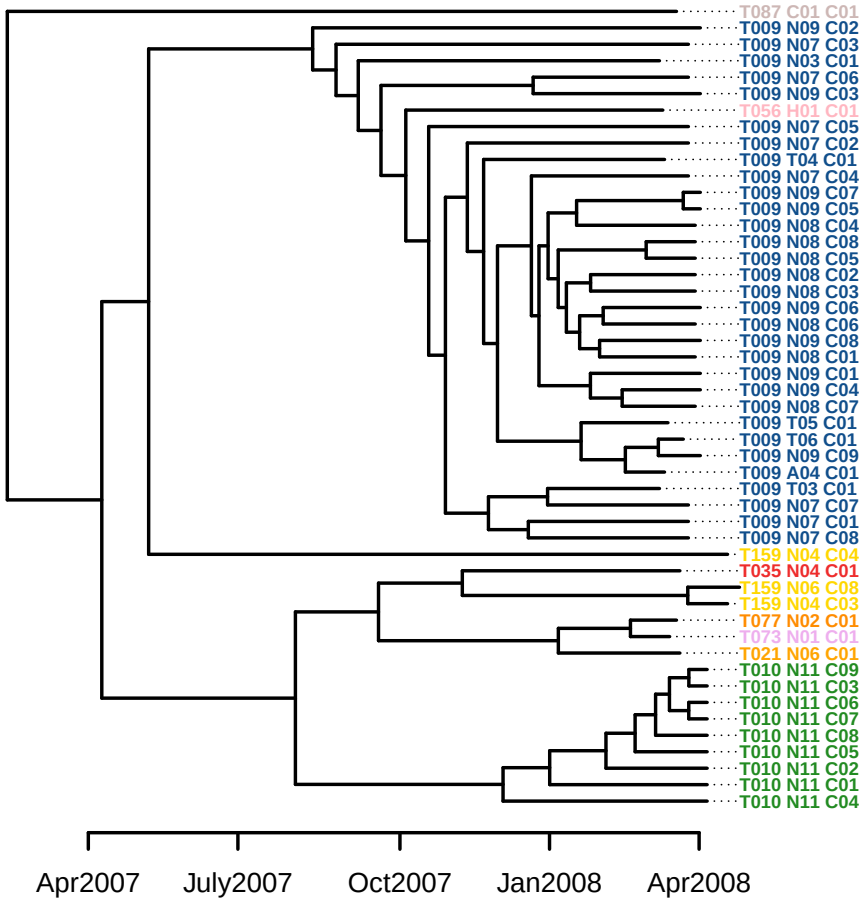

Figure 7: Group2, just the interesting clade

2.6 Group5

Groups 6 has 34 sequences from 9 people. Mostly from the nose (just one from the throat). The nurse T059 was sequenced on March 7th, April 25th and May 23rd. All her sequences cluster here. Nurse T353 was sequenced only on May 23rd. Yet the most recent common ancestor of the two dates back to more than one year before, while the diversity within T059 is much more recent. The diversity within T059 looks more like multiple infections than like a single infection. Yet it is not clear why multiple infections would be all clustering together. It points to some ecological dynamics within this nurse.

Patient T069 is an infant sequenced in the first half of March and discharged home on March 18th. His sequence clusters with the sequences belonging to T327 and T330, but the times are so long that we cannot safely infer a transmission.

T327 and T330 are two adults who stayed in the ICU ward in mid May. In particular T327 from May 16th to May 20th coming directly from home and T330 from May18th to May 20th coming from another hospital ward (S4). Considering that T327 was negative on admission (on his first swab) on May 17th and positive on May 20th and T330 was positive on admission on May 19th we would guess that T330 infected T327. On the other hand T327 stayed in Bed 5

