Exploration of Shared SNPs in Thaps
Chr1-unfiltered

July 26, 2017

Rambling exploration of SNP positions shared between two or more of the isolates. Code is included to document it thoroughly, (even if largely uninteresting to anyone else), and I will summarize it as I go.

Contents

1 History **2**

2 Preliminaries **2**

3 Major Analysis/Performance Parameters. **2**

4 Refined SNP Calls **6**
4.1 Method . . . 6
4.2 Save them . . . 8
4.3 Examples: Consistent . . . 10
4.4 Examples: Inconsistent . . . 11
4.5 Examples: Homozygous nonref . . . 12

5 Table 1 stats **16**
5.1 Table 1 Data . . . 20

6 Shared-SNPs P-Value **20**
6.1 SNP Concordance . . . 20
6.2 Notes . . . 23
6.3 P-Value: The Bottom Line . . . 24

7 Sharing **24**
7.1 Code . . . 25
7.2 Sanity Checks . . . 26
7.3 Main Analysis . . . 31

8 Trees **35**

9 Semi-Automated Tree-Building **46**
9.1 Bootstrap . . . 52

10 Notes **55**

11 Appendix: Old Trees, etc. **56**
11.1 HWE Sharing . . . 56
11.2 Old Tree Stuff . . . 56

1 History

This was added to SVN 1/26/2014; not sure when it was started, but earliest related emails I see are from 1/21/14.

```
r413 | ruzzo | 2014-01-26 08:22:37 -0800 (Sun, 26 Jan 2014) | 2 lines

adding shared-snp analysis.
```

2 Preliminaries

NOTE: Some comments in code and some parts of the text, especially specific numbers and general conclusions, are based on Unqfiltered, Chr1, Medium stringency (i.e., "[[2]]" below) analysis. The broad picture does not appear to change with other choices, but details do, and the text is neither fully parameterized nor fully updated, so proceed with caution.

Load utility R code; do setup:

```
source('../../../R/wlr.R') # load util code; path relative this folder or sibling in scripts/larrys

## Running as: ruzzo @ bicycle.cs.washington.edu; SVN Id, I miss you.  $Id: wlr.R  2017-07-21 or later $

setup.my.wd('shared-snps') # set working dir; UPDATE if this file moves, or if COPY/PASTE to new file
setup.my.knitr('figs-knitr/')
generic.setup('figs-mine/')
```

3 Major Analysis/Performance Parameters.

Choices here control how this file is processed, what data is analyzed, speed, etc. Set them carefully before running "make." Major choices are:

1. WHICH SNP TABLES ARE LOADED??? The logical vector `load.tb` selects the desired combination of SNP tables to load, in the order `full.unfiltered`, `chr1.unfiltered`, `full.qfiltered`, `chr1.qfiltered`. E.g., `load.tb=(T, F, T, F)` loads *full* tables for *both* q- and un-qfiltered data. Primary analysis is only performed on one of them, but the others are retained for comparison/debugging.

2. WHICH MAIN ANALYSIS??? If multiple tables are loaded, which is used for the main analysis? Parameter `pri` is a permutation of 1:4, corresponding to `load.tb`; the first loaded table in that order becomes the analysis focus. The default `pri=c(1,2,3,4)` looks at un-q-filtered data in preference to q-filtered, and full tables in preference to Chr1 within each group.

   (Choice of data for the "Table 1" coverage summary in section 5 is independent of this; full genome data is prefered over Chr 1 for both q- and unq-filtered reads; change `tset.picker` calls near the end of that section to modify this.)

3. CLEAR CACHE??? `clear.cache=T` forces Knitr cache removal at the start of the run; especially important if the previous parameters have changed since the last run.

4. HOW MANY BOOTSTRAP REPLICATES??? The variable `nboot` is a major performance factor; 1000 reps takes several hours. Set to 5 for debug and quick look; 100 or more for final run.

5. TRUNCATE TABLES TO Chrs ONLY??? I.e., remove mitochondrial-, plastid-, and BD- contigs.

The following code chunk sets the first four parameters based on where it's run. To prototype/debug on a laptop, faster is better—run on Chr1 with small `nboot`; when run on the linux servers, I typically do full genomes, more replicates. Just override them if these defaults don't work for you.

```
# for Makefile, params can be command line args, else base on system; see wlr.r for details.
# load.tb order: full.un, chr1.un, full.qfil,  chr1.qfil

params <- pick.params(
  mac   = list(load.tb=c(F,T,F,F), pri=1:4, clear.cache=F, nboot=  1, trunc.tables=T), # quick on lap
 #linux = list(load.tb=c(F,F,F,T), pri=1:4, clear.cache=F, nboot=  5, trunc.tables=T), # quick qfil on server
  linux = list(load.tb=c(T,F,T,F), pri=1:4, clear.cache=T, nboot=101, trunc.tables=T)  # full on server
)

# Alternatively, edit/uncomment the following to override the above as needed
#params<-pick.params(default=list(load.tb=c(T,T,T,T),pri=1:4,clear.cache=T,nboot=1000,trunc.tables=T))
print(params)

# $load.tb
# full.unf chr1.unf  full.qf  chr1.qf
#    FALSE     TRUE    FALSE     TRUE
#
# $pri
# [1] 1 2 3 4
#
# $clear.cache
# [1] TRUE
#
# $nboot
# [1] 5
#
# $trunc.tables
# [1] FALSE
```

CLEAR CACHE??!! Some code chunks use the knitr cache, but extent of cache consistency checks unknown. If in doubt, delete "cache/" (knitr's) directory to force rebuild. T/F set in params above will/won't force removal (actually, rename):

```
decache(params$clear.cache)

# Rename of 'cache' to 'cache96093' returned TRUE .
```

If still in doubt, also manually remove "00common/mycache/" (mine).

Load the main SNP data file(s) based on the parameters set in section 3.

```
# short names to keep the following chunk compact
tb <- params$load.tb
tset <- list(NULL, NULL, NULL, NULL) # tset = 'table set'
```

```
# see wlr.R for load paths
if(tb[1]){tset[[1]] <- load.snp.tables(use.chr1.tables = FALSE, data.name='full.tables.01.26.14')}
if(tb[2]){tset[[2]] <- load.snp.tables(use.chr1.tables = TRUE , data.name='full.tables.01.26.14')}

# Loading ../00common/mycache/snp.tables.chr1.unqfiltered.rda ...Loaded.

if(tb[3]){tset[[3]] <- load.snp.tables(use.chr1.tables = FALSE, data.name='full.tables.02.25.15')}
if(tb[4]){tset[[4]] <- load.snp.tables(use.chr1.tables = TRUE , data.name='full.tables.02.25.15')}

# Loading ../00common/mycache/snp.tables.chr1.qfiltered.rda ...Loaded.
# Bandaiding qfiltered tables...
```

Grrr! I should have excluded non-Chr contigs from full genome runs. Rather than change tons of code below to add mask params, I'm just going to truncate the tables, as follows. (See notes in wlr.r::make.mask for assumptions.)

```
if(params$trunc.tables){
  for(i in 1:4){
    if(!is.null(tset[[i]])){
      first.mito <- match("mitochondria.fasta", tset[[i]][[7]]$Chr)
      if(!is.na(first.mito)){ # will be NA for Chr1 tables
```

```
        for(j in 1:7){
          # hmmm... slow; wonder whether head(tset[[i]][[j]],first.mito-1) is faster;
          # ok, simple tests suggest not: system.time(head(data.frame(1:1e7,1:1e7),5e6))
          tset[[i]][[j]] <- tset[[i]][[j]][1:(first.mito-1),]
        }
      }
    }
  }
} else {
    cat('***\n*** DID YOU *REALLY* WANT UNTRUNCATED TABLES???\n***\n')
}

# ***
# *** DID YOU *REALLY* WANT UNTRUNCATED TABLES???
# ***
```

The tersely-named `tset` list is sometimes convenient, but give them more descriptive names, too.

```
snp.tables.full.unfiltered <- tset[[1]]; names(tset)[1] <- 'snp.tables.full.unfiltered'
snp.tables.chr1.unfiltered <- tset[[2]]; names(tset)[2] <- 'snp.tables.chr1.unfiltered'
snp.tables.full.qfiltered  <- tset[[3]]; names(tset)[3] <- 'snp.tables.full.qfiltered'
snp.tables.chr1.qfiltered  <- tset[[4]]; names(tset)[4] <- 'snp.tables.chr1.qfiltered'
```

Main analysis may just use one of the potentially 4 table sets. Pick it according to the priority specified in section 3, using the shorter name ’snp.tables’ for this default choice.

```
snp.tables <- tset.picker(priority=params$pri, table.set=tset)
```

```
# Sanity check: unlike unqfiltered tables, bug in early code gave qfiltered ones different numbers
# of rows per strain, which breaks much code.  Verify this is no longer happening.
check.eq.nrows <- function(tables){
  if(!is.null(tables)){
    nrow.snp.tables <- unlist(lapply(tables,nrow))
    print(nrow.snp.tables)
    if(all(nrow.snp.tables == nrow.snp.tables[1])){
      cat('OK, all strains have same number of rows.\n')
    } else {
      cat('***\n*** Warning: Different strains have different numbers of rows! ***\n***\n')
    }
  }
}

dummy<-lapply(tset, check.eq.nrows)

#    1007    1012    1013    1014    1015    3367    1335
# 3042585 3042585 3042585 3042585 3042585 3042585 3042585
# OK, all strains have same number of rows.
#    1007    1012    1013    1014    1015    3367    1335
# 3042585 3042585 3042585 3042585 3042585 3042585 3042585
# OK, all strains have same number of rows.
```

Which tables have we got?:

```
# 'which.snp.tables' return summary of which tables, either as a char string (default), e.g.
# "Chr1-qfiltered", or as vector of 2 strings, e.g. c("full","unfiltered").
cat('This analysis uses: (', paste(unlist(lapply(tset,which.snp.tables)),collapse=', '), ') SNP tables.\n')

# This analysis uses: ( NULL, Chr1-unfiltered, NULL, Chr1-qfiltered ) SNP tables.

cat('Main shared SNP analysis focuses on', which.snp.tables(snp.tables), '\n')

# Main shared SNP analysis focuses on Chr1-unfiltered
```

A LaTeX hack: I want which.snp.tables info in doc title/page headers, but it is unknown until now, so the following writes a command definition `\whichsnptables` into the .aux file, which is read during the *next* LaTeX run, when `\begin{document}` is processed:

```
\makeatletter
\immediate\write\@auxout{\noexpand\gdef\noexpand\whichsnptables{Chr1-unfiltered}}
\makeatother
```

Subsequent analysis was initially all directed at Chr1. In general, I have *not* updated the discussion to reflect genome-wide analysis.

```
if(exists('snp.tables.chr1.qfiltered') && exists('snp.tables.chr1.unqfiltered')){
  # If have both, where is new unequal to old?
  uneq <- snp.tables.chr1.qfiltered[[1]]$Ref[1:chr1.len] != snp.tables.chr1.unqfiltered[[1]]$Ref[1:chr1.len]
  cat('Sum uneq:', sum(uneq,na.rm=T), '\n')
  cat('Sum NA:  ', sum(is.na(uneq)),  '\n')
  print(which(is.na(uneq))[1:10])
  seecounts(which(is.na(uneq))[1:4],snp.tables=snp.tables.qfiltered,debug=F)
}
```

In brief, “`snp.tables`” will be a list of 7 data frames, one per strain, giving read counts for each nucleotide at each position, SNP calls, etc.:

```
names(snp.tables)

# [1] "1007" "1012" "1013" "1014" "1015" "3367" "1335"

str(snp.tables[[1]])

# 'data.frame': 3042585 obs. of  15 variables:
#  $ chr   : Factor w/ 66 levels "BD10_65","BD11_74",..: 39 39 39 39 39 39 39 39 39 39 ...
#  $ pos   : int  1 2 3 4 5 6 7 8 9 10 ...
#  $ snp   : int  0 0 0 0 0 0 0 0 0 0 ...
#  $ Chr   : chr  "Chr1" "Chr1" "Chr1" "Chr1" ...
#  $ Pos   : int  1 2 3 4 5 6 7 8 9 10 ...
#  $ Ref   : chr  "T" "C" "C" "A" ...
#  $ Cov   : num  1 3 4 5 7 7 10 12 13 15 ...
#  $ a     : num  0 0 1 0 0 0 0 0 1 0 ...
#  $ g     : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ c     : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ t     : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ n     : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ .match: num  1 3 3 5 7 7 10 12 12 15 ...
#  $ exon  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
#  $ indel : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Just for background, also load the desert tables:

```
# from svn+ssh://ceg1.ocean.washington.edu/var/svn/7_strains/trunk/code/snpNB/data
#load('../../../data/ungit-data/des.rda')
load('../../../data/des.rda')
```

What’s the total length of all deserts in each strain? Big deserts (defined as “big.threshold” or longer)?

```
some.desert.stats <- function(big.threshold=0){
  desert.len <- unlist(lapply(des,function(x){sum(unlist(lapply(x,function(y){sum(y[,'Length'])})))}))
  bigdes.len <- unlist(lapply(des,function(x){sum(unlist(lapply(x,function(y){
                                          sum(y[y[,'Length']>=big.threshold,'Length'])})))}))
  rbind(desert.len, desert.pct=round( desert.len / genome.length.constants()$genome.length.trunc * 100),
        bigdes.len, bigdes.pct=round( bigdes.len / genome.length.constants()$genome.length.trunc * 100))
}
some.desert.stats(big.threshold=50000)

#              tp1007   tp1012  tp1013  tp1014   tp1015 thapsIT   tp1335
# desert.len 11146526 11332566 5801763 9464213 11251426 6780300 10883723
# desert.pct       36       36      19      30       36      22       35
# bigdes.len  3495805  3936973   55365 3627235  3727061   57119  4046934
# bigdes.pct       11       13       0      12       12       0       13
```

I.e., looking at all deserts, about 1/3 of L-clade, 1/5 of H-clade are in deserts, whereas, looking at the largest deserts ($> 50k$), only about 12% in L-clade (and none in H-clade). Note that the rough stats above include artifactual "deserts" created by gaps in the reference sequence, large genomic deletions, etc. A more careful analysis of this is found in nc-snps.rnw.

4 Refined SNP Calls

4.1 Method

It is appropriate that SNP calls should be conservative, to avoid many false positives, but, when a position is called a SNP in one isolate, we often see a significant number of reads for the same non-reference nucleotide at that position in other isolates, even if they are not called as SNPs. On the other hand, we sometimes see a position called a SNP in two or more isolates, but with *different* pairs of nucleotides, potentially suggesting technical errors. Analysis in this section attempts to refine the SNP calls by looking for issues such as these by looking at all 7 isolates jointly, at each position called a SNP in any of them.

For a given strain, the following function returns a vector of 0:4 to indicate which nonreference nucleotide has the maximum read count at the corresponding position. The values 1..4 indicate that the max count occurred at A, G, C, T, resp. (Ties are resolved arbitrarily ($a < g < c < t$), which possibly deserves further attention.) The value 0 means all nonreference counts are below threshold, based *either* on absolute count *or* as a fraction of coverage. Default only excludes 0 counts.

```
nref.nuc.new <- function(strain=1, mask=T, thresh.count=0, thresh.rate=0.0){
        # get read count for max nonref nuc
        nref <- apply(snp.tables[[strain]][mask, c('a', 'g', 'c', 't')], 1, max)
        # where does nref count match a (g,c,t, resp) count
        as <- ifelse(nref == snp.tables[[strain]][mask,'a'],1,0)
        gs <- ifelse(nref == snp.tables[[strain]][mask,'g'],2,0)
        cs <- ifelse(nref == snp.tables[[strain]][mask,'c'],3,0)
        ts <- ifelse(nref == snp.tables[[strain]][mask,'t'],4,0)
        # most positions will show 3 zeros and one of 1:4, so max identifies max nonref count;
        # ties broken arbitrarily  (a<g<c<t)
        merge <- pmax(as,gs,cs,ts)
        # but if max nonref count is zero or below threshold, return 0
        merge[nref == 0 | nref < thresh.count] <- 0
        merge[nref/snp.tables[[strain]][mask,'Cov'] < thresh.rate] <- 0
        return(merge)
}
```

Get union and intersection of the sets of called SNPs. ("$snp" is 0/1.) Also, 5-way (L-clade) and 4-way (L-excluding Gyre).

```
# 4-way union/intersection
u4.snps <- snp.tables[[1]]$snp
i4.snps <- snp.tables[[1]]$snp
for(i in c(2,5,7)) {
        u4.snps <- pmax(u4.snps, snp.tables[[i]]$snp)
        i4.snps <- pmin(i4.snps, snp.tables[[i]]$snp)
}
# 5-way: add gyre
u5.snps <- pmax(u4.snps, snp.tables[[4]]$snp)
i5.snps <- pmin(i4.snps, snp.tables[[4]]$snp)
# 7-way
union.snps     <- pmax(u5.snps, snp.tables[[3]]$snp, snp.tables[[6]]$snp)
intersect.snps <- pmin(i5.snps, snp.tables[[3]]$snp, snp.tables[[6]]$snp)
nu4snps <- sum(u4.snps)
nu5snps <- sum(u5.snps)
ni4snps <- sum(i4.snps)
ni5snps <- sum(i5.snps)
nusnps  <- sum(union.snps)
nisnps  <- sum(intersect.snps)
c(n4u=nu4snps, n5u=nu5snps, n7u=nusnps, n4i=ni4snps, n5i=ni5snps, n7i=nisnps)
```

```
#   n4u   n5u   n7u   n4i   n5i   n7i
# 18564 18696 47499 14365  7628  1641
```

There are nusnps=47499 positions called as SNPs in one or more strains (but only nisnps=1641 that are shared among all 7). Note that the 4-way union is only modestly larger (1.2923077 times larger) than the 4-way intersection, emphasizing the inherent similarities among these SNP sets. The corresponding 5-way numbers show that Gyre adds relatively little to the 5-way union vs the 4-way union, whereas it removes a fair bit from the 5-way intersection. However, much of that loss is simply because Gyre has fewer called SNPs: only 8331 vs 14365 in the 4-way intersection, and they are highly concordant:

```
sum(snp.tables[[4]]$snp*i4.snps)/sum(snp.tables[[4]]$snp)

# [1] 0.9156164
```

So, a likely source of the Gyre's difference in called SNPs is technical (lower read coverage, higher read error rate) rather than biological.

Inclusion of the 2 H-clade members, however, causes more dramatic changes in both union and intersection numbers. I examine all these relationships in more detail below, but first I examine what I believe to be a significant source of technical error in these comparisons—erroneous SNP calls, especially false negative calls.

It is appropriate that SNP calls should be conservative, to avoid many false positives, but, if a position is called a SNP in one strain, we often see a significant number of reads for the same non-reference nucleotide at that position in other strains, even if they are not called as SNPs. For my purposes below, these will be considered "shared SNPs," based on three different levels of permissiveness. Note that, e.g., $\geq$ 84% of all positions have zero reads for any non-reference nucleotide, and only a small fraction have 2 or more non-reference reads:

```
nonmatch <- rbind(
  unlist(lapply(snp.tables,function(x){sum(x$Cov-x$.match == 0)})),
  unlist(lapply(snp.tables,function(x){sum(x$Cov-x$.match == 1)})),
  unlist(lapply(snp.tables,function(x){sum(x$Cov-x$.match == 2)})),
  unlist(lapply(snp.tables,function(x){sum(x$Cov-x$.match == 3)})),
  unlist(lapply(snp.tables,function(x){sum(x$Cov-x$.match >= 4)})),
  unlist(lapply(snp.tables,function(x){sum((x$Cov-x$.match)[union.snps==0] >= 4)}))
)/nrow(snp.tables[[1]])*100
rownames(nonmatch) <- c('% ==0','% ==1','% ==2','% ==3','% >=4', '% >=4, nonSNP')
nonmatch

#                        1007       1012       1013        1014       1015       3367       1335
# % ==0           92.49473721 88.8682485 84.9571992 86.51028648 90.6336881 86.4695974 85.2300922
# % ==1            6.37162150  9.3939528 12.1458891 11.81896315  7.9041670 10.9489792 11.6810541
# % ==2            0.41921590  0.8698853  1.3586145  1.09222914  0.6199991  1.1349231  1.7612326
# % ==3            0.07950476  0.1349182  0.2267480  0.17849953  0.1118128  0.2029196  0.4273997
# % >=4            0.63492063  0.7329951  1.3115492  0.40002169  0.7303329  1.2435807  0.9002214
# % >=4, nonSNP    0.06267697  0.1274903  0.2943878  0.06037629  0.1217715  0.2716440  0.3085534
```

Build a table of max non-reference nucleotides at each position in the union.snps set. The three criteria are

- [[1]]: any non-zero count at any coverage is considered significant
- [[2]]: (count $\geq$ 2 and count/coverage $\geq$ 0.05) is considered significant
- [[3]]: (count $\geq$ 4 and count/coverage $\geq$ 0.10) is considered significant

In all three cases, the nonref nucleotide must also be consistent across all strains passing that threshold; see below.

```
non.refs <- vector('list',4)
for(i in 1:4){
  non.refs[[i]] <- matrix(0, nrow=nusnps, ncol=7)
  colnames(non.refs[[i]]) <- names(snp.tables)
  rownames(non.refs[[i]]) <-
             paste(snp.tables[[1]]$chr[union.snps==1], ':', snp.tables[[1]]$pos[union.snps==1], sep='')
}
for(j in 1:7){
  non.refs[[1]][,j] <- nref.nuc.new(j, mask=union.snps==1, thresh.count=0, thresh.rate=0.00)
  non.refs[[2]][,j] <- nref.nuc.new(j, mask=union.snps==1, thresh.count=2, thresh.rate=0.05)
  non.refs[[3]][,j] <- nref.nuc.new(j, mask=union.snps==1, thresh.count=4, thresh.rate=0.10)
}
```

For comparison, I want to look at unfiltered SAMTools SNP calls. In complete opposition to the measures of consistency imposed above, I'm going to simply force this into the "non.refs" structure constructed above by imagining that any position called a SNP in any strain has its max nonref count on "A", so any given position called a SNP in any strain will automatically be declared "consistent." This will allow the tree-code, etc. given below to work in a uniform way (even though interpretation of the results is different.) Results will be jammed into a 4th component of the "non.refs" list; i.e., we have a 4th criterion:

- [[4]]: all called SNPs at a given position are considered "consistent."

As this case was a late addition to the analysis, the commentary throughout this document has not necessarily been updated to reflect that this case is distinct from the first three.

```
for(j in 1:7){
  non.refs[[4]][,j] <- snp.tables[[j]]$snp[union.snps==1]
}
```

```
str(non.refs[[4]])


#  num [1:47499, 1:7] 0 0 0 0 0 0 0 0 1 0 ...
#  - attr(*, "dimnames")=List of 2
#   ..$ : chr [1:47499] "Chr1:333" "Chr1:417" "Chr1:435" "Chr1:438" ...
#   ..$ : chr [1:7] "1007" "1012" "1013" "1014" ...
```

"non.refs" indicates, among those positions in the union of all called SNPS having any non-reference read count above the thresholds listed above, the non-ref nucleotide having the highest read count in each strain. If, for a given position, the max of this code is the same as the min (among non-zero values), then every strain having over-threshold nonref reads in that position, in fact has most non-reference reads on the *same* nucleotide. These are defined as the "consistent" SNPs.

