Exploration of Shared SNPs in Thaps

January 24, 2014

Some rather raw ramblings on SNP positions shared between two or more of the isolates. I've included my code, but I presume it will be largely uninteresting to you. I will summarize it as we go.

Load the data file, and prune it to just Chromosome 1:

```
load("~/Documents/s/papers/Thaps/tonys-svn/data/full.tables.ch1p100.rda")
tables <- lapply(full.tables.ch1, function(x) {
    x[x$chr == "Chr1", ]
})
```

Is brief, "tables" will be a list of 7 data frames, one per strain, giving read counts for each nucleotide at each position, SNP calls, etc.

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 value 0 means all nonreference counts are 0; the values 1..4 indicate that the max count occurred at A, G, C, T, resp. (Ties are resolved arbitrarily, and there is no attempt to control for low covergae levels. Both issues possibly deserve further attention.)

```
nref.nuc <- function(strain = 1, mask = T) {
    # get read count for max nonref nuc
    nref <- apply(tables[[strain]][mask, c("a", "g", "c", "t")], 1, max)
    # where does nref count match a (g,c,t, resp) count
    as <- ifelse(nref == tables[[strain]][mask, "a"], 1, 0)
    gs <- ifelse(nref == tables[[strain]][mask, "g"], 2, 0)
    cs <- ifelse(nref == tables[[strain]][mask, "c"], 3, 0)
    ts <- ifelse(nref == 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
    merge <- pmax(as, gs, cs, ts)
    # but if max nonref count is zero, return 0
    merge[nref == 0] <- 0
    return(merge)
}
```

Get union and intersection of the sets of called SNPs. ("$snp" is 0/1.)

```
union.snps <- tables[[1]]$snp
intersect.snps <- tables[[1]]$snp
for (i in 2:7) {
    union.snps <- pmax(union.snps, tables[[i]]$snp)
    intersect.snps <- pmin(intersect.snps, tables[[i]]$snp)
}
nusnps <- sum(union.snps)
nisnps <- sum(intersect.snps)
```

There are nusnps=47499 positions called as SNPs in one or more strains (but only nisnps=1641 that are shared among all 7). 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." This may seem an overly permissive definition, but, e.g., $> 85\%$ of all positions have zero reads for any non-reference nucleotide:

```
unlist(lapply(tables, function(x) {
    sum(x$Cov == x$.match)
}))/nrow(tables[[1]])

## tp1007 tp1012 tp1013 tp1014 tp1015 tp3367 tp1335
## 0.9249 0.8887 0.8496 0.8651 0.9063 0.8647 0.8523
```

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

```
non.refs <- matrix(0, nrow = nusnps, ncol = 7)
for (i in 1:7) {
    non.refs[, i] <- nref.nuc(i, union.snps == 1)
}
row.names(non.refs) <- paste(tables[[1]]$chr[union.snps == 1], ":", tables[[1]]$pos[union.snps ==
    1], sep = "")
```

"non.refs" indicates non-ref nucleotide has 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 any nonref reads in that position, in fact has most non-reference reads on the *same* nucleotide. These are defined as the "consistent" SNPs.

```
non.refs.max <- apply(non.refs, 1, max)
non.refs.min <- apply(non.refs, 1, function(x) {
    min(x[x > 0])
})
consistent <- non.refs.min == non.refs.max
sum(consistent)

## [1] 36040
```

Of the 47499 positions in which a SNP is called, 36040 are consistent. (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.)

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 whith 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.

The analysis is broken into cases based on how many strains share a particular SNP, counted as follows:

```
snp.counts <- apply(non.refs > 0, 1, sum)
```

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

```
sum(consistent & snp.counts == 7)  # 8593 on Chr1

## [1] 8593
```

I.e., of the 36040 consistent positions, 23.8% are shared by all 7 strains.

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

```
sum(consistent & snp.counts == 1)  #  9669 on chr1

## [1] 9669
```

```
singles <- vector("integer", 7)
names(singles) <- names(tables)
for (i in 1:7) {
    singles[i] <- sum(non.refs[consistent & snp.counts == 1, i] > 0)
}
print(singles)

## tp1007 tp1012 tp1013 tp1014 tp1015 tp3367 tp1335
##     10     29   4954     22     90   4551     13
```

The high counts for Italy and Wales suggest that they have been separated from each other and from the rest for a long time. Conversely, the low counts for the other 5 suggest that none of them has been isolated for very long (if at all).