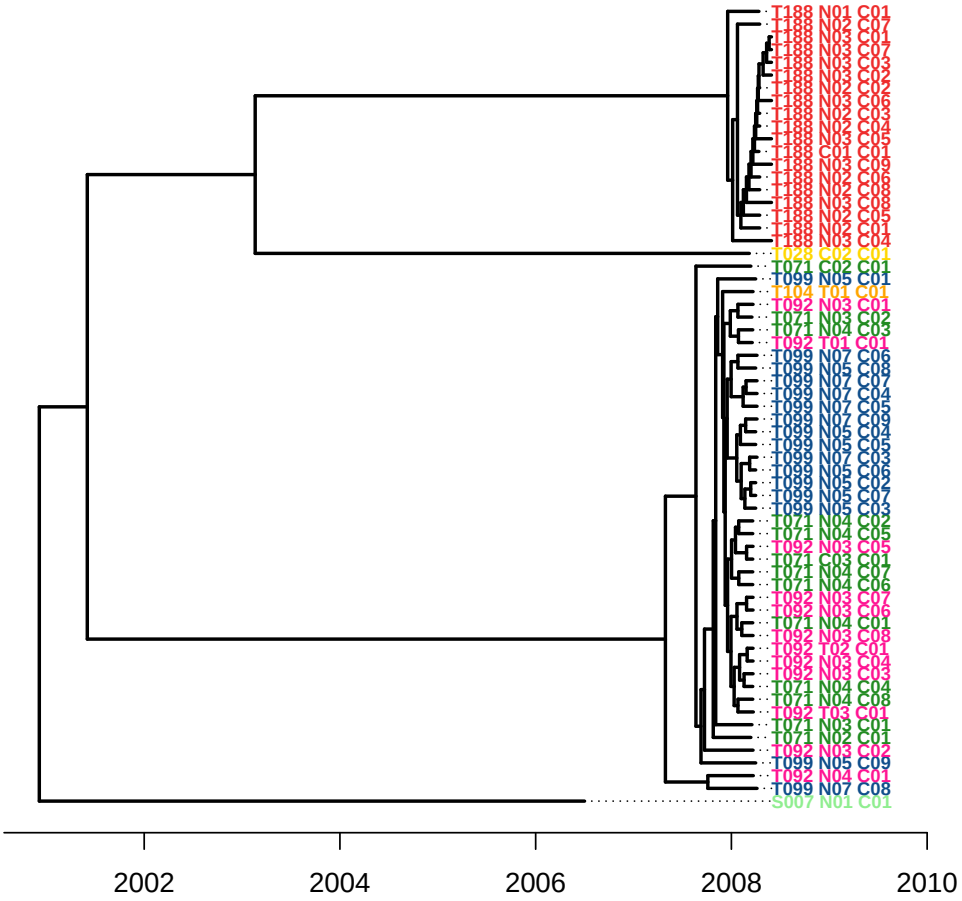

Figure 8: Group4

on May 17th, there is no data for May 18th and T330 is in the same bed on May 19 and 20. Also the only T327 sequence seems to be an ancestor to most of the T330 sequences. Still the times are very short. Both circumstances would favour T327 infecting T330.

It is relevant to point out that both patients were admitted in the afternoon/early evening but the first swab was taken the following morning. Nonetheless T327 is negative on the first swab and positive 3 days after. Does this tell us anything on the bottleneck? I mean, how probable is it that T327 got infected in the admission process and it took 3 days for the bacteria to grow and outnumber his previously negative bacteria?

It is also interesting to note that most of the T330 sequences belong to the same sample taken on May 19th and that if we had sequences just one of the 7 colonies, had we got 5, 6 or 1 the phylogeny would have been totally coherent with the epidemiology (T330 infecting T327), while the other 4 would have been incoherent. Anyway the times are so short that nothing could have been ruled out. On the other hand it would have been interesting to see what we would have seen having other 6 colonies from T327....