```
find.consistent <- function(nr){
  nr.max <- apply(nr,1,max)
  nr.min <- apply(nr,1,function(x){ifelse(max(x)==0,0,min(x[x>0]))})
  return(nr.min == nr.max)
}
consistent  <- lapply(non.refs, find.consistent)
```

4.2 Save them

```
# wrap this in a data structure to be cached:
Description <- [2753 chars quoted with ''']

refined.snps <-
  list(Description=Description,

       Data=list(
         based.on.which.snp.tables=which.snp.tables(),
         number.union.snps=nusnps,
         number.intersection.snps=nisnps,
         non.ref.nucleotide=non.refs,
         consistent.snps=consistent),

       Code=list(
         get.snps = function(strain, stringency=2){
           # return nusnps x 1 Bool vector of consistent SNPs @ specified stringency & strain
           return(refined.snps$Data$consistent.snps[[stringency]] &
                  refined.snps$Data$non.ref.nucleotide[[stringency]][,strain] > 0)
         }
         ,
         get.snp.locs.char = function(strain, stringency=2){
           # return char vector of locations of consistent SNPs @ specified stringency & strain
           snps <- refined.snps$Code$get.snps(strain, stringency)
           return(names(snps)[snps])
         }
         ,
         get.snp.locs.df = function(strain, stringency=2){
           # return data frame (Chr/Pos) of locations of consistent SNPs @ specified stringency & strain
           snplist <- strsplit(refined.snps$Code$get.snp.locs.char(strain, stringency), ':', fixed=TRUE)
           # strsplit returns long list of 2-vectors, 1st=chr, 2nd=char position
           df <- data.frame(Chr=           unlist(lapply(snplist,function(x){return(x[1])})),
                            Pos=as.integer(unlist(lapply(snplist,function(x){return(x[2])}))),
                            stringsAsFactors = FALSE)
           return(df)
         }
       )
  )

# dont't clobber existing .rda, but save if absent.  (delete to re-save)
# result for trunc, unfiltered tables saved to "data" else "mycache"
if(which.snp.tables() == 'trunc-unfiltered'){
  rda.refined <- '../../../data/refined.snps-trunc-unfiltered.rda'
} else {
  rda.refined <- paste('../00common/mycache/refined.snps', which.snp.tables(), 'rda', sep='.')
}
if(file.exists(rda.refined)){
  cat('Pre-existing file', rda.refined, 'unchanged.\n')
} else {
  cat('Saving', rda.refined, '...')
  save(refined.snps, file=rda.refined, compress=TRUE)
  cat('Saved.\n')
}

# Saving ../00common/mycache/refined.snps.Chr1-unfiltered.rda ...Saved.
```

Knitr seems to be failing to format the long char string above, which says:

```
cat(refined.snps$Description)

# Contents of this .rda file:
#
#   * Description: this text
#
#   * Data -- 5 items defining refined SNPs, at 4 different stringency levels, as defined
#     in shared-snps.rnw:
#
#     * based.on.which.snp.tables: {"Chr1","full","trunc"}-{"unfiltered","qfiltered"},
#       depending on which snp tables were used to build this data. ("trunc" = all Chrs.)
#
#     * number.union.snps: the total number of SNPs (SAMtools calls) in the union of SNPs
#       across all 7 strains.
#
#     * number.intersection.snps: similar, for the 7-way intersection.
#
#       nusnps/nisnps are easily recalculated from the data below, but their inclusion
#       may be convenient, e.g., to quickly see if the .rda represents the full genome
#       (nusnps=488848), or the chr 1 subset (nusnps=47499); (redundant with "based.on...";
#       numbers above are for unfiltered, perhaps slightly different if qfiltered)
#
#     * non.ref.nucleotide: 4 arrays, each nusnps x 7, of values 0..4 (0..1 in the 4th
#       array).  In the 1st 3 arrays, 0 means the given position in the given strain did
#       not have nonreference read counts above the corresponding filtering threshold,
#       i.e., is NOT a refined SNP in that strain, whereas 1..4 mean that it did pass
#       threshold, for A,C,G,T resp.  In the 4th array, this value is just 1/0,
#       indicating is/is not a called SNP in that strain.
```

```
#
#     * consistent.snps: 4 Bool vectors of length nusnps flagging positions whose nonref
#       nucs (wrt to the 4 filtering criteria) are deemed *consistent* across
#       all 7 strains.  For the 1st 3, this means all nonzero entries of non.ref.nuc
#       are equal, i.e., nonref read counts passing threshold are on the SAME nonref
#       nucleotide in all strains having over-threshold counts.  Just for comparison
#       and uniformity of data structures, the 4th is all TRUE, i.e., union of SNPs
#       across all strains, without any regard for thresholds or consistency.
#
#       In short, the refined SNPs according to our medium filtering criteria are
#       strains/positions where consistent.snps[[2]]==TRUE and non.ref.nucleotide[[2]]>0.
#
#       Rownames in both non.ref.nucs and consistent define location, e.g. "Chr1:333".
#
#   * Code -- simple routines to extract refined SNPs in (potentially) convenient formats:
#
#     * get.snps(strain, stringency=2)
#       returns nusnps x 1 Bool vector of consistent SNPs @ specified stringency in
#       given strain
#
#     * get.snp.locs.char(strain, stringency=2)
#       returns n x 1 char vector of locations of consistent SNPs @ specified stringency
#       in given strain, e.g. "Chr1:1234", where n == sum(get.snps(...))
#
#     * get.snp.locs.df(strain, stringency=2){
#       As above, but returns data frame (char vector Chr, int vector Pos) with the same info.
```

```
str(consistent[[1]])

#  Named logi [1:47499] TRUE FALSE TRUE FALSE TRUE TRUE ...
#  - attr(*, "names")= chr [1:47499] "Chr1:333" "Chr1:417" "Chr1:435" "Chr1:438" ...
```

```
consistent.count <- unlist(lapply(consistent, sum)) ; consistent.count

# [1] 36040 46896 47174 47499

inconsistent.count <- consistent.count[4] - consistent.count; inconsistent.count

# [1] 11459   603   325     0

inconsistent.percent <- inconsistent.count/consistent.count[4]*100; inconsistent.percent

# [1] 24.1247184  1.2695004  0.6842249  0.0000000
```

I.e., of the 47499 positions in which a SNP is called, 36040 are consistent by my loose definition, and 47174 are consistent by my tightest definition. The increase in concordance supports the view that the loose definition is too loose. Perhaps misleadingly, these counts include positions that are "consistent SNPs" in only one strain; more below. (*TODO* I suspect, but have not yet systematically checked, that most of the rest are positions with low coverage and/or very low read counts on the mixture of non-reference nucleotides.)

4.3 Examples: Consistent

Here are a few (nonrandomly selected) prototypical consistent SNPs:

```
esnps <- names(consistent[[2]][consistent[[2]]])
esnps2 <- as.integer(unlist(lapply(strsplit(esnps[c(7,11:13,92)],':',fixed=TRUE),function(x){x[2]})))
seecounts(esnps2,snp.tables=snp.tables)

#     chr  pos Ref Strain  A   G  C   T SNP  exon indel nrf rat
# 1  Chr1  567   T
# 2                  1007  0   0  2  29   0  TRUE FALSE
# 3                  1012  0   0 15  44   1  TRUE FALSE
# 4                  1013  0   0 15  97   0  TRUE FALSE
# 5                  1014  0   1  0  33   0  TRUE FALSE
# 6                  1015  0   0  9  43   1  TRUE FALSE
# 7                  3367  0   0 16  46   1  TRUE FALSE
# 8                  1335  0   0  2 116   0  TRUE FALSE
# 9  Chr1 1053   A
# 10                 1007 39   0  0   5   0  TRUE FALSE
# 11                 1012 55   0  0  12   0  TRUE FALSE
# 12                 1013 17   1  0  40   0  TRUE FALSE
# 13                 1014 25   0  0   5   0  TRUE FALSE
# 14                 1015 38   0  1  20   1  TRUE FALSE
# 15                 3367 13   0  0   9   0  TRUE FALSE
# 16                 1335 71   1  0  46   1  TRUE FALSE
# 17 Chr1 1055   G
# 18                 1007  0  41  0   2   0  TRUE FALSE
# 19                 1012  1  63  0   8   0  TRUE FALSE
# 20                 1013  1  62  0   8   0  TRUE FALSE
# 21                 1014  1  26  0   8   1  TRUE FALSE
# 22                 1015  0  44  0  14   0  TRUE FALSE
# 23                 3367  0  27  0   0   0  TRUE FALSE
# 24                 1335  0  78  0  40   1  TRUE FALSE
# 25 Chr1 1176   G
# 26                 1007  2  67  0   0   0 FALSE FALSE
# 27                 1012  1  68  0   0   0 FALSE FALSE
# 28                 1013 29  73  0   0   0 FALSE FALSE
# 29                 1014  1  52  0   0   0 FALSE FALSE
# 30                 1015  4 103  0   0   0 FALSE FALSE
# 31                 3367 11   8  0   0   1 FALSE FALSE
# 32                 1335  1 206  0   0   0 FALSE FALSE
# 33 Chr1 8670   A
# 34                 1007 19   0  0   7   0  TRUE FALSE
# 35                 1012 36   0  0  12   0  TRUE FALSE
# 36                 1013 44   0  0  12   0  TRUE FALSE
# 37                 1014 10   0  0   7   0  TRUE FALSE
# 38                 1015 24   0  0  11   1  TRUE FALSE
# 39                 3367 18   0  0   0   0  TRUE FALSE
# 40                 1335 27   0  0   6   0  TRUE FALSE
```

4.4 Examples: Inconsistent

Here is a brief look at some *in*-consistent positions. E.g., Chr1:2013 shows nontrivial counts on 3 alleles in Wales, as do 2319, 3286, 5002, 5433, whereas 7878 shows a different alternate allele in Italy than in Wales.

```
unc <- names(consistent[[2]][!consistent[[2]]])
unc2 <- as.integer(unlist(lapply(strsplit(unc[1:10],':',fixed=TRUE),function(x){x[2]})))
seecounts(unc2,snp.tables=snp.tables)
```

```
#     chr   pos Ref Strain  A   G   C  T SNP  exon indel nrf rat
# 1  Chr1  2013   T
# 2                   1007  4   0   0 20   0  TRUE FALSE
# 3                   1012  8   0   0 34   0  TRUE FALSE
# 4                   1013  9  12   0 16   1  TRUE FALSE
# 5                   1014  1   0   0 19   0  TRUE FALSE
# 6                   1015 13   0   0 24   1  TRUE FALSE
# 7                   3367 10   0   0 36   0  TRUE FALSE
# 8                   1335 20   0   0 68   1  TRUE FALSE
# 9  Chr1  2319   C
# 10                  1007  0  29  22  0   1  TRUE FALSE
# 11                  1012  0  54  26  0   1  TRUE FALSE
# 12                  1013 19  19  18  0   1  TRUE FALSE
# 13                  1014  0  25  19  0   1  TRUE FALSE
# 14                  1015  0  54  29  0   1  TRUE FALSE
# 15                  3367  5   0  43  0   0  TRUE FALSE
# 16                  1335  0 132  48  0   1  TRUE FALSE
# 17 Chr1  3286   T
# 18                  1007  4   0   1 17   0  TRUE FALSE
# 19                  1012  9   0   3 45   0  TRUE FALSE
# 20                  1013 39   1  38 12   1  TRUE FALSE
# 21                  1014  4   0   6 27   0  TRUE FALSE
# 22                  1015 11   0   7 37   0  TRUE FALSE
# 23                  3367  8   0  39 10   0  TRUE FALSE
# 24                  1335 15   0   4 75   0  TRUE FALSE
# 25 Chr1  5002   T
# 26                  1007  0  15   0 12   0  TRUE FALSE
# 27                  1012  1  23   0 26   1  TRUE FALSE
# 28                  1013 21  11   0 39   0  TRUE FALSE
# 29                  1014  0   8   0 12   0  TRUE FALSE
# 30                  1015  0  19   0 16   1  TRUE FALSE
# 31                  3367  0   0   0 35   0  TRUE FALSE
# 32                  1335  0  57   0 60   0  TRUE FALSE
# 33 Chr1  5433   G
# 34                  1007  0  50   0  3   0  TRUE FALSE
# 35                  1012  0  78   0  5   0  TRUE FALSE
# 36                  1013 18  47   0 14   1  TRUE FALSE
# 37                  1014  9  19   0  0   1  TRUE FALSE
# 38                  1015  7  63   0  2   0  TRUE FALSE
# 39                  3367  8  54   0  0   0  TRUE FALSE
# 40                  1335  6 109   0  4   0  TRUE FALSE
# 41 Chr1  7858   C
# 42                  1007  0   0  48  0   0  TRUE FALSE
# 43                  1012  0   1  61  0   0  TRUE FALSE
# 44                  1013  0   0 131 10   0  TRUE FALSE
# 45                  1014  0   0  34  0   0  TRUE FALSE
# 46                  1015  0   0  74  0   0  TRUE FALSE
# 47                  3367 20   0   8  0   1  TRUE FALSE
# 48                  1335  0   0 120  0   0  TRUE FALSE
# 49 Chr1  8914   A
# 50                  1007 23   0   0  2   0  TRUE FALSE
# 51                  1012 29   0  15  0   1  TRUE FALSE
# 52                  1013 25   0   6  0   0  TRUE FALSE
# 53                  1014 22   0   0  0   0  TRUE FALSE
# 54                  1015 31   0   5  2   0  TRUE FALSE
# 55                  3367  8   0   0  1   0  TRUE FALSE
# 56                  1335 68   0   7  0   0  TRUE FALSE
# 57 Chr1  8974   C
# 58                  1007  0   2   6  0   0  TRUE FALSE
# 59                  1012  0   2  17  0   0  TRUE FALSE
# 60                  1013 10  22   4  0   1  TRUE FALSE
# 61                  1014  0   1  10  0   0  TRUE FALSE
# 62                  1015  0   2  15  0   0  TRUE FALSE
# 63                  3367  2   0   3  0   0  TRUE FALSE
# 64                  1335  0  11  49  0   0  TRUE FALSE
# 65 Chr1 10099   T
# 66                  1007 17   0   0 29   0  TRUE FALSE
# 67                  1012 48   0   0 36   0  TRUE FALSE
# 68                  1013  0   2   6 68   0  TRUE FALSE
# 69                  1014 34   0   0 26   0  TRUE FALSE
# 70                  1015 41   0   0 38   0  TRUE FALSE
# 71                  3367  0   1   0 14   0  TRUE FALSE
# 72                  1335 55   0   0 68   1  TRUE FALSE
# 73 Chr1 15154   A
# 74                  1007 25   0   0  0   0 FALSE FALSE
# 75                  1012 56   0   0  1   0 FALSE FALSE
# 76                  1013 10   0  38 10   1 FALSE FALSE
# 77                  1014 26   0   0  0   0 FALSE FALSE
# 78                  1015 37   0   0  0   0 FALSE FALSE
# 79                  3367 19   0   0 13   1 FALSE FALSE
# 80                  1335 70   0   0  3   0 FALSE FALSE
```

4.5 Examples: Homozygous nonref

And at some *homozygous nonreference* positions (defined to be those with nonref fraction > 0.75):

```
hnr <- lapply(snp.tables, function(x){x$.match/x$Cov < 0.25})      # find them
hnr <- lapply(hnr, function(x){ifelse(is.na(x),FALSE,x)})           # remove NA
unlist(lapply(hnr,sum))                                             # count per strain

# 1007 1012 1013 1014 1015 3367 1335
#   65   65 6281   23   62 7071   40
```

Hmm, in L-clade, excluding the ref isolate (1335) this tracks time-in culture to some degree; Maybe many of these are in hemizygous regions. Next two chunks lifted from nc-snps to get tables for hemi-deletion.

```
cnv.chronly <- load.cnv.tables('../../../data/cnv.txt', chrs.only=TRUE)

str(cnv.chronly)

# 'data.frame': 1956 obs. of  11 variables:
#  $ strain   : Factor w/ 7 levels "IT","tp1007",..: 3 3 3 3 3 3 3 3 3 3 ...
#  $ chr      : Factor w/ 65 levels "BD1_7","BD10_65",..: 38 38 38 38 38 38 38 38 38 38 ...
#  $ start    : int  10601 112001 215001 358901 536501 554801 673401 781801 806901 853201 ...
#  $ end      : int  13500 116500 221100 370300 538600 559300 685000 787400 811100 855600 ...
#  $ length   : int  2900 4500 6100 11400 2100 4500 11600 5600 4200 2400 ...
#  $ filtered : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
#  $ type     : Factor w/ 1 level "CNVnator": 1 1 1 1 1 1 1 1 1 1 ...
#  $ cov_ratio: num  0.63738 1.54893 1.65381 0.00204 0.68486 ...
#  $ dup_frac : num  0.41188 0.00908 0.01178 0.97997 0.0211 ...
#  $ iStart   : num  10601 112001 215001 358901 536501 ...
#  $ iEnd     : num  13500 116500 221100 370300 538600 ...

cnv.chronly[c(1:4,nrow(cnv.chronly)+c(-1,0)),]                     ## first/last few rows

#      strain   chr  start    end length filtered     type  cov_ratio   dup_frac   iStart     iEnd
# 1    tp1012  Chr1  10601  13500   2900    FALSE CNVnator 0.63738000 0.41187900    10601    13500
# 2    tp1012  Chr1 112001 116500   4500    FALSE CNVnator 1.54893000 0.00907677   112001   116500
# 3    tp1012  Chr1 215001 221100   6100    FALSE CNVnator 1.65381000 0.01178470   215001   221100
# 4    tp1012  Chr1 358901 370300  11400     TRUE CNVnator 0.00204431 0.97997300   358901   370300
# 1955 tp1335 Chr24 259901 278000  18100    FALSE CNVnator 1.41458000 0.38091100 31264334 31282433
# 1956 tp1335 Chr24 286901 289800   2900    FALSE CNVnator 1.74941000 0.74228100 31291334 31294233
```

```
get.cnv.dels <- function(cov.thresh.lo = 0.0,
                         cov.thresh.hi = 0.8,
                         cnv,
                         snp.tables = NULL,
                         DEBUG = FALSE
){
  # build list of 7 Bool vectors of genome length, with i-th == T iff
  #  * i-th pos is 'NA' in genome seq (if snp.tables are provided), or
  #  * in CNVnator call for coverage in half-open [cov.thresh.lo, hi), and
  #  * not marked 'filtered' by CNVnator
  cnv.deletions <- vector(mode='list',7)                # make list of bool vectors
  if(is.null(snp.tables)){
    # if no tables, assume full
    t.len <- genome.length.constants()$genome.length.trunc
  } else {
    t.len <- nrow(snp.tables[[1]])
  }
  for(st in 1:7){
    if(is.null(snp.tables)){
      cnv.deletions[[st]] <- logical(t.len)                        # all F
    } else  {
      cnv.deletions[[st]] <- is.na(snp.tables[[st]]$Pos[1:t.len])  # NA positions in genome
    }
  }
  strain.names <- c(paste('tp10',c('07',12:15),sep=''),'IT','tp1335')
  names(cnv.deletions) <- strain.names
  for(i in 1:nrow(cnv)){
    if(!cnv$filtered[i] &&
       cnv$cov_ratio[i] >= cov.thresh.lo &&
       cnv$cov_ratio[i] <  cov.thresh.hi)
    {
      if(DEBUG){
        print(cnv[i,])
        print(as.character(cnv$strain[i]))
      }
      # following ASSUMES no CNVnator call crosses a chromosome bdry, & that
      # t.len ends at chr end (typically chr1 or chr24)
      if(cnv$iEnd[i] <= t.len){
        cnv.deletions[[as.character(cnv$strain[i])]][cnv$iStart[i]:cnv$iEnd[i]] <- TRUE
      }
    }
  }
  return(cnv.deletions)
}

# sanity check:
cnv.dels.38 <- get.cnv.dels(0.3, 0.8, cnv.chronly, snp.tables = NULL)
unlist(lapply(cnv.dels.38,sum)) # does it match low.length.38 in tic ?

#  tp1007  tp1012  tp1013  tp1014  tp1015      IT  tp1335
# 1672500 1781500 1383600 1313700  988400  320900 1453000

# 1672500 1781500 1399400 1313700 988400 336500 1453000 <== low.length.38 from tic (circa page 8)
# 1672500 1781500 1399400 1313700 988400 336500 1453000 <== low.length.38 from tic (pg9, 6/28/17)
rm(cnv.dels.38)
```

Slight discrepancy in H-clade that I should hunt down, but basically OK. (hmm; maybe untrunc tbls.)

```
# the ones we want for the current analysis:
hemi.masks <- get.cnv.dels(0.3, 0.8, cnv.chronly, snp.tables=snp.tables)

rbind(
  homnr        = unlist(lapply(hnr,sum)),
  hemi         = unlist(lapply(hemi.masks, sum)),
  homnr.unhemi = unlist(lapply(list(1,2,3,4,5,6,7), function(i){sum(hnr[[i]] & !hemi.masks[[i]])}))
)

#                1007  1012  1013  1014  1015  3367  1335
# homnr            65    65  6281    23    62  7071    40
# hemi          11761 16653 24337 11603 19824 49254 15367
# homnr.unhemi    65    64  6193    23    52  6962    40
```

```
# based on the thought that hnr in 1335 may reflect errors in the ref seq,
# are they shared with others?
unlist(lapply(hnr, function(x){sum(x & hnr[[7]])}))                  # hnr shared with 1335

# 1007 1012 1013 1014 1015 3367 1335
#   11   15   18   11   14   19   40

# answer: around 300 in each strain, of 558 in NY, genomewide,
# so that seems like a plausibly important factor.

hnr.lclade <- hnr[[1]] | hnr[[2]] | hnr[[4]] | hnr[[5]] | hnr[[7]]   # union over L-clade
sum(hnr.lclade)                                                      # count all in L-clade

# [1] 152

sum(hnr[[3]] | hnr[[6]])                                             # present in H-clade

# [1] 10348

sum(hnr[[3]] & hnr[[6]])                                             # shared in H-clade

# [1] 3004
# look at a few in L-clade
w.hnr.l <- which(hnr.lclade)
seecounts(w.hnr.l[1:10],snp.tables=snp.tables)

#     chr    pos  Ref Strain    A    G    C    T SNP  exon indel nrf rat
# 1  Chr1   5397    C
# 2                     1007    0    0   24   27   1  TRUE FALSE
# 3                     1012    0    0   34   40   1  TRUE FALSE
# 4                     1013    0    0   12   42   0  TRUE FALSE
# 5                     1014    1    0   30   28   1  TRUE FALSE
# 6                     1015    0    0   33   35   1  TRUE FALSE
# 7                     3367    0    0   20   38   1  TRUE FALSE
# 8                     1335    0    0   29   98   1  TRUE FALSE
# 9  Chr1  20071    T
# 10                    1007   22    0    0   15   1 FALSE FALSE
# 11                    1012  109    0    0   41   1 FALSE FALSE
# 12                    1013   28    0    0   33   1 FALSE FALSE
# 13                    1014   76    0    0   29   1 FALSE FALSE
# 14                    1015  130    0    0   28   1 FALSE FALSE
# 15                    3367   27    0    0   28   0 FALSE FALSE
# 16                    1335   95    0    0   57   0 FALSE FALSE
# 17 Chr1  25350    G
# 18                    1007  104   31    0    0   1 FALSE FALSE
# 19                    1012  171   53    0    0   1 FALSE FALSE
# 20                    1013  209   87    1    0   0 FALSE FALSE
# 21                    1014   19   32    0    0   0 FALSE FALSE
# 22                    1015   91   44    0    0   1 FALSE FALSE
# 23                    3367  397   94    0    0   0 FALSE FALSE
# 24                    1335   80   64    0    0   0 FALSE FALSE
# 25 Chr1  26205    T
# 26                    1007   50    0    0   20   1 FALSE FALSE
# 27                    1012  104    0    0   33   1 FALSE FALSE
# 28                    1013  224    0    0   69   1 FALSE FALSE
# 29                    1014   23    0    1   16   1 FALSE FALSE
# 30                    1015   88    0    0   33   1 FALSE FALSE
# 31                    3367  143    0    0   41   1 FALSE FALSE
# 32                    1335  196    0    0   67   1 FALSE FALSE
# 33 Chr1  90942    C
# 34                    1007    0    0   15    0   0 FALSE FALSE
# 35                    1012    0    0   33    0   0 FALSE FALSE
# 36                    1013    0    0   46    0   0 FALSE FALSE
# 37                    1014    0    0   16    0   0 FALSE FALSE
# 38                    1015    0    0    7   25   1 FALSE FALSE
# 39                    3367    0    0   56    0   0 FALSE FALSE
# 40                    1335    0    0   70    0   0 FALSE FALSE
# 41 Chr1 149447    T
# 42                    1007    0    0    0    1   0 FALSE FALSE
# 43                    1012    0    1    0    0   0 FALSE FALSE
# 44                    1013    0    0    0    1   0 FALSE FALSE
# 45                    1014    0    0    0    8   0 FALSE FALSE
# 46                    1015    0    0    0    2   0 FALSE FALSE
# 47                    3367    0    0    0    1   0 FALSE FALSE
# 48                    1335    0    1    0    1   0 FALSE FALSE
# 49 Chr1 149457 <NA>
# 50                    1007 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 51                    1012 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 52                    1013 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 53                    1014 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 54                    1015 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 55                    3367 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 56                    1335 <NA> <NA> <NA> <NA>   0 FALSE FALSE
# 57 Chr1 156248    A
# 58                    1007    2    0   39    0   0 FALSE FALSE
# 59                    1012    7    0   67    0   0 FALSE FALSE
# 60                    1013    5    0   53    0   0 FALSE FALSE
# 61                    1014   11    0   13    0   1 FALSE FALSE
# 62                    1015    6    0   44    0   0 FALSE FALSE
# 63                    3367    9    0   66    0   0 FALSE FALSE
# 64                    1335   62    0   31    0    1 FALSE FALSE
# 65 Chr1 176517    C
# 66                    1007    0    0    0    1    0  TRUE FALSE
# 67                    1012    0    0    2    0    0  TRUE FALSE
# 68                    1013    0    0    4    0    0  TRUE FALSE
# 69                    1014    0    0    6    0    0  TRUE FALSE
# 70                    1015    0    0    0    0    0  TRUE FALSE
# 71                    3367    0    0    4    0    0  TRUE FALSE
# 72                    1335    0    0   11    0    0  TRUE FALSE
# 73 Chr1 193761    C
# 74                    1007    0    0   20   14    1 FALSE FALSE
# 75                    1012    0    0   19   31    1 FALSE FALSE
# 76                    1013    0    1    4    6    1 FALSE FALSE
# 77                    1014    0    0    9    4    1 FALSE FALSE
# 78                    1015    0    0   28   39    1 FALSE FALSE
# 79                    3367    0    0    7   11    0 FALSE FALSE
# 80                    1335    0    0   10   43    1 FALSE FALSE

# one of those is a little weird:
xx<-snp.tables[[1]][149457,]
for (i in 2:7){xx <- rbind(xx,snp.tables[[i]][149457,])}
row.names(xx)<-names(snp.tables)
# My guess is that Chr/Pos/Ref are left as NA if coverage is zero.
xx

#       chr    pos snp  Chr    Pos  Ref Cov a g c t n .match  exon indel
# 1007 Chr1 149457   0 <NA>      NA <NA>   0 0 0 0 0 0      0 FALSE FALSE
# 1012 Chr1 149457   0 <NA>      NA <NA>   0 0 0 0 0 0      0 FALSE FALSE
# 1013 Chr1 149457   0 <NA>      NA <NA>   0 0 0 0 0 0      0 FALSE FALSE
# 1014 Chr1 149457   0 Chr1 149457    G   1 0 0 0 0 0      1 FALSE FALSE
# 1015 Chr1 149457   0 <NA>      NA <NA>   0 0 0 0 0 0      0 FALSE FALSE
# 3367 Chr1 149457   0 <NA>      NA <NA>   0 0 0 0 0 0      0 FALSE FALSE
# 1335 Chr1 149457   0 Chr1 149457    G   1 0 0 1 0 0      0 FALSE FALSE
```

5 Table 1 stats

Here is a brief summary of per-strain SNP counts, pairwise overlaps, and other conveniently available stats, such as those shown in Table 1 of the paper.