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

```
sum(consistent & snp.counts == 2)  # 7641 on chr1

## [1] 7641

pairs <- matrix(0, nrow = 7, ncol = 7)
rownames(pairs) <- names(tables)
colnames(pairs) <- names(tables)
for (i in 1:6) {
    for (j in (i + 1):7) {
        pairs[i, j] <- sum(non.refs[consistent & snp.counts == 2, i] > 0 & non.refs[consistent &
            snp.counts == 2, j] > 0)
    }
}
print(pairs)

##         tp1007 tp1012 tp1013 tp1014 tp1015 tp3367 tp1335
## tp1007      0      9    105      2      5     93      1
## tp1012      0      0    165      7      5    150      3
## tp1013      0      0      0    222    125   5920    243
## tp1014      0      0      0      0     10    179      6
## tp1015      0      0      0      0      0    141     11
## tp3367      0      0      0      0      0      0    239
## tp1335      0      0      0      0      0      0      0
```

I.e., of the 7641 paired SNPs, 5920 or 77.5% are found between Italy and Wales, with comparatively few shared between any pair *not* including one of the European isolates.

SNPs shared among exactly 3 isolates are relatively rare, and the 5 trios containg both Italy and Wales predominate.

```
sum(consistent & snp.counts == 3)  # 1438 on chr1

## [1] 1438

triples <- NULL
for (i in 1:5) {
    for (j in (i + 1):6) {
        for (k in (j + 1):7) {
            temp <- sum(non.refs[consistent & snp.counts == 3, i] > 0 & non.refs[consistent &
                snp.counts == 3, j] > 0 & non.refs[consistent & snp.counts ==
                3, k] > 0)
            if (temp > 0) {
                triples <- rbind(triples, data.frame(i = names(tables)[i], j = names(tables)[j],
                  k = names(tables)[k], count = temp))
            }
```

```
        }
    }
}
print(triples[order(triples[4], decreasing = T), ])
```

```
##         i      j      k count
## 29 tp1013 tp3367 tp1335   327
## 25 tp1013 tp1014 tp3367   324
## 17 tp1012 tp1013 tp3367   227
## 27 tp1013 tp1015 tp3367   185
## 7  tp1007 tp1013 tp3367   134
## 23 tp1012 tp3367 tp1335    21
## 26 tp1013 tp1014 tp1335    20
## 32 tp1014 tp3367 tp1335    17
## 15 tp1012 tp1013 tp1014    13
## 18 tp1012 tp1013 tp1335    12
## 6  tp1007 tp1013 tp1015    11
## 14 tp1007 tp3367 tp1335    11
## 21 tp1012 tp1015 tp3367    11
## 24 tp1013 tp1014 tp1015    11
## 1  tp1007 tp1012 tp1013    10
## 3  tp1007 tp1012 tp3367     9
## 16 tp1012 tp1013 tp1015     9
## 20 tp1012 tp1014 tp3367     9
## 28 tp1013 tp1015 tp1335     9
## 30 tp1014 tp1015 tp3367     9
## 33 tp1015 tp3367 tp1335     9
## 8  tp1007 tp1013 tp1335     8
## 5  tp1007 tp1013 tp1014     7
## 31 tp1014 tp1015 tp1335     7
## 2  tp1007 tp1012 tp1015     6
## 22 tp1012 tp1015 tp1335     6
## 10 tp1007 tp1014 tp3367     4
## 13 tp1007 tp1015 tp1335     4
## 12 tp1007 tp1015 tp3367     3
## 4  tp1007 tp1012 tp1335     2
## 9  tp1007 tp1014 tp1015     1
## 11 tp1007 tp1014 tp1335     1
## 19 tp1012 tp1014 tp1015     1
```

