700.1-Examples

Mac Campbell

2/5/2020

Let's show a few results to update the crew.

Here is a representative, but not exhaustive sampling. I have:
1 Excluded inversions
2 Removed tetrasomically-pairing chromosome arms (to reduce paralogy and incomplete lineage sorting)
3 Removed linked-snps for compatability with multispecies coalescent-based approach
4 Conducted some model choice

This alignment has 32 sequences with 5917 columns, 5899 distinct patterns, 3408 parsimony-informative, 2509 singleton sites. If not removing linked SNPs, our alignment has 23759 columns, 10452 distinct patterns, 15680 parsimony-informative, 8079 singleton sites. That's substantially different.

What have I been thinking about?

(1) Homology - hence the exclusion of several tetrasomic pairing regions - Campbell et al. (2019), Blumstein et al. (Recently rejected, :(, resubmission soon to G3).

(2) Incomplete Lineage Sorting (ILS) - Exclusion tetrasomic regions decreases ILS Campbell et al. (In Review). But we can also incorporate a multispecies coalescent model. Since SNAPP is awful, I have used the one other approach I know that models ILS. I am not satisfied with it. The data are not compatible with methods that don't conduct joint estimation of the species tree and gene trees. STARBEAST2 has faulted every time I have tried it with the alignments made.

(3) As hybridization & ILS may be present, networks may show this. It is important here.

(4) Admixture with an archaic lineage is a good idea and I haven't excluded this idea yet and have not explicitly addressed it. At the moment, I'm going with MRRT are a distinct lineage as network analyses don't show admixture.

What do I want to conclude?

There is evidence for four major lineages (1) California Golden Trout Complex (2) Columbia River Redbands (3) Coastal Rainbow Trout + Eagle Lake Rainbow and (4) McCloud River Redband Trout. Relationships among these four lineages is hard to conclude on as there are few sites that may inform the branching patterns.

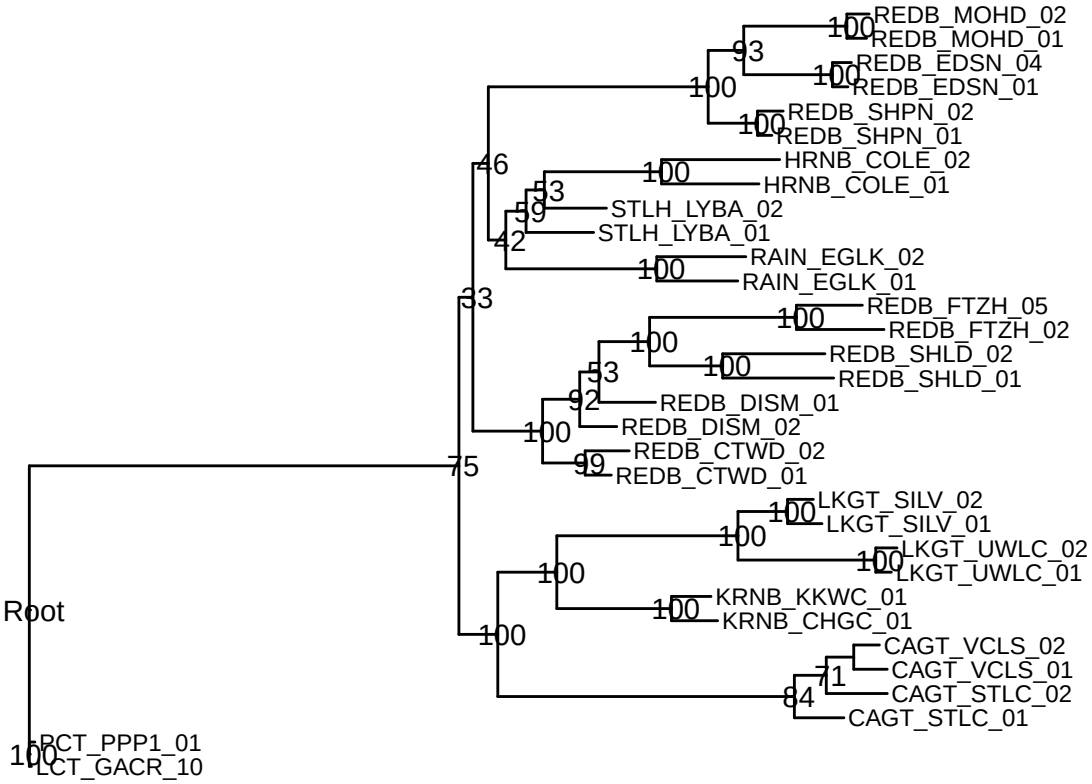
McCloud River Redband Trout is distinctive, and the application of a subspecific name does seem appropriate. McCloud River Redband Trout is not at all closely related to geographically proximate redband populations found in the Northern Sacramento River basin, or Columbia River Redbands as a whole. A new official common name should be applied to this trout.