```
snp.counts    <- matrix(NA,7,4)
snp.pctofny   <- matrix(NA,7,4)
snp.pctofself <- matrix(NA,7,4)
snp.inter  <- matrix(NA,7,7)
snp.union  <- matrix(NA,7,7)
rownames(snp.counts)    <- names(snp.tables)
rownames(snp.pctofny)   <- names(snp.tables)
rownames(snp.pctofself) <- names(snp.tables)
rownames(snp.inter)  <- names(snp.tables)
colnames(snp.inter)  <- names(snp.tables)
rownames(snp.union)  <- names(snp.tables)
colnames(snp.union)  <- names(snp.tables)
for(stringency in 1:4){
  cat('\nStringency', stringency, ifelse(stringency==4,'(i.e. raw SAMTools SNP calls)',''),
      ':\n--------------\n')
  for(i in 1:7){
    f.snps.i <- refined.snps$Code$get.snps(i, stringency)
    snp.counts[i,stringency] <- sum(f.snps.i)
    for(j in i:7){
      f.snps.j <- refined.snps$Code$get.snps(j, stringency)
      snp.inter[i,j] <- sum(f.snps.i & f.snps.j)
      snp.union[i,j] <- sum(f.snps.i | f.snps.j)
    }
  }
  snp.pctofny  [,stringency] <- snp.inter[,7]/snp.counts[7,stringency]
  snp.pctofself[,stringency] <- snp.inter[,7]/snp.counts[ ,stringency]
  cat('Union Counts:\n');                  print(snp.union)
  cat('Intersect Counts:\n');              print(snp.inter)
  cat('Intersect as percent of union:\n'); print(snp.inter/snp.union*100,digits=3)
}

#
# Stringency 1  :
# --------------
# Union Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 17466 18310 30476 18476 18305 30017 18612
# 1012    NA 17836 30594 18774 18574 30130 18883
# 1013    NA    NA 24999 30148 30643 31962 30648
# 1014    NA    NA    NA 16338 18733 29644 18990
# 1015    NA    NA    NA    NA 17814 30160 18852
# 3367    NA    NA    NA    NA    NA 24021 30175
# 1335    NA    NA    NA    NA    NA    NA 18039
# Intersect Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 17466 16992 11989 15328 16975 11470 16893
# 1012    NA 17836 12241 15400 17076 11727 16992
# 1013    NA    NA 24999 11189 12170 17058 12390
# 1014    NA    NA    NA 16338 15419 10715 15387
# 1015    NA    NA    NA    NA 17814 11675 17001
# 3367    NA    NA    NA    NA    NA 24021 11885
# 1335    NA    NA    NA    NA    NA    NA 18039
# Intersect as percent of union:
#      1007  1012  1013  1014  1015  3367  1335
# 1007  100  92.8  39.3  83.0  92.7  38.2  90.8
# 1012   NA 100.0  40.0  82.0  91.9  38.9  90.0
# 1013   NA    NA 100.0  37.1  39.7  53.4  40.4
# 1014   NA    NA    NA 100.0  82.3  36.1  81.0
# 1015   NA    NA    NA    NA 100.0  38.7  90.2
# 3367   NA    NA    NA    NA    NA 100.0  39.4
# 1335   NA    NA    NA    NA    NA    NA 100.0
#
# Stringency 2  :
# --------------
# Union Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 18089 18521 37510 18365 18729 36679 18497
# 1012    NA 18342 37599 18538 18771 36774 18566
# 1013    NA    NA 30797 36124 37777 41302 37426
# 1014    NA    NA    NA 14194 18733 35137 18202
# 1015    NA    NA    NA    NA 18551 36940 18754
# 3367    NA    NA    NA    NA    NA 29496 36558
# 1335    NA    NA    NA    NA    NA    NA 17819
# Intersect Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 18089 17910 11376 13918 17911 10906 17411
# 1012    NA 18342 11540 13998 18122 11064 17595
# 1013    NA    NA 30797  8867 11571 18991 11190
# 1014    NA    NA    NA 14194 14012  8553 13811
# 1015    NA    NA    NA    NA 18551 11107 17616
# 3367    NA    NA    NA    NA    NA 29496 10757
# 1335    NA    NA    NA    NA    NA    NA 17819
# Intersect as percent of union:
#      1007  1012  1013  1014  1015  3367  1335
# 1007  100  96.7  30.3  75.8  95.6  29.7  94.1
# 1012   NA 100.0  30.7  75.5  96.5  30.1  94.8
# 1013   NA    NA 100.0  24.5  30.6  46.0  29.9
# 1014   NA    NA    NA 100.0  74.8  24.3  75.9
# 1015   NA    NA    NA    NA 100.0  30.1  93.9
# 3367   NA    NA    NA    NA    NA 100.0  29.4
# 1335   NA    NA    NA    NA    NA    NA 100.0
#
# Stringency 3  :
```

```
# --------------
# Union Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 17107 18179 36938 17408 18347 36105 17958
# 1012    NA 17881 37265 18044 18505 36432 18258
# 1013    NA    NA 30364 34100 37453 41272 36889
# 1014    NA    NA    NA  9791 18257 33046 17322
# 1015    NA    NA    NA    NA 18095 36626 18420
# 3367    NA    NA    NA    NA    NA 29135 36011
# 1335    NA    NA    NA    NA    NA    NA 16984
# Intersect Counts:
#       1007  1012  1013 1014  1015  3367  1335
# 1007 17107 16809 10533 9490 16855 10137 16133
# 1012    NA 17881 10980 9628 17471 10584 16607
# 1013    NA    NA 30364 6055 11006 18227 10459
# 1014    NA    NA    NA 9791  9629  5880  9453
# 1015    NA    NA    NA   NA 18095 10604 16659
# 3367    NA    NA    NA   NA    NA 29135 10108
# 1335    NA    NA    NA   NA    NA    NA 16984
# Intersect as percent of union:
#      1007  1012  1013  1014  1015  3367  1335
# 1007  100  92.5  28.5  54.5  91.9  28.1  89.8
# 1012   NA 100.0  29.5  53.4  94.4  29.1  91.0
# 1013   NA    NA 100.0  17.8  29.4  44.2  28.4
# 1014   NA    NA    NA 100.0  52.7  17.8  54.6
# 1015   NA    NA    NA    NA 100.0  29.0  90.4
# 3367   NA    NA    NA    NA    NA 100.0  28.1
# 1335   NA    NA    NA    NA    NA    NA 100.0
#
# Stringency 4 (i.e. raw SAMTools SNP calls) :
# --------------
# Union Counts:
#       1007  1012  1013  1014  1015  3367  1335
# 1007 16530 17707 35005 16864 17989 34289 17382
# 1012    NA 17019 35294 17276 18074 34563 17577
# 1013    NA    NA 25412 30445 35599 39448 34479
# 1014    NA    NA    NA  8331 17634 29704 16078
# 1015    NA    NA    NA    NA 17397 34876 17881
# 3367    NA    NA    NA    NA    NA 24613 33699
# 1335    NA    NA    NA    NA    NA    NA 15582
# Intersect Counts:
#       1007  1012  1013 1014  1015  3367  1335
# 1007 16530 15842  6937 7997 15938  6854 14730
# 1012    NA 17019  7137 8074 16342  7069 15024
# 1013    NA    NA 25412 3298  7210 10577  6515
# 1014    NA    NA    NA 8331  8094  3240  7835
# 1015    NA    NA    NA   NA 17397  7134 15098
# 3367    NA    NA    NA   NA    NA 24613  6496
# 1335    NA    NA    NA   NA    NA    NA 15582
# Intersect as percent of union:
#      1007  1012  1013  1014  1015  3367  1335
# 1007  100  89.5  19.8  47.4  88.6  20.0  84.7
# 1012   NA 100.0  20.2  46.7  90.4  20.5  85.5
# 1013   NA    NA 100.0  10.8  20.3  26.8  18.9
# 1014   NA    NA    NA 100.0  45.9  10.9  48.7
# 1015   NA    NA    NA    NA 100.0  20.5  84.4
# 3367   NA    NA    NA    NA    NA 100.0  19.3
# 1335   NA    NA    NA    NA    NA    NA 100.0

vs.stringency <- cbind(snp.counts, matrix(NA,7,1), round(snp.counts[,1:3]/snp.counts[,4]*100,1))
colnames(vs.stringency) <- c('[[1]]', '[[2]]', '[[3]]', '[[4]]', '----', '[[1]]%', '[[2]]%', '[[3]]%')

# SNPs vs filtering stringency (raw counts and as % of [[4]]).  Medium filter
# adds 10-20% in most cases.  Big exception is Gyre, where low coverage,
# high err rate and SAMTools conservatism seemed to seriously undercall:
print(vs.stringency)

#      [[1]] [[2]] [[3]] [[4]] ---- [[1]]% [[2]]% [[3]]%
# 1007 17466 18089 17107 16530   NA  105.7  109.4  103.5
# 1012 17836 18342 17881 17019   NA  104.8  107.8  105.1
# 1013 24999 30797 30364 25412   NA   98.4  121.2  119.5
# 1014 16338 14194  9791  8331   NA  196.1  170.4  117.5
# 1015 17814 18551 18095 17397   NA  102.4  106.6  104.0
# 3367 24021 29496 29135 24613   NA   97.6  119.8  118.4
# 1335 18039 17819 16984 15582   NA  115.8  114.4  109.0

# Intersect NY as % of self (vs stringency):
print(snp.pctofself*100, digits=3)

#        [,1]  [,2]  [,3]  [,4]
# 1007  96.7  96.3  94.3  89.1
# 1012  95.3  95.9  92.9  88.3
# 1013  49.6  36.3  34.4  25.6
# 1014  94.2  97.3  96.5  94.0
# 1015  95.4  95.0  92.1  86.8
# 3367  49.5  36.5  34.7  26.4
# 1335 100.0 100.0 100.0 100.0

# Intersect NY as % of NY (vs stringency):
print(snp.pctofny*100, digits=3)

#        [,1]  [,2]  [,3]  [,4]
# 1007  93.6  97.7  95.0  94.5
# 1012  94.2  98.7  97.8  96.4
# 1013  68.7  62.8  61.6  41.8
# 1014  85.3  77.5  55.7  50.3
# 1015  94.2  98.9  98.1  96.9
# 3367  65.9  60.4  59.5  41.7
# 1335 100.0 100.0 100.0 100.0
```

Quick look at coverage. Are there any NA?:

```
nacount <- NULL
for(i in 1:4){
  if(!is.null(tset[[i]])){
    nacount <- rbind(nacount,
                     unlist(lapply(tset[[i]], function(x){sum(is.na(x$Cov))}))
                     )
    rownames(nacount)[nrow(nacount)] <- names(tset)[i]
  }
}
nacount

#                             1007 1012 1013 1014 1015 3367 1335
# snp.tables.chr1.unfiltered    0    0    0    0    0    0    0
# snp.tables.chr1.qfiltered     0    0    0    0    0    0    0
```

Seemingly no. What's average in unq- vs q-filtered:

```
snp.tables.unqfil <- tset.picker(c(1,2), table.set = tset)
snp.tables.qfil   <- tset.picker(c(3,4), table.set = tset)
cov.unqfil <- unlist(lapply(snp.tables.unqfil, function(x){mean(x$Cov)}))
cov.qfil   <- unlist(lapply(snp.tables.qfil,   function(x){mean(x$Cov,na.rm=T)}))
cov.both <- rbind(cov.unqfil,cov.qfil,cov.qfil/cov.unqfil)
i <- 1
if(!is.null(snp.tables.unqfil)){
  rownames(cov.both)[i] <- which.snp.tables(snp.tables.unqfil)
  i <- i+1
}
if(!is.null(snp.tables.qfil)){
  rownames(cov.both)[i] <- which.snp.tables(snp.tables.qfil)
  i <- i+1
}
if(i==3){
  rownames(cov.both)[i] <- 'Ratio'
}
cat('Mean Coverage:\n'); cov.both

# Mean Coverage:
#                       1007       1012       1013       1014       1015       3367        1335
# Chr1-unfiltered 36.2816326 68.2005811 66.6908911 31.2663216 59.4704151 62.3834535 103.9124774
# Chr1-qfiltered  27.5849516 49.2557296 43.2293645 12.4319866 46.9874722 43.4403699  78.7789843
# Ratio            0.7603007  0.7222186  0.6482049  0.3976159  0.7900983  0.6963444   0.7581282
```

5.1 Table 1 Data

Throw together the conveniently-available Table 1 data, *in Table 1 row order*:

```
# if coverage unavailable, build NA vector
if(!is.null(cov.unqfil)){cov.unqfilv <- cov.unqfil} else {cov.unqfilv <- rep(NA,times=7)}
if(!is.null(cov.qfil  )){cov.qfilv   <- cov.qfil  } else {cov.qfilv   <- rep(NA,times=7)}
t1data.df <- data.frame(
  id        = st.locs(1:7, id=T, loc=F, date=F),
  loc       = st.locs(1:7, id=F, loc=T, date=F),
  date      = st.locs(1:7, id=F, loc=F, date=T),
  cov.unq   = cov.unqfilv,
  cov.q     = cov.qfilv,
  SNPs.4    = snp.counts[,4],
  SNPs.2    = snp.counts[,2],
  olap.ny.4 = snp.pctofny[,4]*100,
  olap.ny.2 = snp.pctofny[,2]*100
)
t1row.order <- c(7,1,2,5,3,6,4)
print(t1data.df[t1row.order,],digits=3)

#            id             loc date cov.unq cov.q SNPs.4 SNPs.2 olap.ny.4 olap.ny.2
# 1335 CCMP1335        New York 1958   103.9  78.8  15582  17819     100.0     100.0
# 1007 CCMP1007        Virginia 1964    36.3  27.6  16530  18089      94.5      97.7
# 1012 CCMP1012    W. Australia 1965    68.2  49.3  17019  18342      96.4      98.7
# 1015 CCMP1015     Puget Sound 1985    59.5  47.0  17397  18551      96.9      98.9
# 1013 CCMP1013           Wales 1973    66.7  43.2  25412  30797      41.8      62.8
# 3367 CCMP3367           Italy 2007    62.4  43.4  24613  29496      41.7      60.4
# 1014 CCMP1014 N. Pacific Gyre 1971    31.3  12.4   8331  14194      50.3      77.5
```

6 Shared-SNPs P-Value

Text of the main paper quotes a "p-value" for the observed degree of SNP sharing in L-clade (and/or L-clade excluding Gyre) under a null model that these isolates were sampled from a population globally in Hardy-Weinberg equilibrium. Details of this analysis are as follows.

6.1 SNP Concordance

Arbitrarily pick one isolate, say, A, as the "template". Arbitrarily pick a heterozygous (aka "SNP") position in A. Let p_1, and $q_1 = 1 - p_1$ be the frequencies in the overall population of the two nucleotides observed at that position in A. (Positions having 3 or 4 nucleotide variants segregating in the population are assumed to be negligibly rare.) Under the HWE null model, a second isolate B will also be heterozygous at the same position with probability $2p_1q_1 \leq 1/2$. Similarly, this position will be heterozygous in a third isolate C with the same probability, *independently*, and so on for isolates D and E. Overall, (assuming HWE) the probability that a heterozygous position in A is simultaneously heterozygous in the other 4 isolates is at most $1/2^4 = 1/16$. Continuing, suppose we pick a second heterozygous position in A, *on a different chromosome* with allele frequencies $p_2, q_2 = 1 - p_2$, say. Again assuming HWE, this position will be a SNP in all of B, C, D and E with probability $(2p_2q_2)^4 \leq 1/16$, and this is independent of the first position, since segregation on different chromosomes is unlinked. Repeat this at 24 heterozygous positions in A, one per chromosome. Then, the number of five-way concordant positions observed should be dominated by the number observed when sampling from a binomial distribution with parameters $n = 24$ and $p = 1/16$, i.e., we expect at most $1/16 = 6.25\%$ of positions to agree, or at most $24/16 = 1.5$ five-way concordant positions in total. In sharp contrast, choosing CCMP 1014 (North Pacific Gyre) as the template, we see many more five-way concordant positions than predicted under these assumptions:

```
gyre.count <- sum(snp.tables[[4]]$snp)
# NOTE: what we now calle "refined" SNPs were once called "filtered" SNPs and I have NOT tried
# to update variable names and annotation in the code below to reflect the terminology change...
# 'unfil.' => unfiltered for consistency; see below.
unfil.fiveway.count   <- sum( snp.tables[[4]]$snp * i4.snps)
unfil.fiveway.percent <- unfil.fiveway.count / gyre.count * 100
unfil.p.value <- pbinom(floor(unfil.fiveway.count/gyre.count*24)-1, 24, 1/16, lower.tail = FALSE)
consistency.comparison <-
  data.frame(
    fiveway.count   = unfil.fiveway.count,
    fiveway.percent = unfil.fiveway.percent,
    p.value         = unfil.p.value
  )
consistency.comparison

#   fiveway.count fiveway.percent      p.value
# 1          7628        91.56164 8.700771e-23
```

Namely, 8331 positions are called as SNPs in CCMP1014, of which 7628 or 91.5616373% are also called as SNPs in *all four* other L-clade isolates. 91.5616373% of 24 is 21.9747929, and the probability of seeing 21 or more "Heads" in 24 flips of a biased coin with $P(\text{Heads}) \leq 1/16$, i.e., our p-value under the HWE null hypothesis, is at most: 8.700771×10^{-23} based on this simple binomial model. This is obviously strong evidence against the null hypothesis.

This analysis is potentially overly-simplistic in four respects, addressed below.

1. "$2pq \leq 1/2$" is conservative. Neutral theory predicts that most variant nucleotides are rare in the population, so $2pq \ll 1/2$ is to be expected. This should make our quoted p-value very conservative.

2. Effect of Erroneous SNP calls. We base our analysis on *predicted* (by SAMTOOLS) heterozygous positions, not absolute-truth, which may affect our conclusions. However,

   - False negatives in A are irrelevant, since we never examine those positions. (This is the motivation for using CCMP1014 as the template; it has the lowest predicted SNP rate, likely due to a high false negative rate in that sequencing run. As noted elsewhere, it had the lowest coverage and lowest sequence quality of the 7 isolates, both of which impare SNP calling.)

   - False negatives in $BCDE$ make such positions appear *non*-concordant. For our purpose, this makes our statistic more conservative since it can only deflate a statistic that we argue is nevertheless unexpectedly large.

   - False positive calls in A are conservatively treated, as well: barring simultaneous false-positive calls in all of $BCDE$, such a position will appear non-concordant, again deflating the statistic. The *false* positive rates in B, C, D and E are unknown, but cannot exceed SAMTOOLS *total* positive rate, which is below 1% in all 7 isolates, suggesting a simultaneous $BCDE$ false positive rate $< 10^{-8}$, which will have a negligible effect.

   - A potentially more serious issue is a true positive in A aligned to false positives in BCD and/or E. (I.e., a position that is polymorphic in the population and heterozygous in A, under the HWE null model is likely to be homozygous for one of the two alleles in one or more of $BCDE$; false positive SNP calls in all of those isolates would make the site appear concordant, i.e., provide evidence against the null model.) However, (a) my impression is that SAMTOOLS is more prone to false negative calls than to false positive calls (see Section 4), and (b) we would need a high rate of false positives to turn a truely heterozygous but non-concordant A call into a false "concordant" call—I'd expect at most half (especially given point 1 above) of $BCDE$ to be heterozygous, but all would need to be falsely declared heterozygous. Such a high false positive rate on $BCDE$ seems unlikely (see previous bullet), and would likely be counterbalanced by a similarly increased rate of false positives on A, which, as noted, tend to deflate our statistic (previous bullet again).

- Systematic errors. If there were, say, a sequence-context-dependent bias in the DNA sequencing, mapping and/or SNP-calling that tended to suggest (or hide) a SNP at some position, we're going to systematically over- (or under-) estimate concordant SNPs across isolates. The discordance of called SNPs between the L- and H-clades and within the H-clade suggests that this is not a major problem, but it is worth noting as a possibility.

3. Discordant nucleotides at "concordant" SNP positions. A "shared" SNP at a given position might be, say, G/C in one isolate vs T/C in another, reflecting an unexpected tri-allelic position in the population or a technical sequencing error. It is inappropriate to count such a "shared" SNP position as evidence against the null hypothesis, since it isn't clear that it is truely shared. Instead, I will identify such inconsistent positions, based on the "stringency [[2]]" criteria established above, and treat each as non-concordant. I.e., a position will be considered to be a "5-way concordant SNP" if and only if it was called as a SNP by SAMTOOLS (independently) in all 5 L-clade isolates, *and* shows the same dominant non-reference nucleotide in all 5, according to criteria [[2]] above. As it turns out, this correction has a very minor effect on the resulting p-value:

```
# 'unfil.' => Ignoring "consistency";  'fil.' => Filtering for "consistency":
fil.fiveway.count   <- sum((snp.tables[[4]]$snp * i4.snps)[union.snps == 1] & consistent[[2]])
fil.fiveway.percent <- fil.fiveway.count / gyre.count * 100
fil.p.value <- pbinom(floor(fil.fiveway.count/gyre.count*24)-1, 24, 1/16, lower.tail = FALSE)
# append new stats to previous table for easy comparison
consistency.comparison <-
  rbind(consistency.comparison,
        data.frame(
          fiveway.count   = fil.fiveway.count,
          fiveway.percent = fil.fiveway.percent,
          p.value         = fil.p.value
        )
  )
rownames(consistency.comparison) <- c('unfiltered', 'consistency.filtered')
consistency.comparison


#                        fiveway.count fiveway.percent      p.value
# unfiltered                      7628        91.56164 8.700771e-23
# consistency.filtered            7534        90.43332 8.700771e-23
```

In particular, it removes 1.1% of five-way consistent positions (only 94 of 7628 positions), and still shows a highly significant p-value.

4. "$P(E[X]) \neq E[P(X)]$". I'm expressing this poorly, but finding the p-value based on the *expected* number of concordant positions is somewhat non-standard. A more typical set-up would use the *actual* value of some statistic, then calculate the probability of observing a value that extreme (or more extreme) under the null model. The fundamental problem is that we have thousands of SNPs, but I don't see an easy way to use more than 24 of them at a time, because potential genetic linkage seemingly destroys statistical independence, which is key to most simple analyses. A somewhat more formal, but still non-standard, approach is the following. Suppose we randomly sample one SNP per chromosome and count the number X of them that are 5-way concordant. What I outlined above calculated the p-value based on $E[X]$, the expected value of X, i.e., $P(E[X])$. Alternatively, we can calculate $E[P(X)]$, the expected p-value. (They are not the same.) In effect, this averages the p-values that would be seen over many different randomly-sampled sets of 24 SNPs. This is not difficult to calculate. First, the probability that we would observe $0 \leq i \leq 24$ concordant positions in a sample of 24, given that 90.43% of positions are concordant follows this binomial distribution:

```
x.equals.i.distribution <- dbinom(0:24, 24, fil.fiveway.percent/100)
print(x.equals.i.distribution, digits=3)


#  [1] 3.45e-25 7.84e-23 8.52e-21 5.90e-19 2.93e-17 1.11e-15 3.32e-14 8.06e-13 1.62e-11 2.72e-10
# [11] 3.86e-09 4.64e-08 4.75e-07 4.15e-06 3.08e-05 1.94e-04 1.03e-03 4.59e-03 1.69e-02 5.04e-02
# [21] 1.19e-01 2.14e-01 2.76e-01 2.27e-01 8.95e-02
```

Second, the p-value corresponding to $0 \leq i \leq 24$ observed concordant positions also follows a different binomial distribution:

```
p.val.of.x.equals.i <- c(1, pbinom(0:23, 24, 1/16, lower.tail = F))
print(p.val.of.x.equals.i, digits=3)

#  [1] 1.00e+00 7.88e-01 4.48e-01 1.87e-01 5.95e-02 1.49e-02 3.01e-03 4.99e-04 6.90e-05 8.02e-06
# [11] 7.89e-07 6.60e-08 4.72e-09 2.87e-10 1.49e-11 6.59e-13 2.46e-14 7.66e-16 1.98e-17 4.14e-19
# [21] 6.88e-21 8.70e-23 7.88e-25 4.56e-27 1.26e-29
```

Finally, the expected (or “average”) p-value is just the weighted average of the latter values, weighted by the former:

```
e.of.p.of.x <- sum(x.equals.i.distribution * p.val.of.x.equals.i)
e.of.p.of.x

# [1] 1.398085e-14
```

This is still highly significant, but weaker than the $P(E[X])$ analysis, basically because $X < E[X]$ has a fair probability of occurring, and the corresponding p-value $P(X)$ rises rapidly as X declines.

Another way to look at the numbers:

```
pvdf <- data.frame(x.density=x.equals.i.distribution,
                   x.cdf=cumsum(x.equals.i.distribution),
                   pval.of.x=p.val.of.x.equals.i)
print(pvdf, digits=4)

#    x.density     x.cdf pval.of.x
# 1  3.454e-25 3.454e-25 1.000e+00
# 2  7.835e-23 7.870e-23 7.875e-01
# 3  8.518e-21 8.596e-21 4.476e-01
# 4  5.904e-19 5.990e-19 1.869e-01
# 5  2.930e-17 2.990e-17 5.950e-02
# 6  1.108e-15 1.138e-15 1.490e-02
# 7  3.317e-14 3.430e-14 3.010e-03
# 8  8.062e-13 8.405e-13 4.994e-04
# 9  1.619e-11 1.704e-11 6.899e-05
# 10 2.722e-10 2.892e-10 8.015e-06
# 11 3.859e-09 4.148e-09 7.887e-07
# 12 4.643e-08 5.058e-08 6.603e-08
# 13 4.755e-07 5.260e-07 4.716e-09
# 14 4.149e-06 4.675e-06 2.875e-10
# 15 3.081e-05 3.549e-05 1.493e-11
# 16 1.942e-04 2.297e-04 6.590e-13
# 17 1.033e-03 1.262e-03 2.456e-14
# 18 4.593e-03 5.855e-03 7.662e-16
# 19 1.689e-02 2.274e-02 1.977e-17
# 20 5.041e-02 7.315e-02 4.143e-19
# 21 1.191e-01 1.923e-01 6.877e-21
# 22 2.145e-01 4.067e-01 8.701e-23
# 23 2.765e-01 6.832e-01 7.884e-25
# 24 2.273e-01 9.105e-01 4.556e-27
# 25 8.951e-02 1.000e+00 1.262e-29
```

E.g., row 9 in that table says that the concordance rate (90%) is so high that a sample of 24 SNPs will almost always have 9 or more five-way concordant positions (probability of fewer is only 1.704e-11), while under the null model, seeing 9 or more is very unlikely (probability at most 6.899e-05). ***AM I OFF-BY-ONE INTERPRETING ROW 9 HERE??***

6.2 Notes

In earlier drafts, an analog of the above analysis was based on the concordance of *refined* SNPs. This now seems to me to be questionable, since the “refined” SNP calling makes SNPs called across L-clade non-independent. OTOH, the above analysis seems valid: SAMTOOLS was run on each isolate independently, and likewise “criterion [[2]]” is evaluated independently in each strain, and is being used here solely to remove SNP predictions, not to add them. “Systematic errors” as outlined above remain a potential problem, but again discordance with/within H-clade suggests that this is of limited concern.

For completeness, I did a similar analysis including a sample of H-clade comparisons: Gyre vs Italy, NY vs Italy, NY vs Italy+Wales, and of Italy vs Wales. As expected, none of these show a statistically significant p-value, although the $\approx 40\%$ concordance in the 2-way comparisons, while $< 1/2$ as predicted, is a bit higher than I expected based on “neutral theory implies many rare variants.” (I did not bother to include “criterion[[2]] filtering” in these calculations.)