Four-way sharing is even less common, with the non-European coastal isolates (i.e., not gyre) dominating.

```
sum(consistent & snp.counts == 4)  # 564 on chr1
```

```
## [1] 564
```

```
quads <- NULL
for (i in 1:4) {
    for (j in (i + 1):5) {
        for (k in (j + 1):6) {
            for (l in (k + 1):7) {
                temp <- sum(non.refs[consistent & snp.counts == 4, i] > 0 &
                  non.refs[consistent & snp.counts == 4, j] > 0 & non.refs[consistent &
                  snp.counts == 4, k] > 0 & non.refs[consistent & snp.counts ==
                  4, l] > 0)
                if (temp > 0) {
                  quads <- rbind(quads, data.frame(i = names(tables)[i], j = names(tables)[j],
                    k = names(tables)[k], l = names(tables)[l], count = temp))
                }
```

```
            }
        }
    }
}
print(quads[order(quads[5], decreasing = T), ])
```

```
##         i      j      k      l count
## 9  tp1007 tp1012 tp1015 tp1335   320
## 30 tp1013 tp1014 tp3367 tp1335    30
## 21 tp1012 tp1013 tp1015 tp3367    25
## 31 tp1013 tp1015 tp3367 tp1335    22
## 23 tp1012 tp1013 tp3367 tp1335    18
## 5  tp1007 tp1012 tp1014 tp1015    16
## 13 tp1007 tp1013 tp1015 tp3367    12
## 28 tp1013 tp1014 tp1015 tp3367    12
## 3  tp1007 tp1012 tp1013 tp3367    11
## 15 tp1007 tp1013 tp3367 tp1335    10
## 4  tp1007 tp1012 tp1013 tp1335     9
## 12 tp1007 tp1013 tp1014 tp3367     9
## 22 tp1012 tp1013 tp1015 tp1335     9
## 2  tp1007 tp1012 tp1013 tp1015     8
## 19 tp1012 tp1013 tp1014 tp3367     8
## 8  tp1007 tp1012 tp1015 tp3367     7
## 25 tp1012 tp1014 tp1015 tp1335     6
## 17 tp1007 tp1015 tp3367 tp1335     5
## 14 tp1007 tp1013 tp1015 tp1335     4
## 29 tp1013 tp1014 tp1015 tp1335     4
## 24 tp1012 tp1014 tp1015 tp3367     3
## 6  tp1007 tp1012 tp1014 tp3367     2
## 11 tp1007 tp1013 tp1014 tp1015     2
## 18 tp1012 tp1013 tp1014 tp1015     2
## 20 tp1012 tp1013 tp1014 tp1335     2
## 27 tp1012 tp1015 tp3367 tp1335     2
## 1  tp1007 tp1012 tp1013 tp1014     1
## 7  tp1007 tp1012 tp1014 tp1335     1
## 10 tp1007 tp1012 tp3367 tp1335     1
## 16 tp1007 tp1014 tp3367 tp1335     1
## 26 tp1012 tp1014 tp3367 tp1335     1
## 32 tp1014 tp1015 tp3367 tp1335     1
```