Some Results

First off, let's look at a concatenated analysis.

```
iqtree -s test6-maf05p-sites.sort.reheadered.min4.asc.phy -st DNA -m MFP+ASC -bb 1000 -alrt 1000 --redo
```

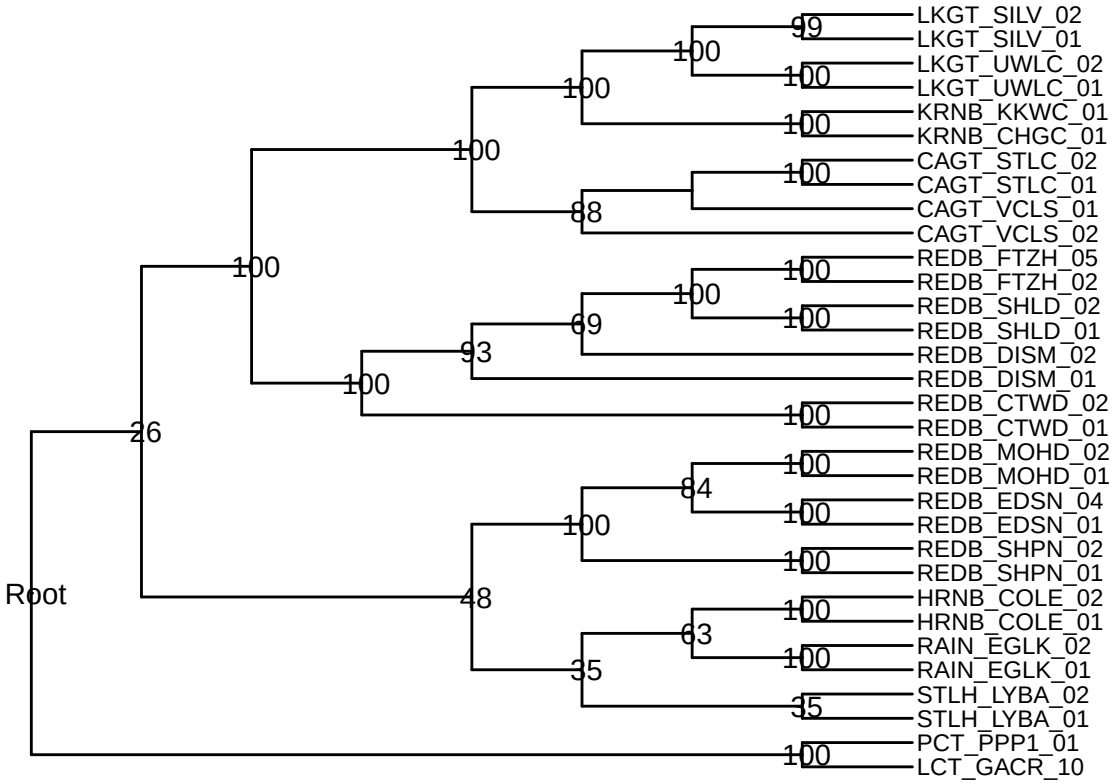
```
library(tidyverse)
library(ggtree)
library(ape)
t<-read.tree(file="outputs/700/test6-maf05p-sites.sort.prune.reheadered.min4.asc.phy.contree")
t<-root(t,c("PCT_PPP1_01","LCT_GACR_10"))
ggtree(t)+geom_tiplab(size=3)+geom_nodelab()+xlim(0,.2)
```



We can leave the linked SNPs in though, as that doesn't violate the ML construction assumptions. I have found that it may increase some nodal support values, and that is important because our results have low nodal support values, but, I don't think it is a great idea, and causes topological changes in this case.