```
# 'gi.twoway' => gyre vs italy 2-way concordance;
# 'ni.twoway' => new york vs italy 2-way concordance;
# not bothering with criterion[[2]] filtering
gi.twoway.count   <- sum(snp.tables[[4]]$snp * snp.tables[[6]]$snp)
gi.twoway.percent <- gi.twoway.count / gyre.count * 100
gi.p.value <- pbinom(floor(gi.twoway.count/gyre.count*24)-1, 24, 1/2, lower.tail = FALSE)
ny.count <- sum(snp.tables[[7]]$snp)
ni.twoway.count   <- sum(snp.tables[[7]]$snp * snp.tables[[6]]$snp)
ni.twoway.percent <- ni.twoway.count / ny.count * 100
ni.p.value <- pbinom(floor(ni.twoway.count/ny.count*24)-1, 24, 1/2, lower.tail = FALSE)
niw.threeway.count   <- sum(snp.tables[[7]]$snp * snp.tables[[6]]$snp * snp.tables[[3]]$snp)
niw.threeway.percent <- niw.threeway.count / ny.count * 100
niw.p.value <- pbinom(floor(niw.threeway.count/ny.count*24)-1, 24, 1/4, lower.tail = FALSE)
it.count <- sum(snp.tables[[6]]$snp)
iw.twoway.count   <- sum(snp.tables[[6]]$snp * snp.tables[[3]]$snp)
iw.twoway.percent <- iw.twoway.count / it.count * 100
iw.p.value <- pbinom(floor(iw.twoway.count/it.count*24)-1, 24, 1/2, lower.tail = FALSE)
consistency.comparison <-
  rbind(consistency.comparison,
        data.frame(
          fiveway.count   = c(gi.twoway.count,   ni.twoway.count,   niw.threeway.count,   iw.twoway.count),
          fiveway.percent = c(gi.twoway.percent, ni.twoway.percent, niw.threeway.percent, iw.twoway.percent),
          p.value         = c(gi.p.value,        ni.p.value,        niw.p.value,          iw.p.value)
        )
  )
colnames(consistency.comparison)[1:2] <- c('552232way.count', '552232way.percent') # old col names misleading
rownames(consistency.comparison)[3:6] <- c('gyre.vs.italy', 'new.york.vs.italy',    # new rows
                                           'ny.vs.it.plus.wales', 'it.vs.wales')
consistency.comparison

#                      552232way.count 552232way.percent      p.value
# unfiltered                      7628          91.56164 8.700771e-23
# consistency.filtered            7534          90.43332 8.700771e-23
# gyre.vs.italy                   3240          38.89089 9.242052e-01
# new.york.vs.italy               6496          41.68913 8.462719e-01
# ny.vs.it.plus.wales             3804          24.41278 7.533516e-01
# it.vs.wales                    10577          42.97323 8.462719e-01
```

6.3 P-Value: The Bottom Line

So, what to say in the body of the paper? $E[P(X)]$ is highly significant, and conservative, but complex to explain. $P(E[X])$ is simpler to explain, but may be criticized as misleading if we aren’t very careful in that explanation. I’m slightly leaning towards the last option, but want to sleep on it and draft the key sentence or two before settling.

7 Sharing

The following analysis looks at the sharing patterns among the consistent SNPs. I assume that shared SNPs reflect shared ancestry, and that SNPs accumulate slowly over time. Then, in outline, the story is consistent with what we have seen in other analyses—there seem to be 3 groups: 1013 (Wales) in one, 3367 (Italy) in another, and the other 5 in a third, with some hints as to the order of divergence. A caveat is that in a sexual population, non-shared SNPs do not immediately imply non-shared ancestry; they may merely reflect Hardy-Weinberg capturing a homozygous state in one isolate vs the other. (Or read errors, etc.) Thus, if we are right that the H-isolates retain sex, then the large number of “private” SNPs in H may be at least partially due to HWE.

Analysis is broken into cases based on how many strains share a particular SNP.

7.1 Code

To categorize SNPs by sharing patterns, first convert the 7-way consistent sharing pattern into a 7-bit binary number, and tabulate based on that:

```r
# convert (n x 7) 0-1 matrix to n vector of 0-127
tobin <- function(x){
  bin <- integer(nrow(x)) # initialized to 0
  for(i in 1:7){
    bin <- bin*2 + as.integer(x[,i]>0)
  }
  return(bin)
}

# get full set of patterns
snp.pattern.all <- lapply(non.refs,tobin)
# prune to just the consistent ones
snp.pattern <- snp.pattern.all
for(i in 1:3){
  snp.pattern[[i]][!consistent[[i]]] <- NA
}

# analogous to built-in ``table'' but simpler.  Count entries in an integer
# vector sharing values in a (smallish) range.  Result is a 2-column matrix with
# the shared values in col 1 and count of occurrences of that value in col 2.
# Out-of-range values cause subscript error.
mytable <- function(vec, therange=range(vec,na.rm=T)){
  counts <- matrix(0,nrow=therange[2]-therange[1]+1,ncol=2,dimnames=list(NULL,c('val','count')))
  counts[1:nrow(counts),1] <- therange[1]:therange[2]
  for(i in 1:length(vec)){
    if(!is.na(vec[i])){
      counts[vec[i]-therange[1]+1,2] <- counts[vec[i]-therange[1]+1,2] + 1
    }
  }
  return(counts)
}

pattern.counts <- lapply(snp.pattern, function(x){mytable(x,c(0,127))})
```

To display the results, build a data frame whose i-th row, $0 \leq i \leq 127$ shows one of the 128 possible sharing patterns, with counts of the numbers of consistent, shared SNPs with that pattern according to criteria c1-c3.

```r
tobitvec <- function(x){
  bitvec <- integer(7)
  for(i in 7:1){
    bitvec[i] <- x %% 2
    x <- x %/% 2
  }
  return(bitvec)
}

flg <- function(x){
  return(ifelse(x==1,'X',''))
}

pat.summary <- function(listOfTbls){
  mydf <- data.frame(pat=0:127,sharedBy=NA,
                     tp1007='',tp1012='',tp1013='',tp1014='',tp1015='',tp3367='',tp1335='',
                     count1=NA,count2=NA,count3=NA,count4=NA,stringsAsFactors=F)

  for(i in 1:128){
    bvec <- tobitvec(i-1)
    mydf[i,'sharedBy']=sum(bvec)
    mydf[i,'tp1007']=flg(bvec[1])
    mydf[i,'tp1012']=flg(bvec[2])
    mydf[i,'tp1013']=flg(bvec[3])
    mydf[i,'tp1014']=flg(bvec[4])
    mydf[i,'tp1015']=flg(bvec[5])
    mydf[i,'tp3367']=flg(bvec[6])
    mydf[i,'tp1335']=flg(bvec[7])
  }

  for(i in 1:length(listOfTbls)){
    tbl <- listOfTbls[[i]]
    if(!is.null(tbl)){
      mydf[,9+i] <- tbl[,2]  ## count1/2/3/4 are columns 10/11/12/13 in mydf
      #for(j in 1:length(tbl)){
      #  k <- as.integer(rownames(tbl)[j]);
      #  mydf[k+1,9+i] <- tbl[j]  ## count1/2/3 are columns 10/11/12
      #}
    }
  }

  mydf$pat <-as.octmode(mydf$pat)  # display bit pattern in octal
  return(mydf)
}

pat.summaries <- pat.summary(pattern.counts)
```

7.2 Sanity Checks

Some sanity checking: table sums equal to number of consistent positions?

```
all(consistent.count == apply(pat.summaries[,10:13],2,sum))

# [1] TRUE
```

More sanity checking: visually inspect a pattern with small counts, specifically pattern 12, i.e., consistent SNPs shared by only strains 1014 and 1015 (2nd and 3 rows from bottom, binary code $12 = 2^3 + 2^2$). There are only 10 such positions on Chr1. Chr1 2524239 has pattern 12 under criteria c1 and c2 but not c3; Chr1 1088766 has in c2 only. Both look good. Neither position is a *called* SNP except in 1015. However, all but 1 nonreference read agree with the called SNP (the exception being one read in Wales). Both 1014 and 1015 have at least 2 non-reference reads, comprising at least 5% of coverage, and in both strains, those reads are on the same non-reference base, satisfying criterion c2. The other strains have higher coverage and/or lower non-reference counts, so they do not satisfy c2. Position 2524239 also satisfies c1, but not c3, since 2 reads out of 35 is below the 10% threshold. (It is pattern 4 inder c3, i.e., a SNP private to 1015.) Position 1088766 is also pattern 4 under c3 (2 reads out of 56 in 1335 is below both thresholds), and it is not consistent under c1, since the single A read in 1013 is discordant with the other non-reference reads.

```
unlist(lapply(snp.pattern,function(x){sum(x==12,na.rm=T)}))

# [1] 10  2  5 12

sp1 <- snp.pattern[[1]]==12
sp2 <- snp.pattern[[2]]==12
sp3 <- snp.pattern[[3]]==12
sp4 <- snp.pattern[[4]]==12
c(sum(sp1,na.rm=T), sum(sp2,na.rm=T), sum(sp3,na.rm=T), sum(sp4,na.rm=T))

# [1] 10  2  5 12

r1 <- rownames(non.refs[[1]])[which(sp1)]
r2 <- rownames(non.refs[[2]])[which(sp2)]
r3 <- rownames(non.refs[[3]])[which(sp3)]
r4 <- rownames(non.refs[[4]])[which(sp4)]

r2
```

```
# [1] "Chr1:1088766" "Chr1:2524239"

c1 <- as.integer(unlist(lapply(strsplit(r1[1:min(20,length(r1))],':',fixed=TRUE),function(x){x[2]})))
c2 <- as.integer(unlist(lapply(strsplit(r2[1:min(20,length(r2))],':',fixed=TRUE),function(x){x[2]})))
c3 <- as.integer(unlist(lapply(strsplit(r3[1:min(20,length(r3))],':',fixed=TRUE),function(x){x[2]})))
c4 <- as.integer(unlist(lapply(strsplit(r4[1:min(20,length(r4))],':',fixed=TRUE),function(x){x[2]})))

c1

#  [1]  198498  914018 1317406 1481838 1501481 1878058 2145849 2388286 2524239 2718093

c2

# [1] 1088766 2524239

c3

# [1]  371484 1210354 1886633 2264683 2898352

c4

#  [1]  518347  691730  767408 1049906 1390437 2072951 2254059 2254789 2264683 2823796 2898352
# [12] 2998868

seecounts(c2,snp.tables=snp.tables)

#     chr     pos Ref Strain A  G  C  T SNP  exon indel nrf rat
# 1  Chr1 1088766   G
# 2                       1007 0 32  1  0   0 FALSE FALSE
# 3                       1012 0 39  1  0   0 FALSE FALSE
# 4                       1013 1 74  0  0   0 FALSE FALSE
# 5                       1014 0 26  2  0   0 FALSE FALSE
# 6                       1015 0 38  9  0   1 FALSE FALSE
# 7                       3367 0 36  1  0   0 FALSE FALSE
# 8                       1335 0 54  2  0   0 FALSE FALSE
# 9  Chr1 2524239   C
# 10                      1007 0  0 37  0   0  TRUE FALSE
# 11                      1012 0  0 47  0   0  TRUE FALSE
# 12                      1013 0  0 62  0   0  TRUE FALSE
# 13                      1014 0  0 33  2   0  TRUE FALSE
# 14                      1015 0  0 11 15   1  TRUE FALSE
# 15                      3367 0  0 41  0   0  TRUE FALSE
# 16                      1335 0  0 95  0   0  TRUE FALSE
```

Position 1088766, however, in a good example of the situation that motivated this analysis—one strain has a G/C SNP and 5 of the other 6 strains have nonreference reads consistent with that SNP. Although, excluding 1015, the nonreference read counts are not high enough to justify a SNP call in any strain considered in isolation, the fact that they *consistently* agree with the 1015 SNP suggests that they are real. One alternative hypothesis is that there is some sequence-dependent bias at this locus that favors misreading a G as a C. On the other hand, one could equally well posit a shared SNP, and a locus-dependant bias that *supresses* C reads, explaining the unbalanced readout that we observe. However, it is hard to reconcile either view with the significant strain-specific patterns that we see in the shared SNPs (as seen below). I think a more likely explanation is that (a) there are some number of relatively rare SNPs present in each of the sampled populations, (b) some of these SNPs happened to be present in one or two cells of the roughly 5-10 cells that we believe constituted the founding population of the culture grown for sequencing, and (c) stochastic effects during culture growth and during sequencing may have further perturbed the apparent frequency of each variant, but the bottom line is that the above-threshold presence of consistent non-reference reads is evidence for shared SNPs at the population level (and the proportions of such reads represent estimates of the population-level frequencies of the variants, albeit a noisy estimate at any specific position).

An aside: I was curious to see whether there is any consistent pattern to positions that are called consistent SNPs in all but Italy, so I repeated the above, basically. My summary is that coverage in Italy tends to be below average in these positions, but otherwise they don't stand out. For the record:

```
abit <- snp.pattern[[2]]==125
abit[is.na(abit)]<-F
```

```
sum(abit)

# [1] 1493

rabit <- rownames(non.refs[[2]])[which(abit)]
rabits <- rabit[1:20]
cabit <- as.integer(unlist(lapply(strsplit(rabits,':',fixed=TRUE),function(x){x[2]})))
cabit

#  [1]  1244  1575  6485  7181  7220  7661  8144  8208  8518  8552  8567  8670  8685 14361 15254
# [16] 15280 16103 17587 18904 25546

seecounts(cabit,snp.tables=snp.tables)

#      chr   pos Ref Strain   A   G   C   T SNP  exon indel nrf rat
# 1   Chr1  1244   G
# 2                    1007   3  30   0   0   0  TRUE FALSE
# 3                    1012   5  54   0   0   0  TRUE FALSE
# 4                    1013  15  47   0   0   1  TRUE FALSE
# 5                    1014   3  30   0   0   0  TRUE FALSE
# 6                    1015  21  68   0   0   1  TRUE FALSE
# 7                    3367   0  10   0   0   0  TRUE FALSE
# 8                    1335 108 108   0   0   1  TRUE FALSE
# 9   Chr1  1575   G
# 10                   1007  26  11   0   0   0  TRUE FALSE
# 11                   1012  49  27   0   0   0  TRUE FALSE
# 12                   1013  19  28   0   0   0  TRUE FALSE
# 13                   1014  15  17   0   0   0  TRUE FALSE
# 14                   1015  46  42   0   0   1  TRUE FALSE
# 15                   3367   0  11   0   0   0  TRUE FALSE
# 16                   1335  37  99   0   0   0  TRUE FALSE
# 17  Chr1  6485   G
# 18                   1007  26  20   0   0   0  TRUE FALSE
# 19                   1012  33  39   0   0   0  TRUE FALSE
# 20                   1013  54  48   0   0   0  TRUE FALSE
# 21                   1014  13  10   0   0   0  TRUE FALSE
# 22                   1015  34  41   0   0   1  TRUE FALSE
# 23                   3367   0  42   0   0   0  TRUE FALSE
# 24                   1335  71  69   0   0   0  TRUE FALSE
# 25  Chr1  7181   G
# 26                   1007   0  37  31   0   0  TRUE FALSE
# 27                   1012   0  66  36   0   0  TRUE FALSE
# 28                   1013   0  30  86   0   0  TRUE FALSE
# 29                   1014   0  19   8   0   0  TRUE FALSE
# 30                   1015   0  44  33   0   1  TRUE FALSE
# 31                   3367   0  33   0   0   0  TRUE FALSE
# 32                   1335   0  94  78   0   0  TRUE FALSE
# 33  Chr1  7220   C
# 34                   1007  17   0  26   6   0  TRUE FALSE
# 35                   1012  45   0  31  14   0  TRUE FALSE
# 36                   1013 112   1  41  14   0  TRUE FALSE
# 37                   1014  16   1  16   2   0  TRUE FALSE
# 38                   1015  66   0  26   7   1  TRUE FALSE
# 39                   3367   0   0  24   0   0  TRUE FALSE
# 40                   1335  68   0  51  25   0  TRUE FALSE
# 41  Chr1  7661   T
# 42                   1007   0   0  10  14   0  TRUE FALSE
# 43                   1012   0   0   7  24   0  TRUE FALSE
# 44                   1013   0   0  32  23   1  TRUE FALSE
# 45                   1014   0   0   8  11   0  TRUE FALSE
# 46                   1015   0   0   6  41   0  TRUE FALSE
# 47                   3367   0   0   0   8   0  TRUE FALSE
# 48                   1335   0   0   8  42   0  TRUE FALSE
# 49  Chr1  8144   G
# 50                   1007  10  16   0   1   0  TRUE FALSE
# 51                   1012  19  28   0   0   1  TRUE FALSE
# 52                   1013  63  67   0   0   0  TRUE FALSE
# 53                   1014   7  12   0   0   0  TRUE FALSE
# 54                   1015  18  28   0   0   0  TRUE FALSE
# 55                   3367   0   7   0   0   0  TRUE FALSE
# 56                   1335  17  58   0   0   1  TRUE FALSE
# 57  Chr1  8208   G
# 58                   1007   0  15   0   8   1  TRUE FALSE
# 59                   1012   0  28   0  16   0  TRUE FALSE
# 60                   1013   0  24   0  63   0  TRUE FALSE
# 61                   1014   0  15   0   4   0  TRUE FALSE
# 62                   1015   0  25   0  13   1  TRUE FALSE
# 63                   3367   0   9   0   1   0  TRUE FALSE
# 64                   1335   0  49   0  21   1  TRUE FALSE
# 65  Chr1  8518   T
# 66                   1007   0   0  20  18   1 FALSE FALSE
# 67                   1012   0   0  45  30   1 FALSE FALSE
# 68                   1013   0   0  57  75   1 FALSE FALSE
# 69                   1014   0   0  10  32   0 FALSE FALSE
# 70                   1015   0   0  41  18   1 FALSE FALSE
# 71                   3367   0   0   0  11   0 FALSE FALSE
# 72                   1335   0   0 120  71   1 FALSE FALSE
# 73  Chr1  8552   G
# 74                   1007   3  13   0   0   0  TRUE FALSE
# 75                   1012  21  31   0   1   0  TRUE FALSE
# 76                   1013  33  35   0   0   1  TRUE FALSE
# 77                   1014   7  15   0   0   0  TRUE FALSE
# 78                   1015  14  22   0   0   0  TRUE FALSE
# 79                   3367   0  28   0   0   0  TRUE FALSE
# 80                   1335  27  59   0   0   0  TRUE FALSE
# 81  Chr1  8567   A
# 82                   1007  16  18   0   0   1  TRUE FALSE
# 83                   1012  34  35   0   0   1  TRUE FALSE
# 84                   1013  66  75   0   0   1  TRUE FALSE
# 85                   1014   9   4   0   0   0  TRUE FALSE
# 86                   1015  17  31   0   0   1  TRUE FALSE
# 87                   3367  29   0   0   0   0  TRUE FALSE
# 88                   1335  59  44   0   0   1  TRUE FALSE
# 89  Chr1  8670   A
# 90                   1007  19   0   0   7   0  TRUE FALSE
# 91                   1012  36   0   0  12   0  TRUE FALSE
# 92                   1013  44   0   0  12   0  TRUE FALSE
# 93                   1014  10   0   0   7   0  TRUE FALSE
# 94                   1015  24   0   0  11   1  TRUE FALSE
# 95                   3367  18   0   0   0   0  TRUE FALSE
# 96                   1335  27   0   0   6   0  TRUE FALSE
# 97  Chr1  8685   G
# 98                   1007   7  16   0   0   0  TRUE FALSE
# 99                   1012  12  37   0   0   0  TRUE FALSE
# 100                  1013  18  30   0   0   1  TRUE FALSE
# 101                  1014   5  32   0   0   0  TRUE FALSE
# 102                  1015  11  35   0   0   1  TRUE FALSE
# 103                  3367   0  12   0   0   0  TRUE FALSE
# 104                  1335   5  45   0   0   0  TRUE FALSE
# 105 Chr1 14361   A
# 106                  1007  29   7   0   0   0 FALSE FALSE
# 107                  1012  54   6   0   0   0 FALSE FALSE
# 108                  1013  28  12   0   0   1 FALSE FALSE
# 109                  1014  22   2   1   0   0 FALSE FALSE
# 110                  1015  51   9   0   0   0 FALSE FALSE
# 111                  3367  12   1   0   0   0 FALSE FALSE
# 112                  1335  64   8   0   0   0 FALSE FALSE
# 113 Chr1 15254   T
# 114                  1007  11   0   0  22   1 FALSE FALSE
# 115                  1012  28   0   0  53   1 FALSE FALSE
# 116                  1013  39   0   0  66   1 FALSE FALSE
# 117                  1014   3   0   0  14   1 FALSE FALSE
# 118                  1015  18   0   0  39   1 FALSE FALSE
# 119                  3367   0   0   0  89   0 FALSE FALSE
# 120                  1335  15   0   0  63   1 FALSE FALSE
# 121 Chr1 15280   T
# 122                  1007   0  14   0  32   1 FALSE FALSE
# 123                  1012   0  31   0  53   1 FALSE FALSE
# 124                  1013   0   6   0 102   0 FALSE FALSE
# 125                  1014   0   3   1  40   0 FALSE FALSE
# 126                  1015   0  22   1  51   1 FALSE FALSE
# 127                  3367   0   0   0  74   0 FALSE FALSE
# 128                  1335   0  26   0 109   1 FALSE FALSE
# 129 Chr1 16103   A
# 130                  1007  12   0  14   0   1 FALSE FALSE
# 131                  1012  50   0  19   0   1 FALSE FALSE
# 132                  1013  29   0  15   0   1 FALSE FALSE
# 133                  1014  28   0   2   0   0 FALSE FALSE
# 134                  1015  37   0  10   0   1 FALSE FALSE
# 135                  3367  41   0   0   0   0 FALSE FALSE
# 136                  1335  56   0  12   0   0 FALSE FALSE
# 137 Chr1 17587   A
# 138                  1007  22   2   0   0   0 FALSE FALSE
# 139                  1012  62   6   0   1   0 FALSE FALSE
# 140                  1013  22  12   0   0   1 FALSE FALSE
# 141                  1014  22   2   0   0   0 FALSE FALSE
# 142                  1015  29   3   0   0   0 FALSE FALSE
# 143                  3367  20   1   0   0   0 FALSE FALSE
# 144                  1335  82  11   0   0   0 FALSE FALSE
# 145 Chr1 18904   T
# 146                  1007   0   5   0  34   0 FALSE FALSE
# 147                  1012   0   4   0  39   0 FALSE FALSE
# 148                  1013   0   9   0  21   0 FALSE FALSE
# 149                  1014   0   3   0  21   0 FALSE FALSE
# 150                  1015   0   9   0  48   0 FALSE FALSE
# 151                  3367   0   5   0  96   0 FALSE FALSE
# 152                  1335   0  27   0  73   1 FALSE FALSE
# 153 Chr1 25546   A
# 154                  1007  31   0   0  14   1 FALSE FALSE
# 155                  1012  64   0   0  22   1 FALSE FALSE
# 156                  1013  20   0   0  50   1 FALSE FALSE
# 157                  1014  22   0   1  18   1 FALSE FALSE
# 158                  1015  64   0   0  18   1 FALSE FALSE
# 159                  3367  73   0   0   0   0 FALSE FALSE
# 160                  1335  80   0   0   5   0 FALSE FALSE
```

More sanity: there are 83 sites on Chr1 shared by zero strains in the tightest condition. (I.e., SAMTOOLS called it a SNP, but the read counts/proportions fall below our 3rd threshold). Are they due to low coverage? Seemingly yes:

```
zp3 <- snp.pattern[[3]] == 0
zr3 <- rownames(non.refs[[3]])[which(zp3)]
zc3 <- as.integer(unlist(lapply(strsplit(zr3[1:min(100,length(zr3))],':',fixed=TRUE),function(x){x[2]})))
zc3

#  [1]   91284  127986  161271  196862  196864  199166  282391  289344  289363  314132  314661
# [12]  438976  447253  475823  501830  501975  504462  652889  657955  692139  709443  762174
# [23]  826899  856950  875379  913014  938651  967184 1036942 1100300 1113225 1181146 1203203
# [34] 1210360 1212223 1224082 1270250 1270251 1348311 1431628 1473437 1516083 1526912 1628300
# [45] 1637082 1686331 1736789 1763837 1782580 1967158 2024930 2075603 2098145 2110716 2194162
# [56] 2242316 2258647 2261176 2325671 2376777 2432898 2441781 2498706 2550796 2554565 2581374
# [67] 2614631 2619528 2659281 2675254 2691279 2703771 2737914 2744068 2802553 2842231 2846930
# [78] 2906880 2931365 2948653 2957936 3014028 3016252

seecounts(zc3[1:5], snp.tables=snp.tables)

#     chr    pos Ref Strain   A G  C  T SNP  exon indel nrf rat
# 1  Chr1  91284   T
# 2                    1007   0 0  0 17   0 FALSE FALSE
# 3                    1012   0 0  0 38   0 FALSE FALSE
# 4                    1013   2 0  0 13   0 FALSE FALSE
# 5                    1014   0 0  0 20   0 FALSE FALSE
# 6                    1015   0 0  0 35   0 FALSE FALSE
# 7                    3367   3 0  0 12   1 FALSE FALSE
# 8                    1335   0 0  0 47   0 FALSE FALSE
# 9  Chr1 127986   A
# 10                   1007  47 0  0  0   0  TRUE FALSE
# 11                   1012  92 0  0  0   0  TRUE FALSE
# 12                   1013  19 1  0  0   0  TRUE FALSE
# 13                   1014  73 0  0  0   0  TRUE FALSE
# 14                   1015  83 0  0  0   0  TRUE FALSE
# 15                   3367  13 3  0  0   1  TRUE FALSE
# 16                   1335 160 0  0  0   0  TRUE FALSE
# 17 Chr1 161271   A
# 18                   1007  31 0  0  0   0  TRUE FALSE
# 19                   1012  47 0  0  0   0  TRUE FALSE
# 20                   1013  18 3  0  0   0  TRUE FALSE
# 21                   1014  30 0  0  0   0  TRUE FALSE
# 22                   1015  59 0  0  0   0  TRUE FALSE
# 23                   3367   8 3  0  0   1  TRUE FALSE
# 24                   1335 102 0  0  0   0  TRUE FALSE
# 25 Chr1 196862   C
# 26                   1007   0 0 10  0   0 FALSE FALSE
# 27                   1012   0 0 22  0   0 FALSE FALSE
# 28                   1013   0 0  8  2   0 FALSE FALSE
# 29                   1014   0 0 14  0   0 FALSE FALSE
# 30                   1015   0 0 18  0   0 FALSE FALSE
# 31                   3367   1 0  4  3   1 FALSE FALSE
# 32                   1335   0 0 18  0   0 FALSE FALSE
# 33 Chr1 196864   T
# 34                   1007   0 0   0 11    0 FALSE FALSE
# 35                   1012   0 0   1 23    0 FALSE FALSE
# 36                   1013   3 0   0  8    1 FALSE FALSE
# 37                   1014   0 0   0 12    0 FALSE FALSE
# 38                   1015   1 0   1 19    0 FALSE FALSE
# 39                   3367   3 0   0  4    1 FALSE FALSE
# 40                   1335   0 0   1 19    0 FALSE FALSE
```

7.3 Main Analysis

Turning to the main analysis, there is a large increase in the number of consistent positions between the loose and medium stringency levels; medium and tight are similar in most respects. The likely interpretation is that the loose criterion is including many “SNPs” induced by read errors, and that either of the tighter criteria are successfully filtering them out. In the interest of simplicity, the narrative below will focus on the shared SNPs at the medium stringency level (the “count2” column in the data frame), although the numbers for all three (sometimes all 4) are displayed. Also note that the prose and some comments in the code were based on the Chr1 analysis, and so may occasionally be off-target for the whole-genome data.