Five-way sharing is much more common, and is strongly dominated by the 5 non-Europeans.

```
sum(consistent & snp.counts == 5)  # 3969 on chr1
```

```
## [1] 3969
```

```
quints <- NULL
for (i in 1:3) {
    for (j in (i + 1):4) {
        for (k in (j + 1):5) {
            for (l in (k + 1):6) {
                for (m in (l + 1):7) {
                  temp <- sum(non.refs[consistent & snp.counts == 5, i] > 0 &
                    non.refs[consistent & snp.counts == 5, j] > 0 & non.refs[consistent &
                    snp.counts == 5, k] > 0 & non.refs[consistent & snp.counts ==
                    5, l] > 0 & non.refs[consistent & snp.counts == 5, m] >
                    0)
                  if (temp > 0) {
                    quints <- rbind(quints, data.frame(i = names(tables)[i],
```

```
                        j = names(tables)[j], k = names(tables)[k], l = names(tables)[l],
                        m = names(tables)[m], count = temp))
                }
            }
        }
    }
}
print(quints[order(quints[6], decreasing = T), ])
```

```
##         i      j      k      l      m count
## 8  tp1007 tp1012 tp1014 tp1015 tp1335  3484
## 5  tp1007 tp1012 tp1013 tp1015 tp1335   201
## 10 tp1007 tp1012 tp1015 tp3367 tp1335   125
## 18 tp1012 tp1013 tp1015 tp3367 tp1335    33
## 4  tp1007 tp1012 tp1013 tp1015 tp3367    30
## 20 tp1013 tp1014 tp1015 tp3367 tp1335    17
## 6  tp1007 tp1012 tp1013 tp3367 tp1335    12
## 14 tp1007 tp1013 tp1015 tp3367 tp1335    11
## 15 tp1012 tp1013 tp1014 tp1015 tp3367     9
## 7  tp1007 tp1012 tp1014 tp1015 tp3367     8
## 1  tp1007 tp1012 tp1013 tp1014 tp1015     7
## 16 tp1012 tp1013 tp1014 tp1015 tp1335     7
## 19 tp1012 tp1014 tp1015 tp3367 tp1335     5
## 11 tp1007 tp1013 tp1014 tp1015 tp3367     4
## 3  tp1007 tp1012 tp1013 tp1014 tp1335     3
## 9  tp1007 tp1012 tp1014 tp3367 tp1335     3
## 13 tp1007 tp1013 tp1014 tp3367 tp1335     3
## 17 tp1012 tp1013 tp1014 tp3367 tp1335     3
## 2  tp1007 tp1012 tp1013 tp1014 tp3367     2
## 12 tp1007 tp1013 tp1014 tp1015 tp1335     2
```

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

```
sum(consistent & snp.counts == 6)  # 4166 on chr1
```

```
## [1] 4166
```

```
sexts <- NULL
for (i in 1:7) {
    temp <- sum(non.refs[consistent & snp.counts == 6, i] == 0)
    if (temp > 0) {
        sexts <- rbind(sexts, data.frame(EXclude = names(tables)[i], count = temp))
    }
}
print(sexts[order(sexts[2], decreasing = T), ])
```

```
##   EXclude count
## 6  tp3367  1756
## 3  tp1013  1343
## 4  tp1014   951
## 7  tp1335    45
## 1  tp1007    43
## 5  tp1015    17
## 2  tp1012    11
```

So, overall, the picture looks like a long shared history (8600 7-way shared positions), followed by a split of the 5 from Europe, then a long shared history in the 5 (3484 quintuples), in parallel with a long shared history in Europe (5920 pairs), then long separate histories in Italy and Wales (>4500), and essentially nolimited differentiation among

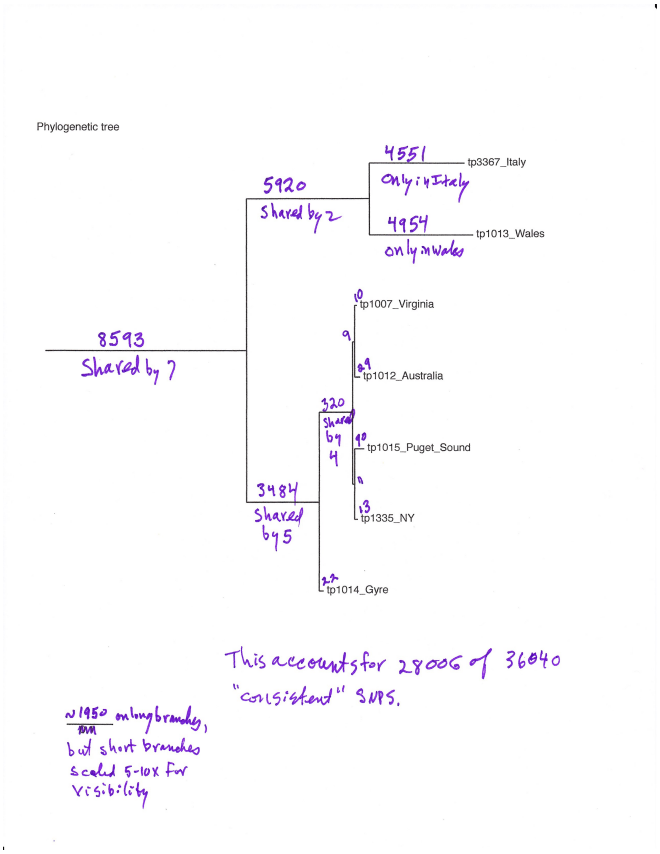


Figure 1: Inferred Tree.