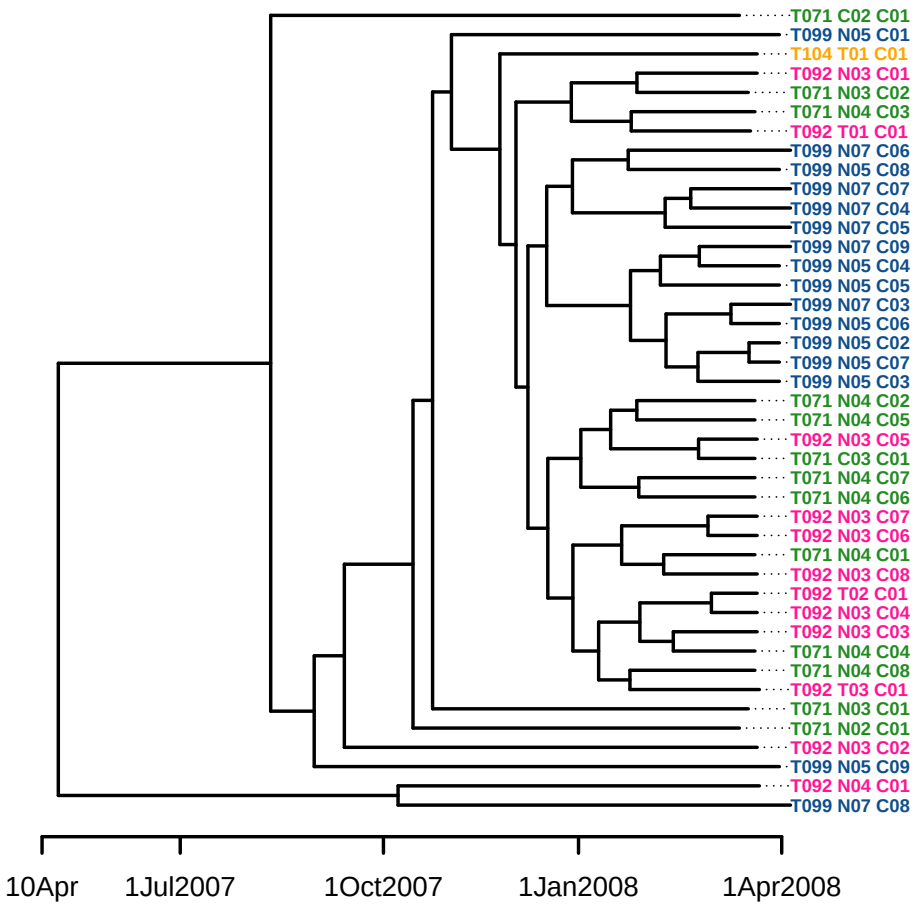

Figure 9: Group4, just the more interesting part

2.7 Group6

Group 6 is the smallest group with 20 sequences from only 8 patients. 4 of the 5 T patients are in the adult unit and one is in the infant unit (T035).

The infant T035 was discharged from the Pediatric ICU the day in which the first of the adults was admitted. His sequence is far from the others

T095 and T107 were directly admitted to the ICU while T099 and T102 were admitted coming from another ward (but not the same ward). Finally all but T095 were negative on admission.

From a temporal point of view T035 should the one that has started the epidemics, the MRCA is more than one year apart. On the other hand T095 comes directly from home and is positive on admission, so it seems that she cannot have been infected in the hospital. It is plausible that she infected both T107 and T102. All of her samples cluster together in this branch, including the only one from the throat too, which points to the fact that she was colonized by the same population both in the nose and in the throat. Yet it is interesting to note that the throat sample is more diverse than the nose ones. T099 was eventually positive just in his second ammission to ICU. Interestingly the 4 adults had procedures related to the

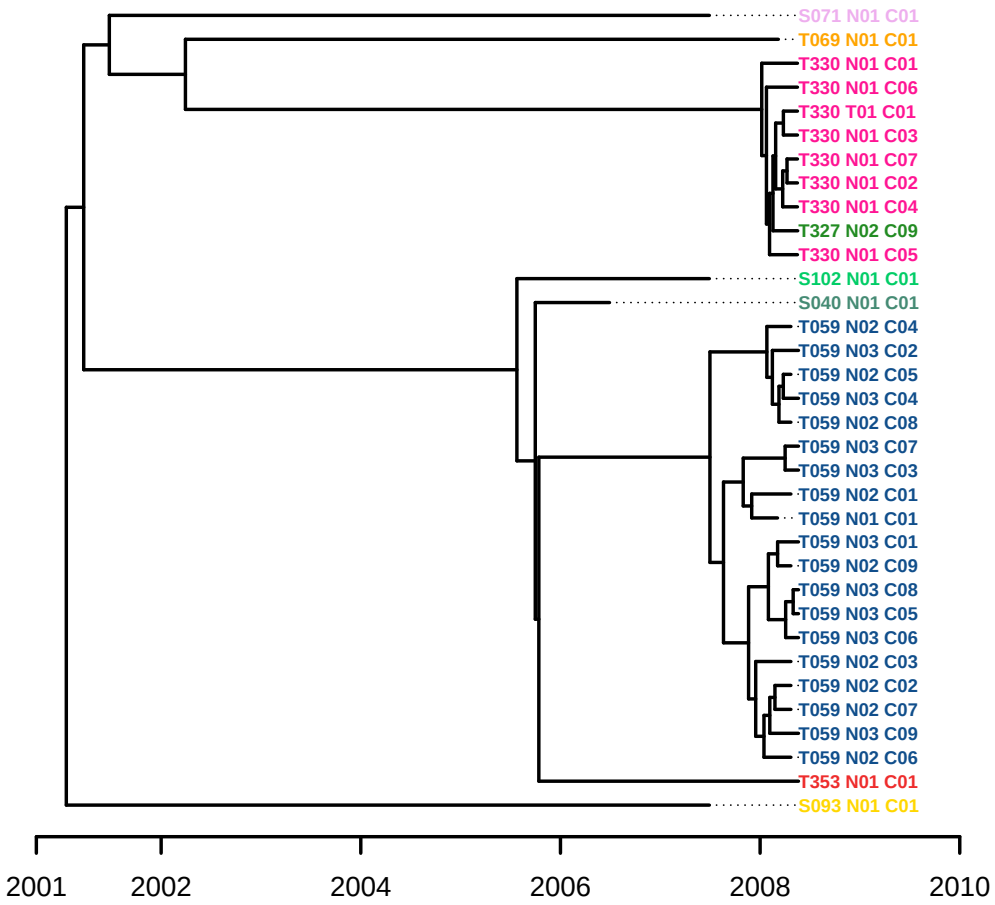

Figure 10: Group5

abdominal region.