```
# Show a subset of pat.summaries, optionally with totals of count_i in last row, and optionally
# aggregating low-count rows as ``Other''
#
#   sharedBy=c(2,4) selects SNPs shared by 2 or 4 strains,
#   subset=as.octmode('35') select those with sharing pattern a subset (optionally proper) of this
#   split=as.octmode('14') additionally restricts to patterns stradling split/subset minus split
#   c2.thresh=42 suppresses printout of rows with count2 < 42
#   restrict.to=c(0,42,127) restrict to these 3 rows
showgroup <- function(p.summ=pat.summaries, sharedBy=0:7, subset=127, split=NULL, proper.subset=F,
                      total=T, c2.thresh=0, fourteenth=F, restrict.to=NULL){
  # pick just those bit patterns that are subsets of 'subset'
  pick <- bitwAnd(0:127,bitwNot(subset))==0
  if(proper.subset){
    pick[subset+1] <- F
  }
  if(!is.null(split)){ # AND that stradle left/right subtrees
    cosplit <- bitwAnd(subset,bitwNot(split))
    pick <- pick & bitwAnd(0:127,split)!=0 & bitwAnd(0:127,cosplit)!=0
  }
  # and have desired shareBy counts
  pick <- pick & (p.summ$sharedBy %in% sharedBy)
  # and are among the set of interest
  if(!is.null(restrict.to)){
    pick <- pick & (0:127 %in% restrict.to)
  }
  # find rows with low counts
  pick.low <- pick & (p.summ$count2 < c2.thresh)
  # now show them
  show <- p.summ[pick & ! pick.low,]
  # rename columns just to narrow the printouts
  colnames(show) <- c('Pat','ShrBy','1007', '1012', '1013', '1014', '1015', '3367', '1335',
                      'count1', 'count2', 'count3','count4')
  show[,1] <- format(show[,1])  # convert octal col to char so can override in last row(2)
  nlow <- sum(pick.low)
  if(nlow > 0){
    n <- nrow(show)+1
    lows <- apply(p.summ[pick.low,10:13],2,sum)
    show[n,10:13] <- lows
    show[n,1:9] <- ''
    row.names(show)[n] <- 'Other'
    if(fourteenth){
      # do this: add 14th col just to hold this comment:
      show <- cbind(show,' '='', stringsAsFactors=F)
      show[n,14] <- paste('(', nlow, 'rows w/ c2 <', c2.thresh, ')')
    } else {
      ## or this (looks a bit funky, but fits across page without line-wrap):
      show[n,1:8] <-c('(', nlow, 'rows', 'w/', 'c2', '<', c2.thresh, ')')
    }
  }
  if(total){
    n <- nrow(show)+1
    tots <- apply(show[,10:13],2,sum)
    show[n,10:13] <- tots
    show[n,1:9] <- ''
    row.names(show)[n] <- 'Total'
    if(ncol(show)==14){show[n,14]<-''}
  }
  return(show)
}
```

First, are there any SNPs that are not "consistent SNPs?" Yes, a few in c3. As noted above, they seem to be mainly low-coverage positions.

```
showgroup(pat.summaries,0,total=F)  # chr1 totals: 0 0 83

#   Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 1   0     0                                         0      0     83      0
```

Next, look at completely shared SNPs, those found in all 7 strains.

```
showgroup(pat.summaries,7,total=F) # Chr1 count1 = 8593, count2 = 7054, count3 = 4790 c4=1641

#     Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 128 177     7    X    X    X    X    X    X    X   8593   7054   4790   1641
```

I.e., of the 46896 consistent positions, 7054 or 15% are shared by all 7 strains.

Next look at singletons, aka private SNPs—SNPs that are called in one strain and no other strain has a significant number of non-ref reads at that position. Presumably these are variants that arose in a given population after it separated from the others.

```
showgroup(pat.summaries,1)  # chr1 totals: 9669  18865  19670  23574

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 2     001     1                                  X     13     41     70    135
# 3     002     1                             X        4551   8813   9159  10949
# 5     004     1                        X               90    207    243    385
# 9     010     1                   X                    22     61     78    113
# 17    020     1              X                       4954   9652   9985  11677
# 33    040     1         X                              29     61     95    174
# 65    100     1    X                                   10     30     40    141
# Total                                                9669  18865  19670  23574
```

The import of shared/private SNPs changes between sexual and asexual populations. Presumably asexuals slowly gain and rarely lose private SNPs; shared ones predate separation of the lineages. In sexual lineages, however, SNPs may be rather freely "gained" or "lost," merely by recombination (converting between homo- and heterozygous in the sample we sequenced). Thus, the low private counts for the 5 L-isolates compared to the large count of het positions overall suggest that (a) they are asexual, and (b) none of them has been isolated from the others for very long (if at all). Conversely, the high counts for Italy and Wales suggest that (a) if asexual, they have been separated from each other and from the rest for a long time, but (b) if sexual, there is little surprise: we have ≈160K SNPs shared between the two (90K just in those two (below), plus 70K shared by all 7), and ≈90K additional positions that are het in one but not the other. These are close to, but not exactly equal to, the 1:2:1 ratios we would naively expect from two samples of a single HWE population. The most parsimonious explanation seems to be that the H-clade is sexual, but perhaps some het positions private to each population separates them.

Aside: counts of "consistent" SNPs minus these singletons yeilds count of shared SNPs:

```
singlets <- apply(pat.summaries[pat.summaries$sharedBy==1,10:13],2,sum)
rbind(consistent=consistent.count,singlets=singlets,shared=consistent.count-singlets)

#            count1 count2 count3 count4
# consistent  36040  46896  47174  47499
# singlets     9669  18865  19670  23574
# shared      26371  28031  27504  23925
```

The slightly higher count of shared positions in the medium case further supports this choice for subsequent analysis.

Next look at consistent SNPs shared between just a pair of isolates.

```
showgroup(pat.summaries,2) # chr 1 counts: 7641   9549   9472  6924

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 4     003     2                             X    X    239     16     28     31
# 6     005     2                        X         X     11     18     35     52
# 7     006     2                        X    X         141     13     21     41
# 10    011     2                   X              X      6      7      7     14
# 11    012     2                   X         X         179     18      2      5
# 13    014     2                   X    X               10      2      5     12
# 18    021     2              X                   X    243      9     11      9
# 19    022     2              X              X        5920   9365   9094   6177
# 21    024     2              X         X              125     12     28     46
# 25    030     2              X    X                   222     18      5      9
# 34    041     2         X                        X      3      4     13     36
# 35    042     2         X                   X         150      1     16     27
# 37    044     2         X              X                5     13     75    155
# 41    050     2         X         X                     7      2      5      7
# 49    060     2         X    X                        165      4     26     28
# 66    101     2    X                             X      1      2     13     20
# 67    102     2    X                        X          93      6      7     25
# 69    104     2    X                   X                5      9     34    107
# 73    110     2    X              X                     2      2      4      7
# 81    120     2    X         X                        105      9      9     31
# 97    140     2    X    X                               9     19     34     85
# Total                                                7641   9549   9472   6924
```

I.e., of the 9549 paired SNPs, 9365 or 98.1% are found between Italy and Wales, with comparatively few shared between any other pairs (only).

SNPs shared among exactly 3 isolates are relatively rare. (The 5 trios containing both Italy and Wales predominate in the loose set, probably because they share many pairs that become triples with the addition of a few read errors.)

```
showgroup(pat.summaries,3) # chr 1 counts: 1438    294    671   1034

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 8     007     3                        X    X    X      9      2      9     10
# 12    013     3                   X         X    X     17      3      2      2
# 14    015     3                   X    X         X      7      6      4      7
# 15    016     3                   X    X    X           9      1      2      2
# 20    023     3              X              X    X    327     20     29     17
# 22    025     3              X         X         X      9      4     12     21
# 23    026     3              X         X    X         185     27     31     32
# 26    031     3              X    X              X     20      2      0      0
# 27    032     3              X    X         X         324     18      8      5
# 29    034     3              X    X    X               11      3      1      1
# 36    043     3         X                   X    X     21      8     14      6
# 38    045     3         X              X         X      6     26    130    131
# 39    046     3         X              X    X          11     12     34     55
# 42    051     3         X         X              X      0      1      2      4
# 43    052     3         X         X         X           9      0      2      1
# 45    054     3         X         X    X                1      6     17     18
# 50    061     3         X    X                   X     12      2     14     12
# 51    062     3         X    X              X         227     17     37     36
# 53    064     3         X    X         X                9      9     36     60
# 57    070     3         X    X    X                    13      1      1      2
# 68    103     3    X                        X    X     11      4      4      8
# 70    105     3    X                   X         X      4      5     25     63
# 71    106     3    X                   X    X           3      9     15     27
# 74    111     3    X              X              X      1      0      1      1
# 75    112     3    X              X         X           4      1      0      0
# 77    114     3    X              X    X                1      2      2      8
# 82    121     3    X         X                   X      8      0      4      4
# 83    122     3    X         X              X         134     10     10     26
# 85    124     3    X         X         X               11      7     20     35
# 89    130     3    X         X    X                     7      1      2      1
# 98    141     3    X    X                        X      2      5     15     40
# 99    142     3    X    X                   X           9      1      9     15
# 101   144     3    X    X              X                6     75    167    355
# 105   150     3    X    X         X                     0      0      4      6
# 113   160     3    X    X    X                         10      6      8     23
# Total                                                1438    294    671   1034
```

Four-way sharing is more common, but dominated by the coastal (i.e., non-Gyre) L-clade isolates. This is likely a reflection of the strong 5-way sharing among the L-clade, from which the Gyre commonly drops out due to the lower coverage/higher error rate in that sequencing run.

```
showgroup(pat.summaries,4) # chr 1 counts: 564   1346   2552  3479

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 16    017     4                   X    X    X    X      1      5      1      2
# 24    027     4              X         X    X    X     22     14     28     24
# 28    033     4              X    X         X    X     30      2      4      6
# 30    035     4              X    X    X         X      4      2      1      0
# 31    036     4              X    X    X    X          12      4      1      3
# 40    047     4         X              X    X    X      2     18     42     60
# 44    053     4         X         X         X    X      1      0      2      2
# 46    055     4         X         X    X         X      6     12     29     36
# 47    056     4         X         X    X    X           3      2      2      5
# 52    063     4         X    X              X    X     18     11     24     21
# 54    065     4         X    X         X         X      9     17     41     48
# 55    066     4         X    X         X    X          25     37    102     76
# 58    071     4         X    X    X              X      2      1      2      0
# 59    072     4         X    X    X         X           8      6      3      2
# 61    074     4         X    X    X    X                2      3      2      7
# 72    107     4    X                   X    X    X      5      5     11     16
# 76    113     4    X              X         X    X      1      2      0      1
# 78    115     4    X              X    X         X      0      4      3      9
# 79    116     4    X              X    X    X           0      1      0      5
# 84    123     4    X         X              X    X     10      9      8      8
# 86    125     4    X         X         X         X      4      3     13     16
# 87    126     4    X         X         X    X          12     20     21     43
# 90    131     4    X         X    X              X      0      1      0      2
# 91    132     4    X         X    X         X           9      1      2      1
# 93    134     4    X         X    X    X                2      2      1      3
# 100   143     4    X    X                   X    X      1      3      5     20
# 102   145     4    X    X              X         X    320   1005   1973   2585
# 103   146     4    X    X              X    X           7     34     65    140
# 106   151     4    X    X         X              X      1      4      5     14
# 107   152     4    X    X         X         X           2      1      1      4
# 109   154     4    X    X         X    X               16     53     49    103
# 114   161     4    X    X    X                   X      9      5      9     18
# 115   162     4    X    X    X              X          11     12     21     33
# 117   164     4    X    X    X         X                8     47     80    163
# 121   170     4    X    X    X    X                     1      0      1      3
# Total                                                 564   1346   2552   3479
```

Five-way sharing is much more common, and is strongly dominated by the 5 L-clade isolates.

```
showgroup(pat.summaries,5) # chr 1 counts: 3969   5047   4624  6125

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 32    037     5              X    X    X    X    X     17     10      8      5
# 48    057     5         X         X    X    X    X      5      7      9     17
# 56    067     5         X    X         X    X    X     33     78    201    104
# 60    073     5         X    X    X         X    X      3      3      6      3
# 62    075     5         X    X    X    X         X      7     11     13     11
# 63    076     5         X    X    X    X    X           9     11     17      7
# 80    117     5    X              X    X    X    X      0      2      1      7
# 88    127     5    X         X         X    X    X     11     17     30     47
# 92    133     5    X         X    X         X    X      3      2      0      0
# 94    135     5    X         X    X    X         X      2      1      3      7
# 95    136     5    X         X    X    X    X           4      6      3      5
# 104   147     5    X    X              X    X    X    125    307    590   1160
# 108   153     5    X    X         X         X    X      3      3      1      7
# 110   155     5    X    X         X    X         X   3484   3912   2642   3228
# 111   156     5    X    X         X    X    X           8     15     15     43
# 116   163     5    X    X    X              X    X     12     18     23     33
# 118   165     5    X    X    X         X         X    201    453    787   1140
# 119   166     5    X    X    X         X    X          30    157    247    254
# 122   171     5    X    X    X    X              X      3      6      2      7
# 123   172     5    X    X    X    X         X           2      6      3      5
# 125   174     5    X    X    X    X    X                7     22     23     35
# Total                                                 3969   5047   4624   6125
```

Six-way sharing is also common, with the sets *ex*cluding Gyre, Italy, or Wales having the most mutually-shared SNPs.

```
showgroup(pat.summaries,6) # chr 1 counts:  4166   4741   5312  4722

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 64    077     6         X    X    X    X    X    X     43     48     60     26
# 96    137     6    X         X    X    X    X    X     11      6     12     14
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 124   173     6    X    X    X    X         X    X     17      9      7      3
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 127   176     6    X    X    X    X    X    X          45    114     77     68
# Total                                                 4166   4741   5312   4722
```

8 Trees

So, overall, the picture looks like a long shared history (7054 7-way shared positions), followed by a split of the 5 L-isolates from the 2 H-isolates, then a long shared history in the 5 (3912 quintuples), in parallel with a long shared history in H- (9365 pairs), then separate histories in Italy and Wales (>8813 "private" SNPs in each, although again if they are sexual, many of these just reflect HWE), and very limited differentiation among the 5 L-isolates.

Branch lengths of course depend on filtering criteria used (and, of course, full vs Chr1 differ by about a factor of 10), but the tree *topology* appears to be fairly stable. Various versions are drawn below, exactly to explore how robust this story is. I think we should go with "medium stringency" SNP filtering (based on un-qfiltered reads).

NOTE: Much of this analysis make less sense for q-filtered read data, since (a) the point of the SNP filtering was to try to correct for noise in the raw reads, which may (or may not; haven't looked closely, yet) be largely fixed by qfiltering (e.g., "loose" or no SNP filtering may be more appropriate, post-q-filtering, esp. if we had re-run SAMTools to call SNPs based on the q-filtered reads), and (b) tree topology *does* appear to change, in that Gyre's coverage has been so sharply reduced by qfiltering that it clearly stands aside from the others (and that's confirmed by bootstrap), but this also seems to be clearly a technical rather than a biological artifact. SO, code below will run on q-filtered data, but *is not tuned to it*. Likewise, most comments in the prose below were made to describe the un-q-filtered data, and *are misleading and in some cases flatly wrong* for qfiltered data, but it doesn't seem worthwhile to bother with a rewrite...

Trees are coded in newick format, which doesn't seem to tolerate line-breaks; print with line-wrap:.

```
# wrap a long char string across multiple lines in printout
cat.hardwrap <- function(str,width=80){
  while(nchar(str)>width){
    cat(substr(str,1,width),'\n')
    str <- substr(str,width+1,nchar(str))
  }
  cat(str,'\n')
}
```

Trees are built as follows. Code for drawing, especially, is specific to the topology of the medium tree, and placement of some of the figure elements have been hand-optimized for this case; drawings for the other variants will not be as pretty.

```
# set up for tree figs

# the newick parser in ape seems to be confused by commas and parens in
# tip names, and blanks are not allowed, so replace by *, <, >, _, resp.
newick.name <- function(name){
  name <- gsub(' ', '_', name, fixed=TRUE)
  name <- gsub(',', '*', name, fixed=TRUE)
  name <- gsub('(', '<', name, fixed=TRUE)
  name <- gsub(')', '>', name, fixed=TRUE)
  return(name)
}
# undo above changes
newick.name.undo <- function(name){
 #name <- gsub('_', ' ', name, fixed=TRUE) # unnecessary; ape plot routine handles this one
  name <- gsub('*', ',', name, fixed=TRUE)
  name <- gsub('<', '(', name, fixed=TRUE)
  name <- gsub('>', ')', name, fixed=TRUE)
  return(name)
}

# make a newick string from tree;  see it below
# 'pre' is prefixed to ccmpid; 'nb' optionally included;
# 'alt' can be used instead of pre/ccmp/nb/where for less formal labeling
# 'newstyle'==T => new node label: [nb_]where[(pre-less-id)]
# 'newstyle'==F => old node label: [nb_][pre id]where
newickize <- function(tree,pre='CCMP',nb=TRUE,alt=F,newstyle=TRUE){
  if(is.null(tree$where)){
    # not a leaf; paste together newick from subtrees
    sub1 <- newickize(tree$sub1,pre=pre,nb=nb,alt=alt,newstyle=newstyle)
    sub2 <- newickize(tree$sub2,pre=pre,nb=nb,alt=alt,newstyle=newstyle)
    new <- paste( '(', sub1, ',', sub2, ')', sep='')
    if(!is.null(tree$length)){
      # internal node, add length
      return(paste(new, ':', tree$length, sep=''))
    } else {
      # top level; escape blanks and add trailing ';'
      return(paste(gsub(' ', '_', new), ';', sep=''))
    }
  } else {
    # a leaf; build label and branch length
    if(alt){
      # label is just alt; if alt omitted, default to where
      new <- newick.name(ifelse( is.null(tree$alt), tree$where, tree$alt ))
    } else {
      if(newstyle){
        # new node label = [nb_]where[(pre-less-id)]
        new <- ifelse( nb && !is.null(tree$nb), paste(tree$nb, '_', sep =''), '' )
        new <- newick.name(paste(new, tree$where, sep=''))
        new <- ifelse( is.null(tree$id), new, paste(new, '_(', tree$id, ')', sep='') )
        new <- newick.name(new)
      } else {
        # old style node label = [nb_][pre id]where
        new <- ifelse( nb && !is.null(tree$nb), paste(tree$nb, '_', sep =''), '' )
        new <- ifelse( is.null(tree$id), new, paste(new, pre, tree$id, '_', sep='') )
        new <- newick.name(paste(new, tree$where, sep=''))
      }
    }
    #add length to either
    new <- paste(new, ':', tree$length, sep='')
  }
  return(new)
}

# Make a tree as nested lists, **based on the chr1, count2 topology**, but using any of the counts.
#   Internal nodes have subtrees sub1/2 and length
#   Root has sub1/2, but no length
#   Leaves have where, length, optionally, id, alt, nb.  (Omit id for 'outgroup'. Use 'alt' for less formal
#     labeling in cartoon version; it defaults to 'where'. Use 'nb' to add abcde annotations for legend.)
# The single parameter v is any of the 4 count vectors contained in pat.summaries (most conveniently
# indexed in octal).  E.g., make.tree(pat.summaries[,'count2']) reproduces the count2 tree.
# (This was previously built by hand-pasting the edge lengths; tree.by.hand is retained in appendix
# for comparison, & its counts are in comments below).
#
make.tree <- function(v){
  pat.count <- function(pat, pat.counts=v){return(pat.counts[1+strtoi(pat,8)])}
  thetree <-
    list(
      sub1 = list(
        sub1 = list(
          sub1 = list(id=3367, length=pat.count('002'), where='Venice, Italy', alt='Venice'), #8813
          sub2 = list(id=1013, length=pat.count('020'), where='Wales, UK'),                   #9652
          length=pat.count('022')),                                                           #9365
        sub2 =  list(
          sub1 = list(
            sub1 = list(
              sub1 = list(id=1007, length=pat.count('100'), nb='e', where='Virginia, USA'),    #30
              sub2 = list(id=1012, length=pat.count('040'), nb='d', where='Perth, W. Australia', alt='Perth'), #61
              length=pat.count('140')),                                                        #19
            sub2 = list(
              sub1 = list(id=1015, length=pat.count('004'),nb='c', where='Washington, USA', alt='Puget Sound'), #207
              sub2 = list(id=1335, length=pat.count('001'), nb='b', where='New York, USA',   alt='NY'), #41
              length=pat.count('005')),                                                        #18
            length=pat.count('145')),                                                          #1005
          sub2 = list(id=1014, length=pat.count('010'), nb='a', where='N. Pacific Gyre'),      #61
          length=pat.count('155')),                                                            #3912
        length=pat.count('177')),                                                              #7054
      sub2 = list(length=0, where='outgroup')
    )
  return(thetree)
}
```

Code to plot a tree given newick description. Again, code is somewhat general, but has some specializations tied to the medium-stringency, full-genome, un-qfiltered data.

```
# run following 2 lines after an R upgrade
# update.packages()
# install.packages("ape")
library(ape)
show.tree <- function(newick.str=newick.medium,
                      col.edge  ='darkblue', lwd.edge =2,
                      col.elabel='darkblue',                cex.elabel=0.8, font.elabel=3,
                      col.arrow ='red',      lwd.arrow=1.5, cex.arrow =0.9, font.arrow =4,
                      col.clade ='black',    lwd.clade=1,   cex.clade =1.0, font.clade =3,
                      col.legbox='beige',                   cex.legend=0.8,
                      col.tip   ='darkblue',                                font.tip   =4,
                      plusx=FALSE, pltdebug=FALSE, total.snps=consistent.count[2],
                      straight.arrow=FALSE){

  ####
  #
  # ADJUST NEWICK & GET LENGTHS, COORDINATES
  #
  newick.str.noout <- sub('outgroup','_',newick.str) # Hide outgroup ('_' prints as blank)
  the.tree <- read.tree(text=newick.str.noout)

  ## nasty hack: ape's newick parser seems to be confused by commas, () in tip labels, so
  ## newickize replaced them by '*<>'; before plotting, I want to convert them back, and hope
  ## this doesn't break anything else...  And if a revised version of ape changes the internal
  ## representation of a tree, this may need to be redone.
  the.tree$tip.label <- newick.name.undo(the.tree$tip.label)

  # extract branch lengths as char string of comma-separated numbers via pattern matching hack:
  # lengths always preceded by colon
  lengths.ch <- strsplit(paste(newick.str,':'),'[^0-9][^:]*:')[[1]]

  # then convert to ints, dropping empty string at front
  lengths.int <- scan(what=integer(),quiet=T,sep=',',text=lengths.ch[-1])

  # then to data frame with named rows; a..g are terminal branches; others are internal.
  # a..e match legend in plot; f/g = wales/italy.  lengths appear in postfix order of
  # newick tree, and ape draws the 1st of them at the bottom of the plot.
  lmed <- data.frame(lengths=lengths.int,
                     row.names=c('g','f','fg','e','d','de','c','b','bc','bcde','a','abcde','all','out'))
```

```
 # extract counts needed for legend:
#leg.counts <- c(      61, 41,207, 61, 30, 1005,   18, 19) #by hand, medium chr1
 leg.counts <- lmed[c('a','b','c','d','e','bcde','bc','de'),1]
 discord <- total.snps - sum(lmed$lengths)

 #tree.labels <- list( ## x,y,text; coords are all picked by eye
 #   3000, 3.62, paste(lmed['all'  ,1], 'shared by 7', sep='\n'), # 7054
 #   8900, 5.75, paste(lmed['abcde',1], 'by 5'       , sep='\n'), # 3912
 #  12000, 1.50, paste(lmed['fg'   ,1], 'shared by 2', sep='\n'), # 9365
 #  21000, 2.00, paste(lmed['f'    ,1], 'only\nin Wales'),         # 9652
 #  21000, 1.00, paste(lmed['g'    ,1], 'only\nin Italy'),         # 8813
 #  11500, 4.50, '*')
 # automating x-placement, below; retain above for comparison...
 tip <- integer(7)  # x coords of tree tips
 tip[1] <-sum(lmed[c('all','fg','g'),1])
 tip[2] <-sum(lmed[c('all','fg','f'),1])
 tip[3] <-sum(lmed[c('all','abcde','bcde','de','e'),1])
 tip[4] <-sum(lmed[c('all','abcde','bcde','de','d'),1])
 tip[5] <-sum(lmed[c('all','abcde','bcde','bc','c'),1])
 tip[6] <-sum(lmed[c('all','abcde','bcde','bc','b'),1])
 tip[7] <-sum(lmed[c('all','abcde','a'),1])

 inode <- integer(5) # x coords of (some) internal nodes
 inode[1] <- 0                                     # root
 inode[2] <- lmed['all',1]                         # lca of all
 inode[3] <- sum(lmed[c('all','fg'),1])            # lca H-clade
 inode[4] <- sum(lmed[c('all','abcde'),1])         # lca L-clade
 inode[5] <- sum(lmed[c('all','abcde','bcde'),1]) # lca L-clade, nonGyre
 tree.labels <- list( ## x,y,text; y coords partially picked by eye
   sum(inode[c(1,2)])/2, 3.62, paste(lmed['all'  ,1], 'shared by 7', sep='\n'), # 7054
   sum(inode[c(2,4)])/2, 5.75, paste(lmed['abcde',1], 'by 5'       , sep='\n'), # 3912
   sum(inode[c(2,3)])/2, 1.50, paste(lmed['fg'   ,1], 'shared by 2', sep='\n'), # 9365
   (inode[3]+tip[2])/2,  2.00, paste(lmed['f'    ,1], 'only\nin 1013'),          # 9652
   (inode[3]+tip[1])/2,  1.00, paste(lmed['g'    ,1], 'only\nin 3367'),          # 8813
   sum(inode[c(4,5)])/2, 4.35, '* ')


 tree.labels <- list( ## x,y,text; y coords partially picked by eye
   sum(inode[c(1,2)])/2, 3.62, paste(lmed['all'  ,1], 'in 7', sep='\n'), # 7054
   sum(inode[c(2,4)])/2, 5.75, paste(lmed['abcde',1], 'in 5', sep='\n'), # 3912
   sum(inode[c(2,3)])/2, 1.50, paste(lmed['fg'   ,1], 'in 2', sep='\n'), # 9365
   (inode[3]+tip[2])/2,  2.00, paste(lmed['f'    ,1], 'only\nin 1013'),   # 9652
   (inode[3]+tip[1])/2,  1.00, paste(lmed['g'    ,1], 'only\nin 3367'),   # 8813
   sum(inode[c(4,5)])/2, 4.35, '* ')

 ####
 #
 # BOGUS PLOT
 #
 # a messy bit: need string widths to set xlim; but strwidth needs x-scale so must plot first.
 # M plot completely invisible, overlay 2nd plot via par(new=F...) .
 #
 # PROVISIONALLY set x.lim here at about 30% wider than tree; fine tune it for the real plot
 # based on strwidth(tip labels) below.
 #
 provisional.tree.x.lim <- 1.3 * max(tip) # <== PROVISIONAL plot width
 plot(0,0, type='n', bty='n', xaxt='n', yaxt='n', xlab='', ylab='', xlim=c(0,provisional.tree.x.lim), ylim=c(0,7))

 tiplabel.x <- integer(7)
 for(i in 1:7){
   # see warning above about internals of the.tree; labels have '_', printed as ' '.
   tiplabel.x[i] <- tip[i]+strwidth(gsub('_',' ',the.tree$tip.label[i],fixed=T), font=font.tip)
 }