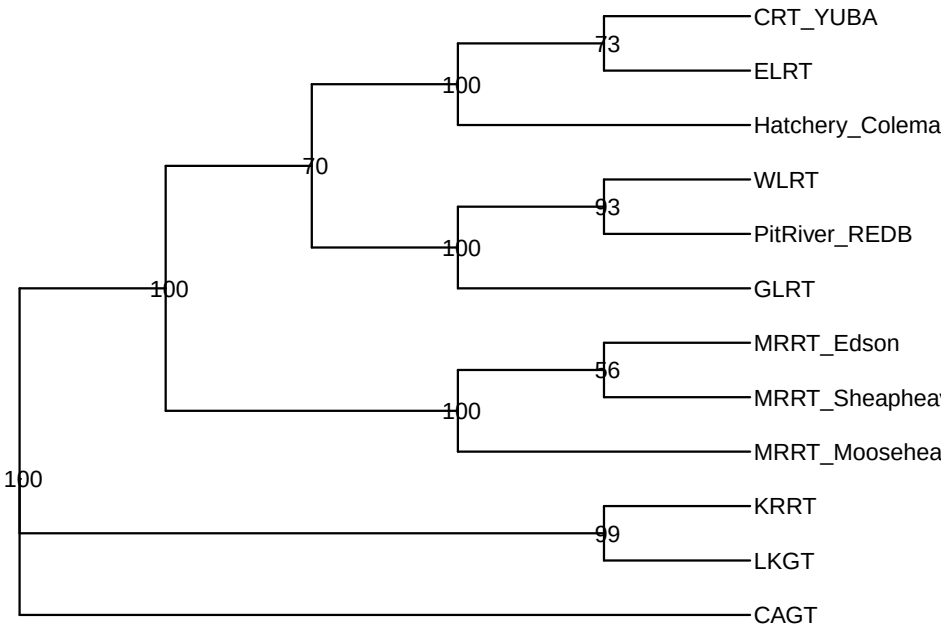
```
iqtree -s test6-maf05p-sites.sort.reheadered.min4.asc.phy -st DNA -m MFP+ASC -bb 1000 -alrt 1000 --redo
```

```
tt<-read.tree(file="outputs/700/test6-maf05p-sites.sort.reheadered.min4.asc.phy.contree")
tt<-root(tt,c("PCT_PPP1_01","LCT_GACR_10"))
ggtree(tt, branch.length = "none")+geom_tiplab(size=3)+geom_nodelab()+xlim(0,10)
```



Here is an example from SVDQuartets incorporating the multispecies coalescent. We don't need the rooting of LCT, so I have dropped it and recalled genotypes.

```
t2<-read.nexus(file="outputs/700/test7/svdq-test.asc.tre")
ggtree(t2,branch.length = "none")+geom_tiplab(size=4)+geom_nodelab()+xlim(0,6)
```



Networks provide a way of showing reticulations in the phylogeny, instead of bifurcations.

Here is a neighbor-net version followed by an alternative distance-based network.