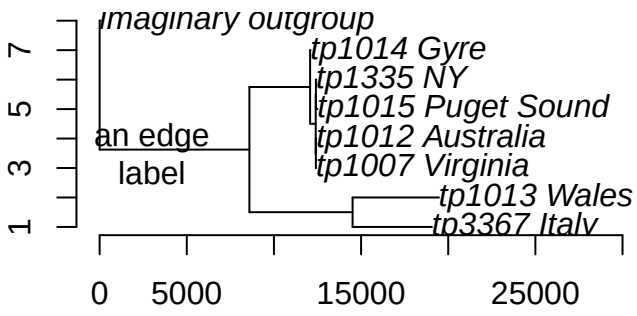
the 5 non-Europeans, but if we split hairs, we get the following tree (newick format):

```
(((tp3367_Italy:4551,tp1013_Wales:4954):5920,(((tp1007_Virginia:10,tp1012_Australia:29):9,
(tp1015_Puget_Sound:90,tp1335_NY:13):11):320,tp1014_Gyre:22):3484):8593);
```

rendered via `http://iubio.bio.indiana.edu/treeapp/treeprint-form.html` as Fig 1. Note: to visually resolve the edges among the 5, I scaled those length by 5x - 10x.

Maybe I can plot a tree:

```
newick <- "(((tp3367_Italy:4551,tp1013_Wales:4954):5920,(((tp1007_Virginia:10,tp1012_Australia:29):9,
library(ape)
plot(read.tree(text = newick))
axis(1)
axis(2)
text(3000, 3.5, "an edge\nlabel")
```



Looking at pairwise counts of shared SNPs (without regard to how many other strains share the SNP), we have:

```
sum(consistent)  # 36040 on chr1

## [1] 36040

pairwise <- matrix(0, nrow = 7, ncol = 7)
rownames(pairwise) <- names(tables)
colnames(pairwise) <- names(tables)
for (i in 1:6) {
    for (j in (i + 1):7) {
        pairwise[i, j] <- sum(non.refs[consistent, i] > 0 & non.refs[consistent,
            j] > 0)
    }
}
print(pairwise)

##         tp1007 tp1012 tp1013 tp1014 tp1015 tp3367 tp1335
## tp1007      0  16992  11989  15328  16975  11470  16893
## tp1012      0      0  12241  15400  17076  11727  16992
## tp1013      0      0      0  11189  12170  17058  12390
## tp1014      0      0      0      0  15419  10715  15387
## tp1015      0      0      0      0      0  11675  17001
## tp3367      0      0      0      0      0      0  11885
## tp1335      0      0      0      0      0      0      0

pw <- pairwise + t(pairwise)
p <- c(1, 2, 5, 7, 4, 3, 6)
print(pw[p, p])

##         tp1007 tp1012 tp1015 tp1335 tp1014 tp1013 tp3367
## tp1007      0  16992  16975  16893  15328  11989  11470
## tp1012  16992      0  17076  16992  15400  12241  11727
## tp1015  16975  17076      0  17001  15419  12170  11675
## tp1335  16893  16992  17001      0  15387  12390  11885
## tp1014  15328  15400  15419  15387      0  11189  10715
## tp1013  11989  12241  12170  12390  11189      0  17058
## tp3367  11470  11727  11675  11885  10715  17058      0
```

Noise: Various sources of "noise" in the data:

1. deep coalescence
2. read errors
3. low reads depth
4. skew because 1335 is the reference
5. varying error rates and sequencing depth among the 7
6. varying numbers of founder cells in the sequencing cultures
7. tri-allelic positions where stochastic fluctuation is sequence sampling promotes the rare allele to prominence

To Do:

1. try filtering out singleton reads
2. any spacial structure to various sub-classes?
3. any association of .8 group to various subclasses?
4. after top level split, should I reanalyze halves of partition in isolation?