 # visually show tip coords & max x to debug placement issues
 plt.debug <- function(tree.x.lim, tip, tiplabel.x, spx=NULL, spy=NULL){
   if(pltdebug){ # F to hide/T to show debug
     cat('Tip labels:', paste(the.tree$tip.label,sep='',collapse='/'), '\n')
     axis(2) # useful only for placing labels
     for(i in 1:7){
       points(c(tip[i],tiplabel.x[i]),c(i,i)) # debug: do I have right tip coordinates?
     }
     lines(rep(tree.x.lim,2),c(0,7)) # where is right edge?
     if(!is.null(spx)){
       points(spx,spy) # show spline control points, for tweaking
     }
   }
 }
```

```
plt.debug(provisional.tree.x.lim, tip, tiplabel.x)

label.end.H <- max(tiplabel.x[1:2])
label.end.L <- max(tiplabel.x[3:7])
clade.dx <- strwidth('x') # space between clade marker line and its label
xdel <- 3*clade.dx        # space between labeled clade tips and marker line

tree.x.lim <- 1.03*(max(tiplabel.x)+xdel)  # <== FINAL plot width
tree.y.lim <- 7
if(pltdebug){cat('Plot width hacking:', provisional.tree.x.lim, tree.x.lim, tree.x.lim/1.03/max(tip), clade.dx)}

par(new=T)  # I.e., NOT starting a new plot

####
#
# REAL PLOT
#
plot(the.tree,
     x.lim = c(0, tree.x.lim),
     y.lim = c(0, tree.y.lim),
     font=font.tip, label.offset=100,             # bold-italic, nudged slightly right
     tip.color=col.tip, edge.color=col.edge,
     edge.width=lwd.edge,
     edge.lty=c(1,1,1,1, 1 ,1,1,1,1,1,1,1,1,0)    # 5th is bottleneck edge; 14th is outgroup
    )
lines(00+c(0,0),c(3.5,6),col='white',lwd=6)       # Hide vertical line to outgroup
axis(1, pos=0.25, at=seq(0,25,by=5)*10^round(log10(max(tip)/25)))

if(pltdebug){text(tip[1]+100, 1.0, 'Venice, Italy (3367)', adj=0, font=font.tip)}

####
#
# BOTTLENECK ANNOTATION
#
# spline/elipse control points (spy/y) & tweaks thereto (dx/y)
dx <- 0.01 * tree.x.lim
dy <- .04
spx <- c(7400, 7400, 9900, 10500) # by eye, chr1, for comparison
spx <- c(inode[2]+dx,inode[2]+dx,inode[4]-3*dx,inode[4]-dx)
spy <- c( 3.8,  3.9,  5.6-dy,   5.6-dy)

plt.debug(tree.x.lim, tip, tiplabel.x, spx, spy)

if(T){
  #elipse version, defined by rect thru 2 middle pts of spx/y
  spf<-function(x){
    ifelse(x <= spx[2], spy[1],
           ifelse(x >= spx[3], spy[4],
                  spy[2]+(spy[3]-spy[2])*sqrt(pmax(0,1-((x-spx[3])/(spx[3]-spx[2]))^2))))
    }
} else {
  # spline version
  spf <- splinefun(spx,spy,method='hyman')
}
serx <- seq(spx[1],spx[length(spx)],length.out=50)
sery <- spf(serx)
tailx <- spx[1]
taily <- spy[1]
headx <- spx[4]
heady <- spy[4]
textx <- (headx+tailx)/2+(headx-tailx)*(-.01)
texty <- (heady+taily)/2+(heady-taily)*(-.10)
bottle.txt <- "inbreeding\nLoH / LoS"
if(!straight.arrow){
  arrows(headx,heady,headx+tree.x.lim*1e-3,heady, length=.1,col=col.arrow,lwd=lwd.arrow)
  lines(rev(serx), rev(sery), lty=c(5,1),col=col.arrow, lwd=lwd.arrow)
  textangle <- 66
  textadj <- c(0,0)
} else {
  # Tweak positioning slightly; visualize a rectangle from 7-node to base of L-clade;
  # center text, rotated, on diagonal towards L-clade; ditto the straight arrow.
  llx <- inode[2] # the aforementioned rectangle
  urx <- inode[4]
  lly <- 3.62
  ury <- 5.75
  # rect(llx,lly,urx,ury) # show rect for debug
  textx <- (llx+urx)/2    # center text
  texty <- (lly+ury)/2
  textangle <- atan(grconvertY(ury-lly,to='dev')/grconvertX(urx-llx,to='dev'))*360/(2*pi)
  textadj <- c(0.50, 0.43) #tweak position; ".5" = center in x , ".43" raises, THEN rotate.
  alpha <- .78 # fraction along diag at which arrow begins
  beta  <- .95 # ... and ends
  arrows((1-alpha)*llx + alpha*urx,
         (1-alpha)*lly + alpha*ury,
         (1-beta)*llx  + beta*urx,
         (1-beta)*lly  + beta*ury, length=.1,col=col.arrow,lwd=lwd.arrow,angle=25)
}
if(T){
  text(textx, texty, bottle.txt, srt=textangle, font=font.arrow, cex=cex.arrow,
       col=col.arrow, adj=textadj)
} else {
  # experiment at wrapping text along curved path; unpretty, but retain for now, maybe revisit
  bottlec <- strsplit(bottle,split=NULL)[[1]]
  for(i in 1:length(bottlec)){
    text(xser[i],yser[i],bottlec[i], srt=65, font=4, cex=.7, col=col.arrow)
  }
}

####
#
# CLADE ANNOTATION
#
clade.L.x <- label.end.L + xdel
clade.H.x <- label.end.H + xdel
dy <-.33
lines(rep(clade.L.x,2),c(3-dy,7+dy),lwd=lwd.clade,col=col.clade)
lines(rep(clade.H.x,2),c(1-dy,2+dy),lwd=lwd.clade,col=col.clade)
text(clade.L.x+clade.dx,5.0,'L-clade',srt=90,font=font.clade,cex=cex.clade,col=col.clade)
text(clade.H.x+clade.dx,1.5,'H-clade',srt=90,font=font.clade,cex=cex.clade,col=col.clade)

####
#
# LEGEND
#
# parameter plusx controls whether we try to annotate b/c (+) and d/e (x) sharing in tree; I think
# it looks cluttered, rather than adding clarity, so I vote no, but code is here, in case.  "Logic,"
# if any, for my symbol choice is that + overlaid on x looks like the * at the next level; this
# analogy is more visible if we use pch 3/4/8 rather than Courier or Helvetica chars, but probably
# should use same in both tree & legend, which will take a modicum of additional work.
legend.text <- c('a: only in 1014  ',
                 'b: only in 1335  ',
                 'c: only in 1015  ',
                 'd: only in 1012  ',
                 'e: only in 1007  ',
                 '*: shared by bcde',
                 paste(ifelse(plusx,'+:','  '),'shared by b/c '),
                 paste(ifelse(plusx,'x:','  '),'shared by d/e ')
)

legend.text <- c('a: only in 1014 ',
                 'b: only in 1335 ',
                 'c: only in 1015 ',
                 'd: only in 1012 ',
                 'e: only in 1007 ',
                 '*: in bcde      ',
                 paste(ifelse(plusx,'+:','  '),'in bc         '),
                 paste(ifelse(plusx,'x:','  '),'in de         '),
                 'Discordant SNPs '
)
legend.text <- paste(legend.text,format(c(leg.counts,discord),width=4),sep=' - ')
legend.text <- paste(legend.text,' ') # add a little more right margin in box
opar <- par(family='mono',cex=cex.legend)
legend('topright', legend=legend.text, cex=cex.legend, inset=c(0.05,0), bg=col.legbox, box.col=col.legbox)
par(opar)
if(plusx){
  points(tree.labels[[16]],tree.labels[[17]]+.14,pch=8,col=col.elabel)
  points(tree.labels[[16]]+200,tree.labels[[17]]+1,pch=3,col=col.elabel)
  points(tree.labels[[16]]+200,tree.labels[[17]]-1,pch=4,col=col.elabel)
}

####
#
# EDGE LENGTHS
#
for(i in seq(1,length(tree.labels)-ifelse(plusx,5,2),by=3)){
  if(F){ # T for \n in edge labels; F to remove (except "by 5")
    text(tree.labels[[i]], tree.labels[[i+1]], tree.labels[[i+2]])
  } else {
    # points(tree.labels[[i]], tree.labels[[i+1]], pch=3,col='green') # for debugging
    text(tree.labels[[i]], tree.labels[[i+1]], sub('\n([^z])',' \\1', tree.labels[[i+2]]),
           pos=3, offset=.4, font=font.elabel, col=col.elabel,cex=cex.elabel)
    }
  }
}
if(FALSE){#for debug convenience
  pdf(paperfig.path, width=8,height=5,onefile=TRUE,family='Helvetica',fonts='Courier',pointsize=10)
  show.tree(newick.medium, total.snps=consistent.count[2], pltdebug=F,straight.arrow=T)
  dev.off()
}
```

```
caption <- function(stringency,which.tables=which.snp.tables(string.val=F)){
  caption.where <- '(UNKNOWN genome subset).'
  if(which.tables[1]=='Chr1') {caption.where <- 'on Chr1.'}
  if(which.tables[1]=='full') {caption.where <- 'genome-wide.'}
  if(which.tables[1]=='trunc'){caption.where <- 'all Chrs.'}
  cap.stringency <- c(
    'loose SNP filters.',
    'medium SNP filters.',
    'strict SNP filters.',
    'unfiltered SNPs.')
  cap <- paste('Tree based on', which.tables[2], 'reads and', cap.stringency[stringency],
               ' ``Lengths\'\' are numbers of shared/private SNPs', caption.where)
  return(cap)
}
```

Trees based on all four SNP filtering criteria are shown below. Their topologies are exactly the same, although the branch lengths are different. In all four, the length of the branch labeled "*" is probably inflated by lower coverage and higher error rate in 1014, which may mask further legitimate sharing between it and the other L-isolates. The branch lengths among the other 4 are too short for their topology to be convincing without a more rigorous analysis (e.g., a bootstrap test), but detail there is irrelevant to the story.

My sense is that the "medium" version is the best for the paper, made here and shown in Fig 1. In theory, this should look exactly like Fig 3, but something is apparently different between Knitr and direct-to-pdf. (Increasing fig.width in Knitr's chunk headers from 8 (as in the pdf call below) to 9 helps somewhat, but probably still best to make the paper fig directly rather than via Knitr.)

```
###
#
# MAKE PROTOTYPE PDF FOR PAPER, *AND* SAVE DATA NEEDED TO BUILD IT
#
w.s.t. <- which.snp.tables()
if(w.s.t. == 'trunc-unfiltered'){
  rda.Description <- 'This .rda contains data to generate Fig 3; see shared.snps.rnw for details.'
  save(rda.Description, w.s.t., pat.summaries, consistent.count, file='Fig3-data.rda')
  paperfig.path <- paste('figs-mine/paperfig-medium-tree-', w.s.t., '--Fig3proto.pdf', sep='')
} else {
  paperfig.path <- paste('figs-mine/paperfig-medium-tree-', w.s.t., '.pdf', sep='')
}
pdf(paperfig.path, width=8,height=5,onefile=TRUE,family='Helvetica',fonts='Courier',pointsize=10)
newick.medium <- newickize(make.tree(pat.summaries[,'count2']))
show.tree(newick.medium, total.snps=consistent.count[2], pltdebug=F,straight.arrow=T)
dev.off()

# pdf
#   2
```

```
# fig.paths for knitr chunks below;  .h for "hand-made" trees; plain for automatic chr1/full versions
myfigpath   <- paste(getwd(), '/figs-knitr/newick-', which.snp.tables(), '-', sep='')
myfigpath.h <- paste(getwd(), '/figs-knitr/newick-', sep='')
```

Figure 2, i.e., criteria [[1]]:

```
newick.loose <- newickize(make.tree(pat.summaries[,'count1']))
show.tree(newick.loose, total.snps=consistent.count[1])
```

Figure 3, i.e. [[2]]:

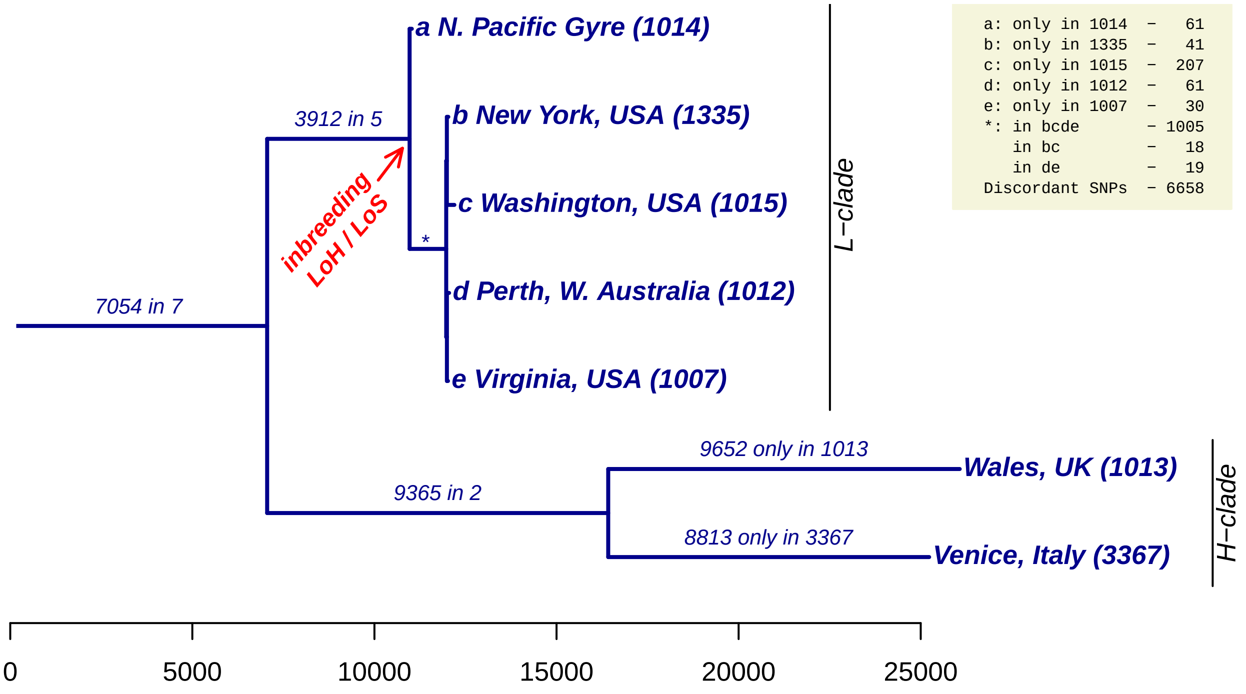

Figure 1: Proposed fig. for paper: Tree based on unfiltered reads and medium SNP filters. “Lengths” are numbers of shared/private SNPs on Chr1.

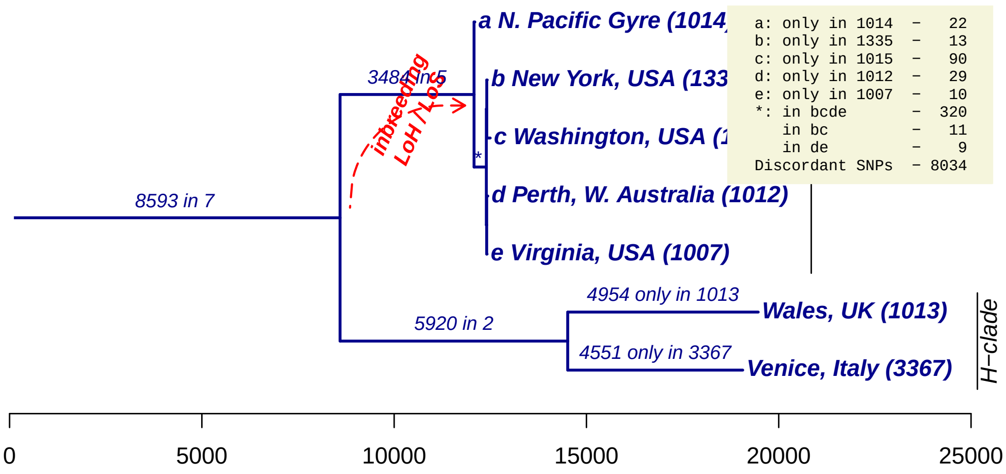

Figure 2: Tree based on unfiltered reads and loose SNP filters. "Lengths" are numbers of shared/private SNPs on Chr1.

```
# newick.medium <- newickize(tree.by.hand)
# simple.newick.medium <- newickize(tree.by.hand,alt=TRUE)
newick.medium <- newickize(make.tree(pat.summaries[,'count2']))
simple.newick.medium <- newickize(make.tree(pat.summaries[,'count2']),alt=TRUE)
show.tree(newick.medium, total.snps=consistent.count[2])
```

Figure 4, i.e. [[3]]:

```
newick.strict <- newickize(make.tree(pat.summaries[,'count3']))
show.tree(newick.strict, total.snps=consistent.count[3])
```

Figure 5, i.e. [[4]]:

```
newick.unfiltered <- newickize(make.tree(pat.summaries[,'count4']))
show.tree(newick.unfiltered, total.snps=consistent.count[4])
```

Some other versions of the trees are included in the appendix.

Counts for all tree edges in the medium tree:

```
#pat.summaries[c(128,110,102,6,97,19,9,2,5,33,65,17,3),]
tree.edges <- c(128,110,102,6,97,19,9,2,5,33,65,17,3)-1
non.edges <- setdiff(0:127, tree.edges)
sg.edges <- showgroup(restrict.to=tree.edges) ; sg.edges

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 2     001     1                                   X     13     41     70    135
# 3     002     1                              X        4551   8813   9159  10949
# 5     004     1                         X               90    207    243    385
# 6     005     2                         X           X     11     18     35     52
# 9     010     1                    X                    22     61     78    113
# 17    020     1               X                       4954   9652   9985  11677
# 19    022     2               X              X        5920   9365   9094   6177
# 33    040     1          X                              29     61     95    174
# 65    100     1     X                                   10     30     40    141
# 97    140     2     X    X                               9     19     34     85
# 102   145     4     X    X              X           X    320   1005   1973   2585
# 110   155     5     X    X         X    X           X   3484   3912   2642   3228
# 128   177     7     X    X    X    X    X    X      X   8593   7054   4790   1641
# Total                                                  28006  40238  38238  37342
```

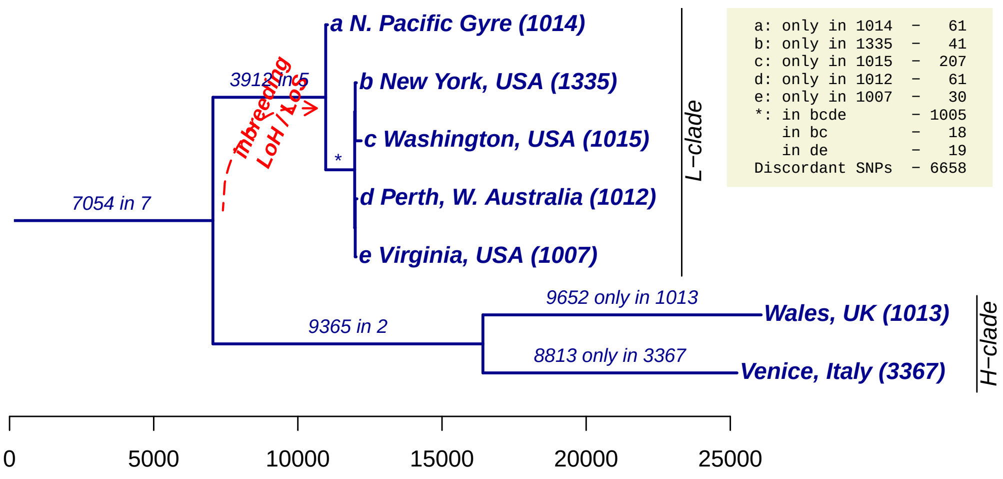

Figure 3: Tree based on unfiltered reads and medium SNP filters. "Lengths" are numbers of shared/private SNPs on Chr1.

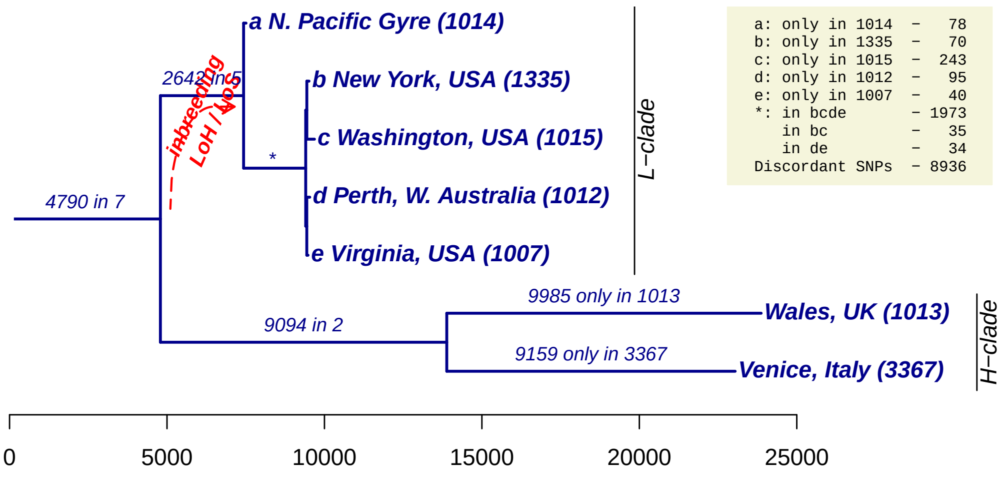

Figure 4: Tree based on unfiltered reads and strict SNP filters. "Lengths" are numbers of shared/private SNPs on Chr1.

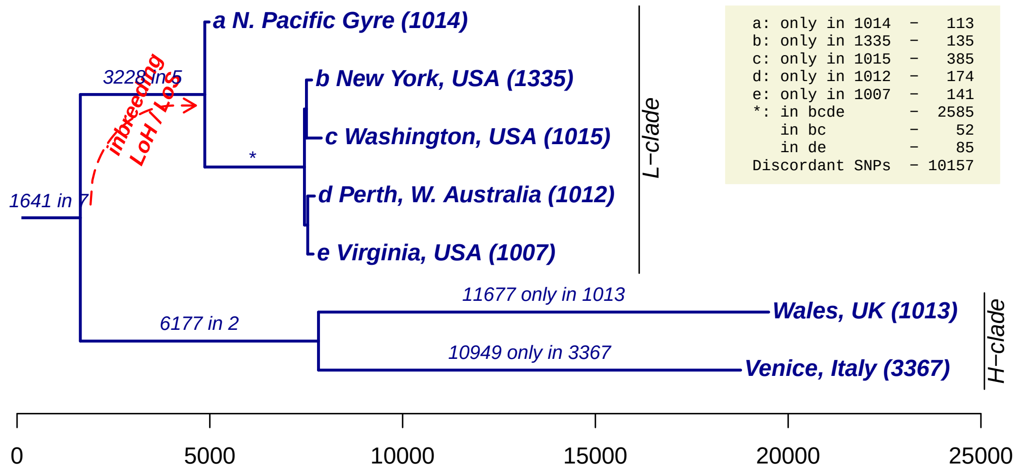

Figure 5: Tree based on unfiltered reads and unfiltered SNPs. "Lengths" are numbers of shared/private SNPs on Chr1.

Counts for the top 10 discordant patterns, i.e., SNPs whose sharing pattern does not match any of the bifurcations in the tree:

```
tenth <- sort(showgroup(restrict.to=non.edges)[-(length(non.edges)+1),'count2'],decreasing=T)[10]
sg.non.edges <- showgroup(restrict.to=non.edges, c2.thresh = tenth) ; sg.non.edges

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 56    067     5         X    X         X    X    X     33     78    201    104
# 101   144     3    X    X              X                6     75    167    355
# 104   147     5    X    X              X    X    X    125    307    590   1160
# 109   154     4    X    X         X    X               16     53     49    103
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 118   165     5    X    X    X         X         X    201    453    787   1140
# 119   166     5    X    X    X         X    X          30    157    247    254
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 127   176     6    X    X    X    X    X    X          45    114     77     68
# Other   (   105 rows   w/   c2    <   53    )        3528    857   1662   2362
# Total                                                8034   6658   8936  10157
```

And percent of discordant SNPs:

```
nsge <- nrow(sg.edges)
discordv <- consistent.count - sg.edges[nsge,c('count1','count2','count3','count4')] ; discordv

#       count1 count2 count3 count4
# Total   8034   6658   8936  10157

discordv.pct <- round(discordv/consistent.count*100,1) ; discordv.pct

#       count1 count2 count3 count4
# Total   22.3   14.2   18.9   21.4
```

In short, the sharing pattern observed at 6658 or 14.2% of the 46896 medium-stringency consistent SNPs positions observed across all 7 isolates are discordant with the medium tree. (The strict tree has slightly more.)

A majority of the discordant SNPs fall into one of three patterns: 6-way sharing excluding Gyre (likely a technical artifact since the low coverage in Gyre reduces our power to detect SNPs there), or 6-way sharing excluding one of the two H-isolates (likely a reflection of sexuality in the H-clade—SNP positions in a population in Hardy-Weinberg equilibrium are fairly likely to be homozygous for the reference allele in a given individual).

```
third.biggest <- sort(showgroup(pat.summaries,6)[-8,'count2'],decreasing=T)[3]
big.three <- showgroup(pat.summaries,6,c2.thresh = third.biggest); big.three

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# Other   (     4 rows   w/   c2    < 1192    )          116    177    156    111
# Total                                                4166   4741   5312   4722

big.three.frac <- sum(big.three[1:3,'count2'])/discordv$count2; big.three.frac

# [1] 0.6854911
```

I.e., 68.5% of discordant SNPs fall into one of these three categories.

Out of curiousity: what is the ratio of full genome to Chr 1 branch lengths. Except for the shortest few, generally ≈10x, as expected given the length of Chr 1:

```
# (vectors derived by editing Newick strings, and in that order)
print(
  c(Italy=86155, Wales=95697, IW=89598, Virg=330,     Aust=632,      VA=1296,
    Puget=2113,  NY=658,      PNY=480,  four=10059,   Gyre=568,      five=39517, all=69526) /
  c(Italy=8813,  Wales=9652,  IW=9365,  Virg=30,      Aust=61,       VA=19,
    Puget=207,   NY=41,       PNY=18,   four=1005,    Gyre=61,       five=3912,  all= 7054),
  digits=3)

# Italy Wales    IW  Virg  Aust    VA Puget    NY   PNY  four  Gyre  five   all
#  9.78  9.91  9.57 11.00 10.36 68.21 10.21 16.05 26.67 10.01  9.31 10.10  9.86

round(genome.length.constants()$genome.length.trunc / genome.length.constants()$chr1.length, digits=4)

# [1] 10.2879
```

9 Semi-Automated Tree-Building

Slightly formalizing the process above: Look for the bifurcation of the 7 strains that maximizes the number of shared SNPs *within* each side of the partition while minimizing the number and fraction of SNPs that are shared by subsets that include at least one strain on each side of the partition. The 2/5 split is the winner, with 6418 SNPs in confict with that partition (16% of the 39842 SNPs not shared by all 7; Chr1 data). The runner-up places the Gyre in a group by itself (7079 = 18% in conflict).