2.8 Group7

297 sequences from 11 people. All the possible bodyparts (except H that we do not know what it is and is present only once in one nurse in group 2).

No tree yet

2.9 Heterozygosity

3 Discussion

3.1 the questions

1.Would we see the same transmission patters if we had sequenced only one colony per sample?

2.Do we see different strains in different body parts?

3.If we assume that the coalescence times are real, can we say something on the bottleneck?

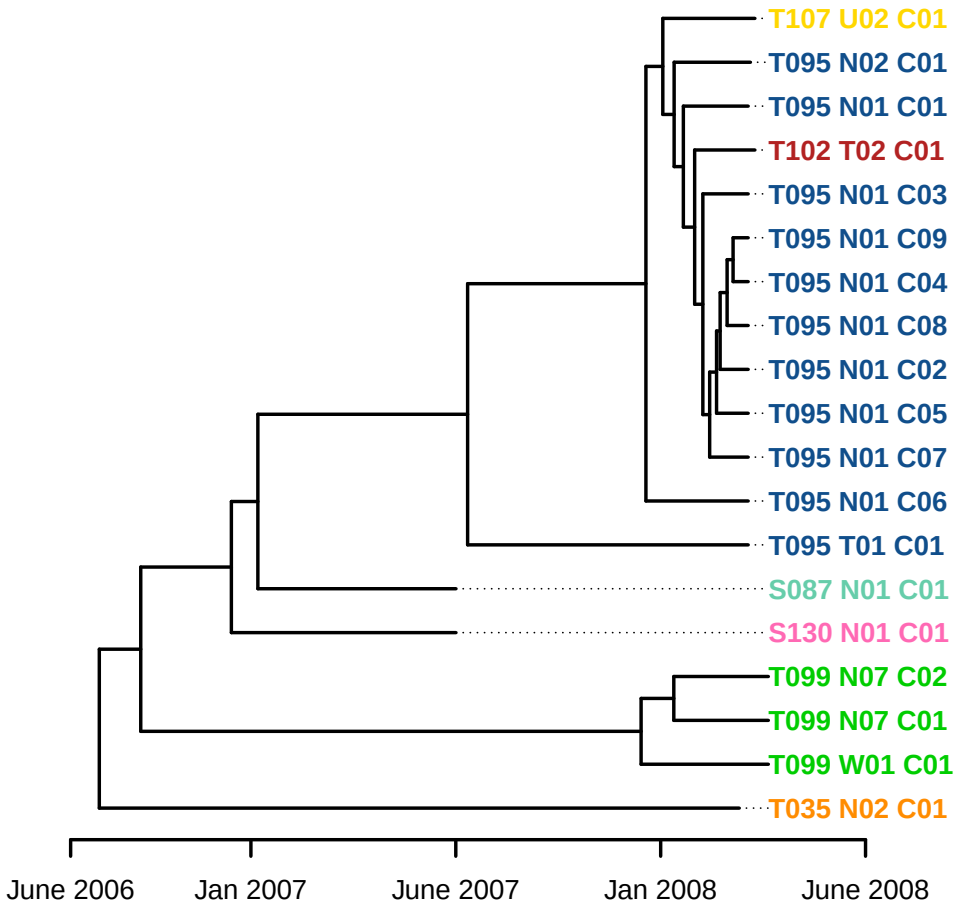

Figure 11: Group6

3.2 T099

T099 was admitted twice in the ICU. On the first admittance he was negative and did not become positive. On the second admittance on March 24 he was negative too.

He gets his first positive swab on March 31. This is $T099\_N05$. We have 7 of the 9 samples in the main clade and 2 pretty distant. The next sample we have is $T099\_N07$. Of this sample, 6 of the 9 samples cluster in the main clade, one is further apart but in the same group and 2 are in group 6, clustering with a sample coming from his wound. *(a plot of the within T099 distances would be helpful).*

It is clear that T099 was infected in the ICU with a main clade. He was infected in the wound with a different clade, which was seen in the same day also in the nose.

We would have needed more colonies from the wound to prove this story. Yet the fact that there are no sequences clustering with the wound ones before the wound was samples does not disprove it. The only registered patient procedures for T099 are *"cut down"*, "revision" and "central line" in the ward on March 19 and 20, so there is no reason why the wound bacteria should appear only on April 5. On the other hand there is only one wound swab and it is positive, so we do not know whether before there was not the wound or it was not swabbed.