```
treepart <- function(p.summ=pat.summaries, root=127, verbose=T, stringency='count2'){
  root.shared <- p.summ[root+1,stringency]
  df<-NULL
  for(i in 1:floor(root/2)){
    if(bitwAnd(i,root)==i && i < root-i){
      l1 <- showgroup(p.summ,subset=i,split=NULL,proper.subset=F,total=T)
      l  <- l1[nrow(l1),stringency]
      r1 <- showgroup(p.summ,subset=root-i,split=NULL,proper.subset=F,total=T)
      r  <- r1[nrow(r1),stringency]
      c1 <- showgroup(p.summ,subset=root,split=i,proper.subset=T,total=T)
      c  <- c1[nrow(c1),stringency]
      df <- rbind(df, data.frame(pat=i,left=l,right=r,both=l+r,cross=c,all=l+r+c,ratio=c/(l+r+c),
                                 best='',stringsAsFactors=F))
    }
  }
  df$pat<-as.octmode(df$pat)
  maxl <- which.max(df$left)
  maxr <- which.max(df$right)
  maxb <- which.max(df$both)
  minc <- which.min(df$cross)
  minr <- which.min(df$ratio)
  df$best[c(maxl,maxr,maxb,minc,minr)] <- '<'
  df$best[maxl] <- paste(df$best[maxl], 'L') # max Left
  df$best[maxr] <- paste(df$best[maxr], 'R') # max Right
  df$best[maxb] <- paste(df$best[maxb], 'B') # max Both (L+R)
  df$best[minc] <- paste(df$best[minc], 'C') # min Cross
  df$best[minr] <- paste(df$best[minr], 'O') # min ratiO (Cross/(Left+Right+Cross)
  if(verbose){
    same <- all(maxl==c(maxr,maxb,minc,minr))
    cat('root:',       format(as.octmode(root),width=3),
        '; shared:',   root.shared,
        '.  max l',    format(as.octmode(df$pat[maxl]),width=3),
        ', max r',     format(as.octmode(df$pat[maxr]),width=3),
        ', max both',  format(as.octmode(df$pat[maxb]),width=3),
        ', min cross', format(as.octmode(df$pat[minc]),width=3),
        ', min ratio', format(as.octmode(df$pat[minr]),width=3),
        '. \nAll the same?:',same,
        '\n')
    cat('\n')
  }
  return(df)
}
```

```
treepart()

# root: 177 ; shared: 7054 .  max l 077 , max r 010 , max both 022 , min cross 022 , min ratio 022 .
# All the same?: FALSE
#    pat   left right  both cross   all     ratio    best
# 1   01     41 29077 29118 10724 39842 0.2691632
# 2   02   8813 17400 26213 13629 39842 0.3420762
# 3   03   8870 10338 19208 20634 39842 0.5178957
# 4   04    207 28345 28552 11290 39842 0.2833693
# 5   05    266 28142 28408 11434 39842 0.2869836
# 6   06   9033  9956 18989 20853 39842 0.5233924
# 7   07   9110  9866 18976 20866 39842 0.5237187
# 8   10     61 32702 32763  7079 39842 0.1776768     < R
# 9   11    109 28694 28803 11039 39842 0.2770694
# 10  12   8892 11759 20651 19191 39842 0.4816776
# 11  13   8959 10160 19119 20723 39842 0.5201295
# 12  14    270 28163 28433 11409 39842 0.2863561
# 13  15    342 28006 28348 11494 39842 0.2884895
# 14  16   9115  9849 18964 20878 39842 0.5240199
# 15  17   9213  9781 18994 20848 39842 0.5232669
# 16  20   9652 16099 25751 14091 39842 0.3536720
# 17  21   9702  9470 19172 20670 39842 0.5187993
# 18  22  27830  5594 33424  6418 39842 0.1610863 < B C O
# 19  23  27916   542 28458 11384 39842 0.2857286
# 20  24   9871  9119 18990 20852 39842 0.5233673
# 21  25   9943  9016 18959 20883 39842 0.5241454
# 22  26  28089   239 28328 11514 39842 0.2889915
# 23  27  28213   175 28388 11454 39842 0.2874856
# 24  30   9731 10772 20503 19339 39842 0.4853923
# 25  31   9790  9303 19093 20749 39842 0.5207821
# 26  32  27945  1520 29465 10377 39842 0.2604538
# 27  33  28045   414 28459 11383 39842 0.2857035
# 28  34   9955  9014 18969 20873 39842 0.5238944
# 29  35  10044  8931 18975 20867 39842 0.5237438
# 30  36  28214   162 28376 11466 39842 0.2877868
# 31  37  28375   110 28485 11357 39842 0.2850510
# 32  40     61 28554 28615 11227 39842 0.2817881
# 33  41    106 28330 28436 11406 39842 0.2862808
# 34  42   8875 10122 18997 20845 39842 0.5231916
# 35  43   8944 10017 18961 20881 39842 0.5240952
# 36  44    281 28125 28406 11436 39842 0.2870338
# 37  45    370 28005 28375 11467 39842 0.2878119
# 38  46   9120  9835 18955 20887 39842 0.5242458
# 39  47   9253  9773 19026 20816 39842 0.5224637
# 40  50   124 28358 28482 11360 39842 0.2851262
# 41  51   177 28189 28366 11476 39842 0.2880377
# 42  52  8956 10008 18964 20878 39842 0.5240199
# 43  53  9036  9926 18962 20880 39842 0.5240701
# 44  54   352 27986 28338 11504 39842 0.2887405
# 45  55   467 27885 28352 11490 39842 0.2883891
# 46  56  9212  9743 18955 20887 39842 0.5242458
# 47  57  9386  9691 19077 20765 39842 0.5211837
# 48  60  9717  9297 19014 20828 39842 0.5227649
# 49  61  9773  9175 18948 20894 39842 0.5244215
# 50  62 27913   396 28309 11533 39842 0.2894684
# 51  63 28024   313 28337 11505 39842 0.2887656
# 52  64  9958  9006 18964 20878 39842 0.5240199
# 53  65 10079  8931 19010 20832 39842 0.5228653
# 54  66 28243   143 28386 11456 39842 0.2875358
# 55  67 28531    93 28624 11218 39842 0.2815622
# 56  70  9799  9180 18979 20863 39842 0.5236434
# 57  71  9866  9087 18953 20889 39842 0.5242960
# 58  72 28037   312 28349 11493 39842 0.2884644
# 59  73 28167   246 28413 11429 39842 0.2868581
# 60  74 10054  8912 18966 20876 39842 0.5239697
# 61  75 10217  8849 19066 20776 39842 0.5214598
# 62  76 28399    73 28472 11370 39842 0.2853772
# 63  77 28807    30 28837 11005 39842 0.2762161     < L
```

Comparing the 5/2 split to the second-place NPG/rest split (below), the former has fewer pattern instances in conflict with the split (6418 vs 7079), as well as somewhat more random distribution of the conflicting patterns (92 vs 62 rows), whereas the 1/6 split has the majority of its conflicts (3912 of 7079, or 55%) concentrated in one pattern—the 5 NE strains. Collectively, these seem to favor the 5/2 split as the correct "history."

```
showgroup(pat.summaries,split=strtoi('022'), subset=127, proper.subset=T, c2.thresh=100)

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 104   147     5    X    X              X    X    X    125    307    590   1160
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 118   165     5    X    X    X         X         X    201    453    787   1140
# 119   166     5    X    X    X         X    X          30    157    247    254
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 127   176     6    X    X    X    X    X    X          45    114     77     68
# Other   (    85 rows   w/   c2    <  100    )        3493    823   1387   1771
# Total                                                7944   6418   8244   9004

showgroup(pat.summaries,split=strtoi('010'), subset=127, proper.subset=T, c2.thresh=100)

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 110   155     5    X    X         X    X         X   3484   3912   2642   3228
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 127   176     6    X    X    X    X    X    X          45    114     77     68
# Other   (    58 rows   w/   c2    <  100    )        1095    368    368    522
# Total                                                7723   7079   4923   6577
```

Below is the full summary of shared SNPs that do *not* directly correspond to tree splits, e.g. deep coalescence, independent coincident mutations, false positives/false negatives in the shared SNP calls, loss of SNPs in hemizygous regions, etc. (Additionally, SAMTools' SNP calls exclude positions it judges to be homozygous, and I think it operates without regard to the reference sequence, so homozygous nonreference positions, while rare except in IT/Wales, often are not called SNPs by SAMTools, but are relevant for this analysis. Provided the position is called a SNP in some other isolate, the consistency filtering we've done above should recover it, but this is still worth keeping in mind when examining the data.)

First, here are SNPs that "coalesce" on the branch from the LCA of bcde, i.e., shared among some nonempty, proper subset of bcde other than bc or de. There are 8 such patterns: any of the 4 choose 3 trios plus any of the 4 pairs having exactly one of bc.

```
sg4 <- showgroup(pat.summaries, subset=strtoi('0145'), split=5, proper.subset = F)
sg4

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 34    041     2         X                        X      3      4     13     36
# 37    044     2         X              X                5     13     75    155
# 38    045     3         X              X         X      6     26    130    131
# 66    101     2    X                             X      1      2     13     20
# 69    104     2    X                   X                5      9     34    107
# 70    105     3    X                   X         X      4      5     25     63
# 98    141     3    X    X                        X      2      5     15     40
# 101   144     3    X    X              X                6     75    167    355
# 102   145     4    X    X              X         X    320   1005   1973   2585
# Total                                                 352   1144   2445   3492

sg4n <- nrow(sg4)
sg4pct <- round(sg4$count2[sg4n-1]/sg4$count2[sg4n]*100,1)
sg4pct

# [1] 87.8
```

So, of the 1144 SNPs found only in bcde, 87.8% have a sharing pattern consistent with the given tree structure.

Similarly, we analyze patterns relative to the root of the L-clade (14 patterns—any nonempty proper subset of bcde together with a):

```
sg5 <- showgroup(pat.summaries,subset=strtoi('0155'), split=8, proper.subset = F)
sg5

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 10    011     2                   X              X      6      7      7     14
# 13    014     2                   X    X               10      2      5     12
# 14    015     3                   X    X         X      7      6      4      7
# 41    050     2         X         X                     7      2      5      7
# 42    051     3         X         X              X      0      1      2      4
# 45    054     3         X         X    X                1      6     17     18
# 46    055     4         X         X    X         X      6     12     29     36
# 73    110     2    X              X                     2      2      4      7
# 74    111     3    X              X              X      1      0      1      1
# 77    114     3    X              X    X                1      2      2      8
# 78    115     4    X              X    X         X      0      4      3      9
# 105   150     3    X    X         X                     0      0      4      6
# 106   151     4    X    X         X              X      1      4      5     14
# 109   154     4    X    X         X    X               16     53     49    103
# 110   155     5    X    X         X    X         X   3484   3912   2642   3228
# Total                                                3542   4013   2779   3474

sg5n <- nrow(sg5)
sg5pct <- round(sg5$count2[sg5n-1]/sg5$count2[sg5n]*100,1)
```

I.e., of the 4013 SNPs found only in abcde, 97.5% have a sharing pattern consistent with the given tree structure.

Finally, how many SNPs have patterns inconsistent with the 5-2 split, i.e., include at least one strain on each side of the 5-2 split, but not shared by all 7?

```
sg7 <- showgroup(pat.summaries, subset=127, split=strtoi('022'), proper.subset=F)
sg7

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 4     003     2                             X    X    239     16     28     31
# 7     006     2                        X    X         141     13     21     41
# 8     007     3                        X    X    X      9      2      9     10
# 11    012     2                   X         X         179     18      2      5
# 12    013     3                   X         X    X     17      3      2      2
# 15    016     3                   X    X    X           9      1      2      2
# 16    017     4                   X    X    X    X      1      5      1      2
# 18    021     2              X                   X    243      9     11      9
# 20    023     3              X              X    X    327     20     29     17
# 21    024     2              X         X              125     12     28     46
# 22    025     3              X         X         X      9      4     12     21
# 23    026     3              X         X    X         185     27     31     32
# 24    027     4              X         X    X    X     22     14     28     24
# 25    030     2              X    X                   222     18      5      9
# 26    031     3              X    X              X     20      2      0      0
# 27    032     3              X    X         X         324     18      8      5
# 28    033     4              X    X         X    X     30      2      4      6
# 29    034     3              X    X    X               11      3      1      1
# 30    035     4              X    X    X         X      4      2      1      0
# 31    036     4              X    X    X    X          12      4      1      3
# 32    037     5              X    X    X    X    X     17     10      8      5
# 35    042     2         X                   X         150      1     16     27
# 36    043     3         X                   X    X     21      8     14      6
# 39    046     3         X              X    X          11     12     34     55
# 40    047     4         X              X    X    X      2     18     42     60
# 43    052     3         X         X         X           9      0      2      1
# 44    053     4         X         X         X    X      1      0      2      2
# 47    056     4         X         X    X    X           3      2      2      5
# 48    057     5         X         X    X    X    X      5      7      9     17
# 49    060     2         X    X                        165      4     26     28
# 50    061     3         X    X                   X     12      2     14     12
# 51    062     3         X    X              X         227     17     37     36
# 52    063     4         X    X              X    X     18     11     24     21
# 53    064     3         X    X         X                9      9     36     60
# 54    065     4         X    X         X         X      9     17     41     48
# 55    066     4         X    X         X    X          25     37    102     76
# 56    067     5         X    X         X    X    X     33     78    201    104
# 57    070     3         X    X    X                    13      1      1      2
# 58    071     4         X    X    X              X      2      1      2      0
# 59    072     4         X    X    X         X           8      6      3      2
# 60    073     5         X    X    X         X    X      3      3      6      3
# 61    074     4         X    X    X    X                2      3      2      7
# 62    075     5         X    X    X    X         X      7     11     13     11
# 63    076     5         X    X    X    X    X           9     11     17      7
# 64    077     6         X    X    X    X    X    X     43     48     60     26
# 67    102     2    X                        X          93      6      7     25
# 68    103     3    X                        X    X     11      4      4      8
# 71    106     3    X                   X    X           3      9     15     27
# 72    107     4    X                   X    X    X      5      5     11     16
# 75    112     3    X              X         X           4      1      0      0
# 76    113     4    X              X         X    X      1      2      0      1
# 79    116     4    X              X    X    X           0      1      0      5
# 80    117     5    X              X    X    X    X      0      2      1      7
# 81    120     2    X         X                        105      9      9     31
# 82    121     3    X         X                   X      8      0      4      4
# 83    122     3    X         X              X         134     10     10     26
# 84    123     4    X         X              X    X     10      9      8      8
# 85    124     3    X         X         X               11      7     20     35
# 86    125     4    X         X         X         X      4      3     13     16
# 87    126     4    X         X         X    X          12     20     21     43
# 88    127     5    X         X         X    X    X     11     17     30     47
# 89    130     3    X         X    X                     7      1      2      1
# 90    131     4    X         X    X              X      0      1      0      2
# 91    132     4    X         X    X         X           9      1      2      1
# 92    133     5    X         X    X         X    X      3      2      0      0
# 93    134     4    X         X    X    X                2      2      1      3
# 94    135     5    X         X    X    X         X      2      1      3      7
# 95    136     5    X         X    X    X    X           4      6      3      5
# 96    137     6    X         X    X    X    X    X     11      6     12     14
# 99    142     3    X    X                   X           9      1      9     15
# 100   143     4    X    X                   X    X      1      3      5     20
# 103   146     4    X    X              X    X           7     34     65    140
# 104   147     5    X    X              X    X    X    125    307    590   1160
# 107   152     4    X    X         X         X           2      1      1      4
# 108   153     5    X    X         X         X    X      3      3      1      7
# 111   156     5    X    X         X    X    X           8     15     15     43
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 113   160     3    X    X    X                         10      6      8     23
# 114   161     4    X    X    X                   X      9      5      9     18
# 115   162     4    X    X    X              X          11     12     21     33
# 116   163     5    X    X    X              X    X     12     18     23     33
# 117   164     4    X    X    X         X                8     47     80    163
# 118   165     5    X    X    X         X         X    201    453    787   1140
# 119   166     5    X    X    X         X    X          30    157    247    254
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 121   170     4    X    X    X    X                     1      0      1      3
# 122   171     5    X    X    X    X              X      3      6      2      7
# 123   172     5    X    X    X    X         X           2      6      3      5
# 124   173     6    X    X    X    X         X    X     17      9      7      3
# 125   174     5    X    X    X    X    X                7     22     23     35
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 127   176     6    X    X    X    X    X    X          45    114     77     68
# 128   177     7    X    X    X    X    X    X    X   8593   7054   4790   1641
# Total                                               16537  13472  13034  10645

sg7n <- nrow(sg7)
sg7pct <- round(sg7$count2[sg7n-1]/sg7$count2[sg7n]*100,1)
sg7pct

# [1] 52.4
```

A more compact version of that table, showing only the larger counts:

```
thresh <- signif(.02 * sg7$count2[sg7n],1)
thresh

# [1] 300

showgroup(pat.summaries, subset=127, split=strtoi('022'), proper.subset=F, c2.thresh = thresh)

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 104   147     5    X    X              X    X    X    125    307    590   1160
# 112   157     6    X    X         X    X    X    X   1343   1192    839   1343
# 118   165     5    X    X    X         X         X    201    453    787   1140
# 120   167     6    X    X    X         X    X    X    951   1879   3320   1852
# 126   175     6    X    X    X    X    X         X   1756   1493    997   1416
# 128   177     7    X    X    X    X    X    X    X   8593   7054   4790   1641
# Other   (    87 rows   w/   c2    <  300    )        3568   1094   1711   2093
# Total                                               16537  13472  13034  10645
```

So, of the 13472 SNPs found both in the L- and H-clades, 52.4% have a sharing pattern consistent with the given tree structure, i.e., are found in all 7 isolates. Among the others, three patterns dominate—(i) the 6-way pattern excluding the Gyre is the largest, plausibly explained by 7-way sharing from which the Gyre drops out due to low coverage/high error rate, (ii) the 6-way excluding Italy, and (iii) ditto for Wales. Origin of the later two cases is unclear, but may partly reflect Hardy-Weinberg—some positions that are *population-level* SNPs in those isolates will be homozygous-reference in the CCMP founder cell for IT or Wales. If I take the 7-way shared SNP count (69526) as a surrogate approximating the number of population-level SNPs in either IT or Wales that are shared with the L-clade, then I might expect, based on HWE, roughly half that number to to be lost (become homozygous) in IT, and a similar number in Wales. However, the observed counts of these positions are lower by $\approx$20K than I might have guessed from HWE, perhaps suggesting that IT and Wales are distinct populations, each with a pool of many thousand private polymorphisms.

In aggregate:

```
untreelike <-
  sg7$count2[sg7n]-sg7$count2[sg7n-1] +
  sg5$count2[sg5n]-sg5$count2[sg5n-1] +
  sg4$count2[sg4n]-sg4$count2[sg4n-1]
untreelike
```

```
# [1] 6658

consistent.count[2]

# [1] 46896

unpct <- round(untreelike/consistent.count[2]*100,1)
unpct

# [1] 14.2
```

I.e., 6658 or 14.2% of the 46896 consistent SNPs identified (by criterion 2) across all 7 isolates are discordant with the assumed tree.

Overall, based on this data, I take the following to be obvious: (a) separation of the the H-isolates from the L-isolates (and from each other??), and (b) near-identity of the L-isolates. Due to the small counts, the exact topology among the L-isolates (esp. bcde) is uncertain, but *any* topology there is consistent with the asexual/clonal/global-expansion hypothesis, so there is little point in examining this subtree more careful. Again, we believe the (apparent) slight separation of the Gyre from the other L-isolates is largely driven by technical artifacts (lower coverage/higher error rates) in the sequencing rather than by biological effects. However, the discord between Gyre SNPs and others is the major substantive ambiguity in the offered tree. Nevertheless, in the next section we show by a bootstrap analysis that the offered placement of Gyre with respect to the other 4 L-isolates is strongly supported by the data.

9.1 Bootstrap

How robust is the inferred tree? Italy/Wales seem clearly related to each other but separate from the other 5. Likewise, the 4 coastal L-isolates seem to be closely related, with little data to separate them (and perhaps little sense in trying). So, the key question here is whether the top level bifurcation is 2/5 or NPG/6. Here, we do a simple bootstrap test (on c2 numbers only) to see whether the 2/5 split is consistently the most parsimonious.

```
n2 <- sum(pattern.counts[[2]][,2]); n2

# [1] 46896
```

Conceptually, we sample, with replacement, n2=46896 SNP positions from among the 46896 positions declared consisent SNPs according to criterion c2, and recalculate the statistics examined above to see whether the 2/5 split again minimizes conflicting sharing patterns. This resampling/calculation is repeated nboot times (set near front of file). Since all that matters is the sharing pattern, this procedure is expedited by actually sampling 46896 independent integers in the range 0:127 with probabilities proportional to the patttern counts given in column 2 of pattern.counts[[2]]. The sample is then tabulated in a 128 row table analogous to pattern.summaries, for analysis by showgroups/treepart, as above.

```
boot.sample <- sample(0:127,n2,replace=T,prob=pattern.counts[[2]][,2])
str(boot.sample)

#  int [1:46896] 16 119 16 2 18 18 2 28 2 109 ...

boot.count <- mytable(boot.sample,c(0,127))
boot.count[c(1:4,125:128),] # show a few rows

#      val count
# [1,]   0     0
# [2,]   1    39
# [3,]   2  8780
# [4,]   3    16
# [5,] 124    15
# [6,] 125  1469
# [7,] 126   118
# [8,] 127  6999

boot.counts <- list(NULL,boot.count,NULL) # dummy list with just c2 summaries
cor(pattern.counts[[2]][,2],boot.counts[[2]][,2]) # just curious - how correlated are they?
```

```
# [1] 0.9999417

boot.summaries <- pat.summary(boot.counts)
showgroup(boot.summaries,c2.thresh=400) #show a few rows

#       Pat ShrBy 1007 1012 1013 1014 1015 3367 1335 count1 count2 count3 count4
# 3     002     1                              X          NA   8780     NA     NA
# 17    020     1              X                          NA   9533     NA     NA
# 19    022     2              X              X          NA   9461     NA     NA
# 102   145     4    X    X              X         X     NA    965     NA     NA
# 110   155     5    X    X         X    X         X     NA   3961     NA     NA
# 112   157     6    X    X         X    X    X    X     NA   1242     NA     NA
# 118   165     5    X    X    X         X         X     NA    480     NA     NA
# 120   167     6    X    X    X         X    X    X     NA   1900     NA     NA
# 126   175     6    X    X    X    X    X         X     NA   1469     NA     NA
# 128   177     7    X    X    X    X    X    X    X     NA   6999     NA     NA
# Other   (   118 rows   w/   c2    <  400    )          NA   2106     NA     NA
# Total                                                  NA  46896     NA     NA
```

Tree partition analysis (and how to pluck out only the best rows based on 3 smallest cross counts and "best" criteria):

```
tp <- treepart(boot.summaries,root=127) ; tp

# root: 177 ; shared: 6999 .  max l 077 , max r 010 , max both 022 , min cross 022 , min ratio 022 .
# All the same?: FALSE
#    pat  left right  both cross   all     ratio    best
# 1   01    39 29036 29075 10822 39897 0.2712485
# 2   02  8780 17278 26058 13839 39897 0.3468682
# 3   03  8835 10204 19039 20858 39897 0.5227962
# 4   04   209 28302 28511 11386 39897 0.2853849
# 5   05   273 28117 28390 11507 39897 0.2884177
# 6   06  9000  9830 18830 21067 39897 0.5280347
# 7   07  9083  9753 18836 21061 39897 0.5278843
# 8   10    61 32669 32730  7167 39897 0.1796376     < R
# 9   11   106 28634 28740 11157 39897 0.2796451
# 10  12  8862 11614 20476 19421 39897 0.4867785
# 11  13  8926 10032 18958 20939 39897 0.5248264
# 12  14   270 28102 28372 11525 39897 0.2888688
# 13  15   347 27959 28306 11591 39897 0.2905231
# 14  16  9083  9719 18802 21095 39897 0.5287365
# 15  17  9187  9665 18852 21045 39897 0.5274833
# 16  20  9533 16175 25708 14189 39897 0.3556408
# 17  21  9578  9450 19028 20869 39897 0.5230719
# 18  22 27774  5617 33391  6506 39897 0.1630699 < B C O
# 19  23 27858   549 28407 11490 39897 0.2879916
# 20  24  9756  9092 18848 21049 39897 0.5275835
# 21  25  9828  9000 18828 21069 39897 0.5280848
# 22  26 28038   244 28282 11615 39897 0.2911246
# 23  27 28171   189 28360 11537 39897 0.2891696
# 24  30  9612 10751 20363 19534 39897 0.4896107
# 25  31  9664  9286 18950 20947 39897 0.5250269
# 26  32 27901  1493 29394 10503 39897 0.2632529
# 27  33 27997   422 28419 11478 39897 0.2876908
# 28  34  9843  8984 18827 21070 39897 0.5281099
# 29  35  9929  8908 18837 21060 39897 0.5278592
# 30  36 28178   166 28344 11553 39897 0.2895706
# 31  37 28344   121 28465 11432 39897 0.2865378
# 32  40    66 28519 28585 11312 39897 0.2835301
# 33  41   107 28303 28410 11487 39897 0.2879164
# 34  42  8847 10012 18859 21038 39897 0.5273078
# 35  43  8911  9913 18824 21073 39897 0.5281851
# 36  44   290 28079 28369 11528 39897 0.2889440
# 37  45   386 27962 28348 11549 39897 0.2894704
# 38  46  9099  9709 18808 21089 39897 0.5285861
# 39  47  9239  9655 18894 21003 39897 0.5264306
# 40  50   129 28308 28437 11460 39897 0.2872396
# 41  51   176 28144 28320 11577 39897 0.2901722
# 42  52  8931  9897 18828 21069 39897 0.5280848
# 43  53  9004  9819 18823 21074 39897 0.5282101
# 44  54   359 27924 28283 11614 39897 0.2910996
# 45  55   485 27826 28311 11586 39897 0.2903978
# 46  56  9190  9615 18805 21092 39897 0.5286613
# 47  57  9375  9570 18945 20952 39897 0.5251523
# 48  60  9602  9274 18876 21021 39897 0.5268817
# 49  61  9649  9150 18799 21098 39897 0.5288117
# 50  62 27867   406 28273 11624 39897 0.2913502
# 51  63 27964   321 28285 11612 39897 0.2910495
# 52  64  9845  8973 18818 21079 39897 0.5283355
# 53  65  9969  8904 18873 21024 39897 0.5269569
# 54  66 28206   142 28348 11549 39897 0.2894704
# 55  67 28495    97 28592 11305 39897 0.2833546
# 56  70  9684  9156 18840 21057 39897 0.5277840
# 57  71  9739  9059 18798 21099 39897 0.5288368
# 58  72 28010   324 28334 11563 39897 0.2898213
# 59  73 28123   254 28377 11520 39897 0.2887435
# 60  74  9943  8874 18817 21080 39897 0.5283605
# 61  75 10109  8815 18924 20973 39897 0.5256786
# 62  76 28386    70 28456 11441 39897 0.2867634
# 63  77 28791    31 28822 11075 39897 0.2775898     < L
```

```
otp <- order(tp[,'cross'])[1:3]     # 3 smallest 'cross' counts
btp <- which(tp[,'best'] != '')     # 'best' by Left/Right/Both/Cross/ratiO
toptp <- unique(c(otp,btp,18,8))    # above, plus 5/2, 6/1 splits
print(tp[toptp,])                   # show the winners

#    pat  left right  both cross   all     ratio    best
# 18  22 27774  5617 33391  6506 39897 0.1630699 < B C O
# 8   10    61 32669 32730  7167 39897 0.1796376     < R
# 26  32 27901  1493 29394 10503 39897 0.2632529
# 63  77 28791    31 28822 11075 39897 0.2775898     < L
```

Now repeat the above nboot times, and summarize results:

```
nboot <- params$nboot #  default from params set in section 2
nboot <- ((nboot+2) %/% 4) * 4 + 1  # summary is cleaner if n mod 4 == 1, so int median/quartiles
cat('***\n*** Doing', nboot, 'bootstrap replicates.\n***\n')

# ***
# *** Doing 5 bootstrap replicates.
# ***

bcor <- numeric(nboot)
b52cross <- integer(nboot)
b61cross <- integer(nboot)
brev <- logical(nboot)
for(i in 1:nboot){
  boot.sample <- sample(0:127,n2,replace=T,prob=pattern.counts[[2]][,2])
  boot.count  <- mytable(boot.sample,c(0,127))
  boot.counts <- list(NULL,boot.count,NULL) # dummy list with just c2 summaries
  boot.summaries <- pat.summary(boot.counts)
  tp <- treepart(boot.summaries,root=127, verbose=F)
  bcor[i] <- cor(pattern.counts[[2]][,2],boot.counts[[2]][,2]) # just curious - how correlated are they?
  b52cross[i] <- tp[18,'cross']
  b61cross[i] <- tp[ 8,'cross']
  brev[i] <- (b52cross[i] > b61cross[i])
  if(brev[i]){
    # show the unexpected ones; probably breaks w/ cache
    otp <- order(tp[,'cross'])[1:3]
    btp <- which(tp[,'best'] != '')
    toptp <- unique(c(otp,btp,18,8))
    print(tp[toptp,])
  }
}
# summarize:
corsummary <- t(as.matrix(c(summary(bcor),sd=sd(bcor))))
row.names(corsummary) <- 'bcor'
bdelta <- b61cross-b52cross
brevp <- 100*brev   # make it percent reversed instead of logical
thesummary <- rbind(summary(b52cross),summary(b61cross),summary(c(bdelta)),summary(brevp))
row.names(thesummary) <- c('b52cross', 'b61cross', 'b61-b52', '% rev')
thesummary <- cbind(thesummary, sd=c(sd(b52cross),sd(b61cross),sd(bdelta),sd(brevp)))
```

SUMMARY: In 5 bootstrap replicates, we saw 0 samples with the 6/1 split having fewer conflicts than the 5/2 split, and the minimum separation between them was ≈ 7.7 sigma, hence highly statistically significant.

```
# 'opt' hacking is trying to force knitr to show more digits of bcor in summary, as Rstudio does, but
# it still fails...  Bottom line is the correlation seems to be  > .999 in all samples, rounds to 1.0,
# as seen in this run of 1001 samples cut/paste from Rstudio:
#          Min.        1st Qu.     Median      Mean        3rd Qu. Max.   sd
# bcor     "   0.9998" "   0.9999" "   0.9999" "   0.9999" "   1"  "   1" "   0.00003462"
# > max(bcor)
# [1] 0.9999915
o.opts <- options(digits=7,width=127)
format(rbind(corsummary,thesummary),scientific=F,digits=4,drop0trailing=T)

#          Min.            1st Qu.         Median          Mean            3rd Qu.         Max.            sd
# bcor     "   0.99987057" "   0.99990268" "   0.99993947" "   0.99992174" "   0.99993965" "   0.99995633" "   0.
# b52cross "6332"          "6399"          "6437"          "6428.6"        "6485"          "6490"          "  65.
# b61cross "7002"          "7030"          "7099"          "7077.4"        "7113"          "7143"          "  59.
# b61-b52  " 565"          " 628"          " 631"          " 648.8"        " 653"          " 767"          "  73.
# % rev    "   0"          "   0"          "   0"          "   0"          "   0"          "   0"          "   0"

options(o.opts)
```

Based on this, it is reasonable to claim that we are confident that the tree topology is as shown in the earlier figures, with the exception of the exact order of the splits with the 4 NE coastal isolates.

10 Notes

This section is a random brain dump of limitations of the current analysis, ideas for improvements, etc. In the main, these may not be worth doing, unless we see significant holes or get pushed by reviewers, etc, but I wanted to catalog before we forget them.
Noise: Various sources of “noise” in the data:

1. Read errors, low read depth — perhaps fixed by medium/strict thresholding

2. Deep coalescence

3. Skew because 1335 is the reference. (Julie notes we could partially fix this by remapping based on discovered SNPs, tho that wouldn’t fix gross misassembly in 1335, e.g. collapsed or misordered tandem duplicates, or segments missing in 1335 that are present in one or more other strains, etc.; much harder to fix those, let’s just hope they are rare...)

4. Varying error rates and sequencing depth among the 7. E.g., plausibly the 1000 SNPs shared by 4 but not by Gyre are a result of lower read depth (we missed a SNP that is actually present) and/or higher error rates (causing a position to appear inconsistent in gyre) in the gyre data. I can’t think of a way to correct for this effect. It might be possible, perhaps by simulation, to estimate the size of the effect and see whether it could explain ≈1000 SNPs.

5. Varying numbers of founder cells in the sequencing cultures. (Again, I made some attempts at modeling this, but nothing very satisfactory yet.)

6. Tri-allelic positions where stochastic fluctuation in sequence sampling promotes the rare allele to prominence. (Julie replies: “isn’t this the same as more than one founder cell? If they are diploid there should only ever be two alleles, unless there were random and very rare, thus unlikely, trisomy events?” I agree, but it is a concrete example of an effect of multiple founders that might be important. Not sure this is the most important such effect...)

7. Gaps/indels - alignments are likely to be of lower quality in the vicinity of an indel, so, maybe lower coverage/more SNPs. We ignored them. Does this add any systematic bias? e.g. if one strain had more indels than another, would this confound other analyses? unclear. Julie suggested a paper titled “Barking up the wrong tree-length: yada yada yada gap penalties”; maybe relevant?

Other Items/Potential To Dos:

1. any spacial structure to various sub-classes?
2. after top level split, should I reanalyze halves of partition in isolation? said another way, I think the tree-building is sensible, but not sure it's optimal.
3. if we believe no sex, then I think gain of SNP should be more common than loss of SNP, since the later can only happen by (a) mutation reverting to reference, (b) second mutation matching nonreference, (c) homologous repair (look for blocks of LOH), or (d) false negative e.g. from low read depth. Does tree-building appropriately weight the gain vs loss cases? (Does it even care?)
4. should we weight coding and/or nonsynonomous SNPs more heavily? Julie says "you do not want to weight the coding or nonsynonomous/coding SNPs because for time you want the more clock-like neutral mutations." I.e., I got this backwards. Maybe should redo tree based on noncoding SNPs only.
5. We could also do an actual parsimony analysis based on 2-state model (homozygous-ref vs not), but I'm not quite sure how to handle this in a mixed sex/nosex case.
6. Might be interesting to look at sharing just within (shared?) deserts. Given tree model above and expectation that bottleneck followed split of H- from L-clades, I would expect little or no sharing of L-clade desert SNPs with H-clade; sharing between It/Wales might suggest "desert" is actually a region under strong purifying selection (e.g. a gene); sharing/non-sharing within L-clade deserts might suggest more about evo history of the 5.

11 Appendix: Old Trees, etc.

Tangents, old stuff of historical interest at best, etc..

11.1 HWE Sharing

Tangent: As a function of nonref allele freq, assuming HWE, what is probability that nonref allele will be seen in k strains, $0 \leq k \leq 4$ (Fig 6).

```
myfigpath.h <- paste(getwd(), '/figs-knitr/', sep='')
```

```
p <- (0:20)/20
q <- 1-p
r <- 2*p*q+p^2
plot(  p, 1*q^0*r^4, type='b',pch='4', ylab="share prob")
points(p, 4*q^2*r^3, type='b',pch='3')
points(p, 6*q^4*r^2, type='b',pch='2')
points(p, 4*q^6*r^1, type='b',pch='1')
points(p, 1*q^8*r^0, type='b',pch='0')
```

11.2 Old Tree Stuff

All based on un-q-filtered reads.

The first pass at the tree analysis was the Chr1 tree, *loose criteria* (c1); it is rendered via `http://iubio.bio.indiana.edu/treeapp/treeprint-form.html` as Fig 7, and in newick format is:

```
newick.chr1.loose <- '(((tp3367_Italy:4551,tp1013_Wales:4954):5920,(((tp1007_Virginia:10,tp1012_Australia:29):9,(
cat.hardwrap(newick.chr1.loose)

# (((tp3367_Italy:4551,tp1013_Wales:4954):5920,(((tp1007_Virginia:10,tp1012_Austra
# lia:29):9,(tp1015_Puget_Sound:90,tp1335_NY:13):11):320,tp1014_Gyre:22):3484):859
# 3,outgroup:0);
```

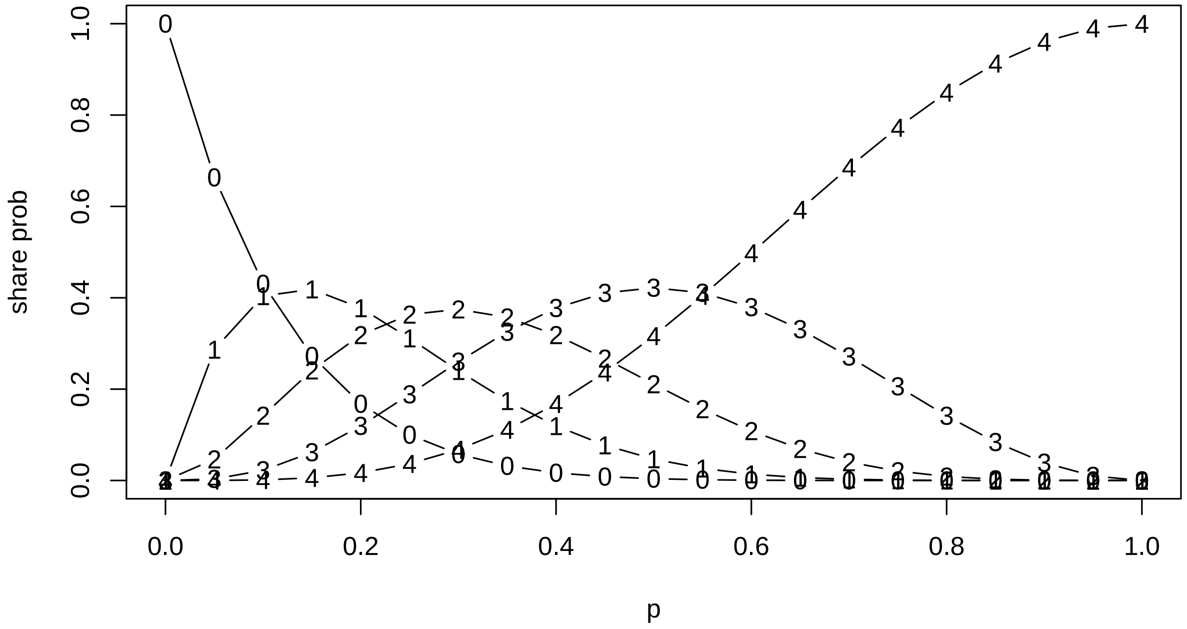

Figure 6: Sharing Probability

Chr 1 tree based on *medium criteria* (c2) has exactly the same topology is, although the branch lengths are different. As noted earlier, the length of the branch labeled “*” is probably inflated by lower coverage and higher error rate in 1014, which may mask further legitimate sharing between it and the other L-isloates. The branch lengths among the other 4 are too short for its topology to be convincing without a more rigorous analysis (e.g., a bootstrap test).

Chr1 tree, medium criteria, in newick format:

```
newick.chr1.med <- '(((tp3367_Italy:8813,tp1013_Wales:9652):9365,(((e_tp1007_Virginia:30,d_tp1012_Australia:61):1
cat.hardwrap(newick.chr1.med)

# (((tp3367_Italy:8813,tp1013_Wales:9652):9365,(((e_tp1007_Virginia:30,d_tp1012_Au
# stralia:61):19,(c_tp1015_Puget_Sound:207,b_tp1335_NY:41):18):1005,a_tp1014_Gyre:
# 61):3912):7054,outgroup:0);
```

NOTE: In early code, tree was not being recalculated; it was defined by constants in the following code chunk, hand-copied from the analysis above.

```
# tree parameters as nested lists
#   Internal nodes have subtrees sub1/2 and length
#   Root has sub1/2, but no length
#   Leaves have where, length, optionally, id, alt, nb.  (Omit id for 'outgroup'. Use 'alt' for less formal
#     labeling in cartoon version; it defaults to 'where'. Use 'nb' to add abcde annotations for legend.)
# This hand-made version is now subsumed by make.tree; retained for comparison
tree.by.hand <-
  list(
    sub1 = list(
      sub1 = list(
        sub1 = list(id=3367, length=8813, where='Venice, Italy', alt='Venice'),
        sub2 = list(id=1013, length=9652, where='Wales, UK'),
        length=9365),
      sub2 =  list(
        sub1 = list(
          sub1 = list(
            sub1 = list(id=1007, length=30, nb='e', where='Virginia, USA'),
```

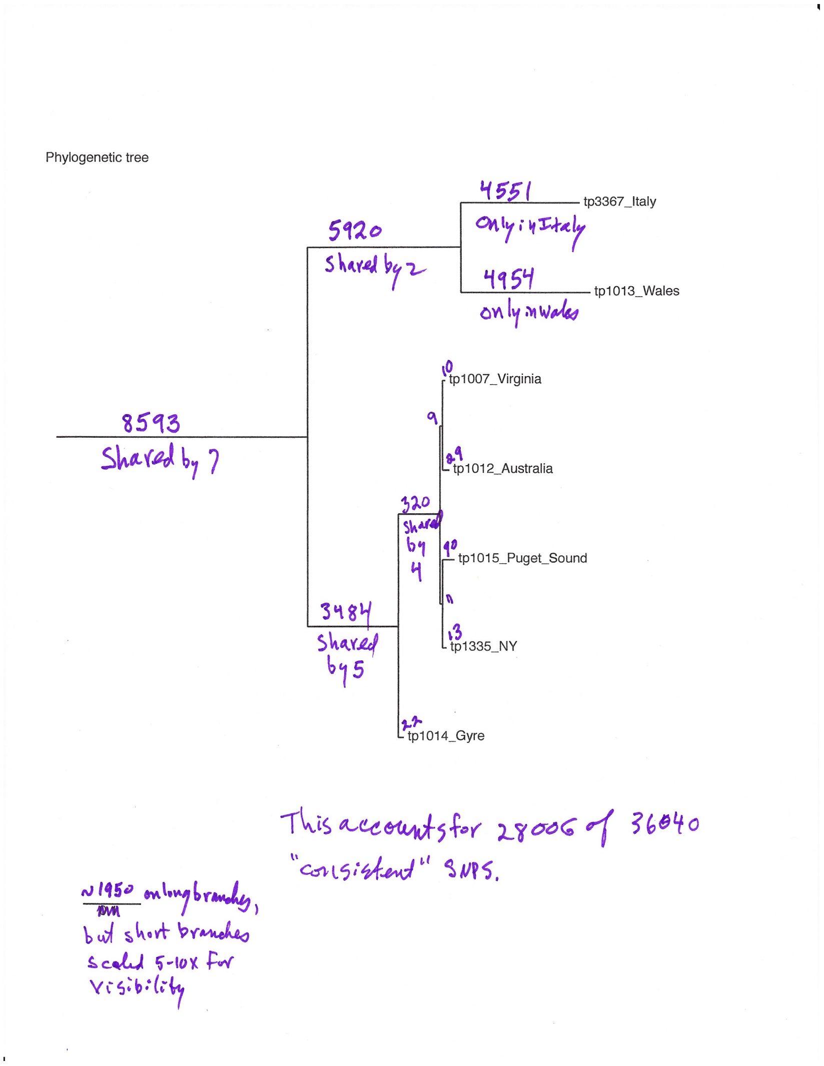

Figure 7: Inferred Tree, based on Chr1, un-q-filtered reads, loose criteria. (Note: to visually resolve the edges among the 5, their lengths were scaled by 5x – 10x in this figure, but not in the newick description shown in the text.)

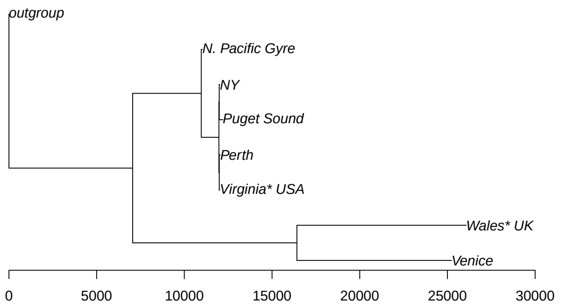

Figure 8: Tree based on unfiltered reads and medium SNP filters. “Lengths” are numbers of shared/private SNPs on Chr1. (no edge labels, nolegend)

```
            sub2 = list(id=1012, length=61, nb='d', where='Perth, W. Australia', alt='Perth'),
            length=19),
          sub2 = list(
            sub1 = list(id=1015, length=207,nb='c', where='Washington, USA', alt='Puget Sound'),
            sub2 = list(id=1335, length=41, nb='b', where='New York, USA',   alt='NY'),
            length=18),
          length=1005),
        sub2 = list(id=1014, length=61, nb='a', where='N. Pacific Gyre'),
        length=3912),
      length=7054),
    sub2 = list(length=0, where='outgroup')
  )

# historical, format example, and debug help:
oldwick.medium <- '(((CCMP3367_Italy:8813,CCMP1013_Wales:9652):9365,(((e_CCMP1007_Virginia:30,d_CCMP1012_Australia:61):19,(c_CCMP
# with simpler labeling for cartoon
simple.oldwick.medium <- '(((Italy:8813,Wales:9652):9365,(((Virginia:30,Australia:61):19,(Puget:207,NY:41):18):1005,Gyre:61):3912
cat.hardwrap(oldwick.medium)

# (((CCMP3367_Italy:8813,CCMP1013_Wales:9652):9365,(((e_CCMP1007_Virginia:30,d_CCM
# P1012_Australia:61):19,(c_CCMP1015_Puget_Sound:207,b_CCMP1335_NY:41):18):1005,a_
# CCMP1014_NPG:61):3912):7054,outgroup:0);

cat.hardwrap(simple.oldwick.medium)

# (((Italy:8813,Wales:9652):9365,(((Virginia:30,Australia:61):19,(Puget:207,NY:41)
# :18):1005,Gyre:61):3912):7054,outgroup:0);
```

Two other versions of the tree, for possible use in figs in the main paper.

Figure 8: [** as of 10/4/2015, this fig and next have stray stars on virginia, wales labels; probably due to hacking with commas in newick; not worth fixing unless we resurrect these trees for some purpose, but if so, see use of newick.name.undo in show.tree as probable fix. **]

```
tree.scale <- ifelse(which.snp.tables(string.val=F)[1]=='Chr1', 1, 10)
tree.x.lim <- 3e4 * tree.scale
the.simple.tree <- read.tree(text=simple.newick.medium)
plot(the.simple.tree, x.lim = tree.x.lim)
axis(1)
```

Figure 9:

```
plot(the.simple.tree, x.lim = tree.x.lim)
axis(1,(0:4)*7000*tree.scale,(0:4)*7000*tree.scale)
```

At some much earlier point, Tony ran the whole-genome version of the then-current code above, and manually entered tree branch lengths/legend for the resuting tree, shown in Fig 10. Code above can now automatically generate such a tree, but retain the following for comparison. The basic story seems clear—same topology and branch lengths

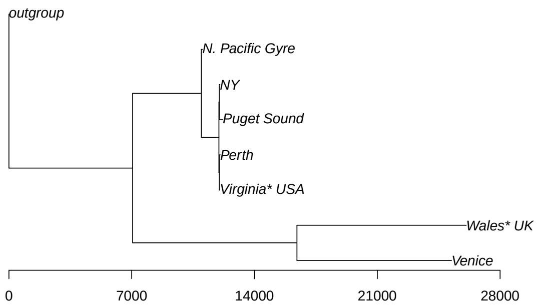

Figure 9: Tree based on unfiltered reads and medium SNP filters. “Lengths” are numbers of shared/private SNPs on Chr1. (no edge labels, no legend, short scale bar)

scaled by about 10x, which is completely reasonable given that Chr1 is about 10% of the genome. Note that this tree is not being recalculated; it is defined by constants in the following code chunk.

```
fullgenome.newick.medium <- '(((3367_Italy:86155,1013_Wales:95697):89598,(((e_1007_VA:330,d_1012_Australia:632):1296,(c_1015_WA:2
cat.hardwrap(fullgenome.newick.medium)

# (((3367_Italy:86155,1013_Wales:95697):89598,(((e_1007_VA:330,d_1012_Australia:63
# 2):1296,(c_1015_WA:2113,b_1335_NY:658):480):10059,a_1014_NPG:568):39517):69526,o
# utgroup:0);

legend.text <- c('a: only in 1014  ',
                 'b: only in 1335  ',
                 'c: only in 1015  ',
                 'd: only in 1012  ',
                 'e: only in 1007  ',
                 '*: shared by bcde',
                 '   shared by b/c ',
                 '   shared by d/e '
)
fullgenome.tree.x.lim <- 300000
fullgenome.counts <- c( 568, 658, 2113, 632, 330, 10059, 480, 1296 )
fullgenome.legend.text <- paste(legend.text,format(fullgenome.counts,width=5),sep=' - ')
fullgenome.tree.labels <- list( ## x,y,text
     41000,3.63,'69526\nshared by 7',
     90000,5.75,'39517\nby 5 (**)',
    115000,1.5, '89598\nshared by 2',
    210000,2.0, '95697 only\nin Wales',
    210000,1.0, '86155 only\nin Italy',
    113500,4.6, '*')
```

Figure 10:

```
library(ape)
the.fullgenome.tree <- read.tree(text=fullgenome.newick.medium)
plot(the.fullgenome.tree, x.lim = fullgenome.tree.x.lim)
axis(1) # ; axis(2) useful only for placing labels
opar <- par(family='mono',cex=.8)
legend('topright', legend=fullgenome.legend.text)
par(opar)
for(i in seq(1,length(fullgenome.tree.labels)-2,by=3)){
  text(fullgenome.tree.labels[[i]], fullgenome.tree.labels[[i+1]], fullgenome.tree.labels[[i+2]])
}
```

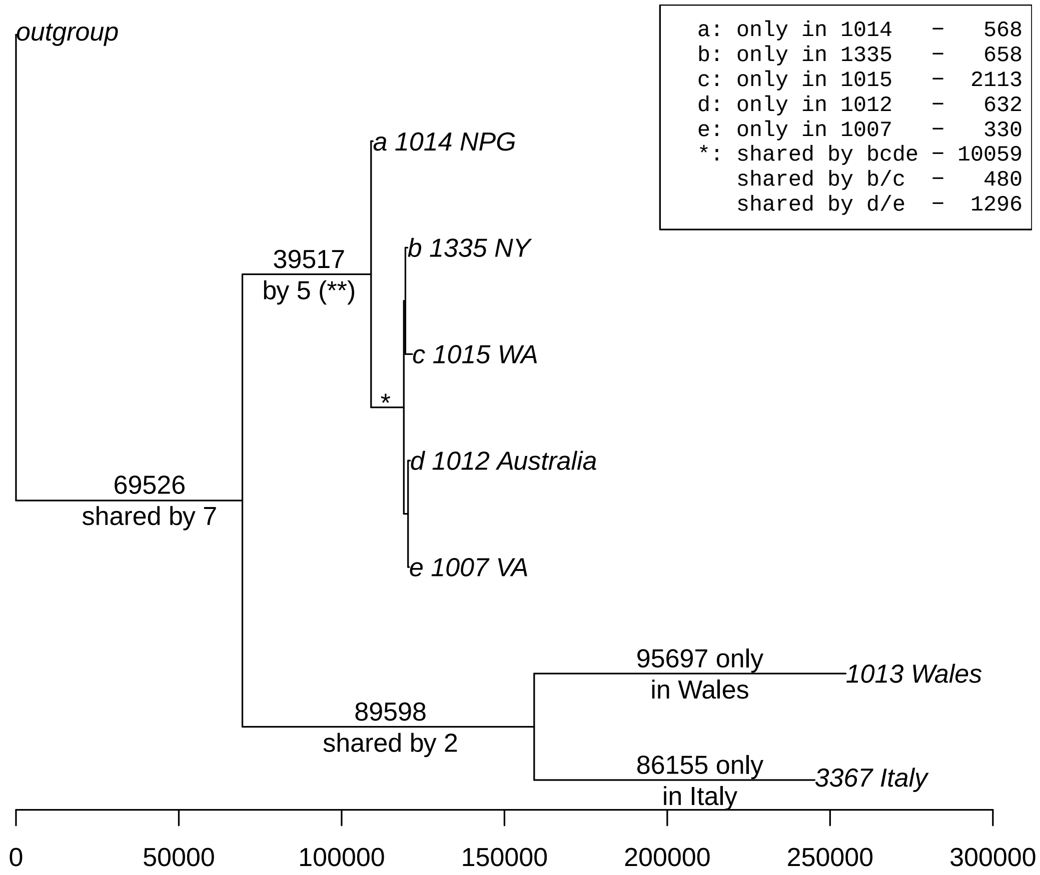

Figure 10: Tree based on unqfiltered reads and medium SNP filters. "Lengths" are numbers of shared/private SNPs genome-wide. (By-hand legacy version)