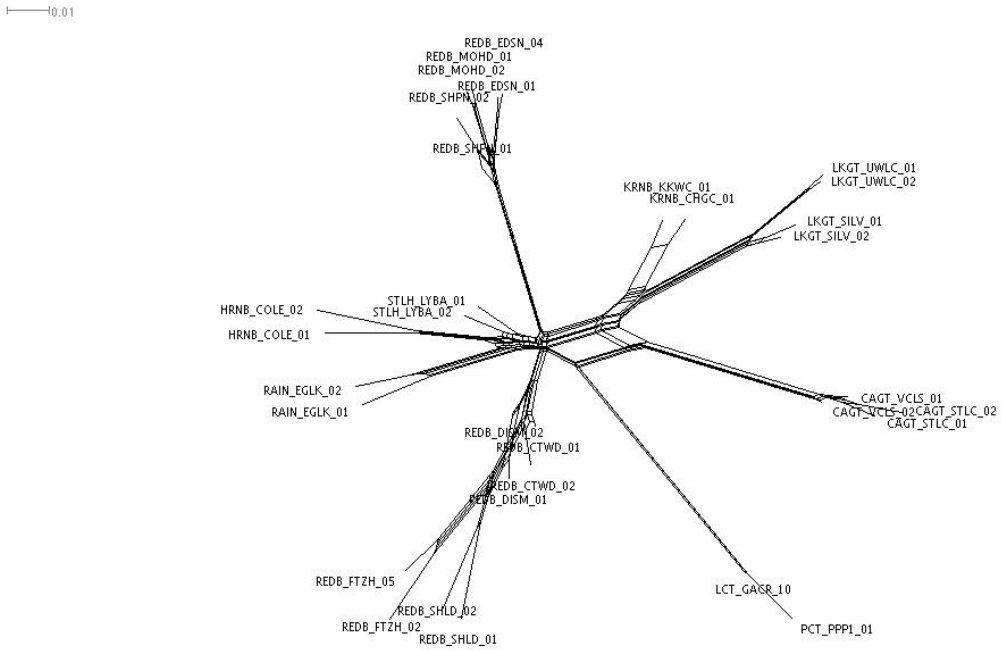


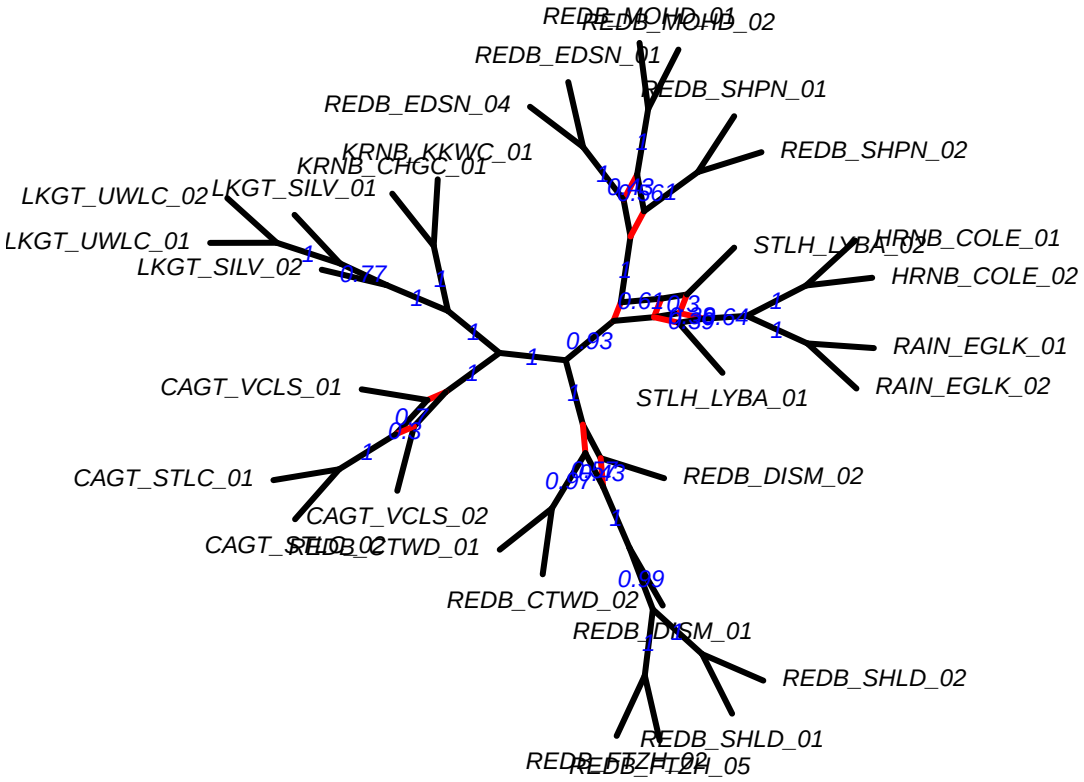
Figure 1: Neighbor-Net

```
library(phangorn)
dna<-read.phyDat("outputs/700/test7-maf05p-sites.sort.prune.reheadered.min4.asc.phy")
dat<-as.phyDat(dna)
set.seed(1)
bs <- bootstrap.phyDat(dat, FUN = function(x)nj(dist.hamming(x)),
    bs=100)
tree <- nj(dist.hamming(dat))
par("mar" = rep(1, 4))
#tree<-root(tree, c("PCT_PPP1_01\t","LCT_GACR_10\t"))
#tree <- plotBS(tree, bs, "phylogram")
#By default prob=0.3
cnet <- consensusNet(bs, .25)
cnet$tip.label<-gsub("\t","",cnet$tip.label)

#plot(cnet, "2D", show.edge.label=TRUE)

edge.col <- createLabel(cnet, tree, "black", nomatch="red")

plot(cnet, show.edge.label = T, "2D", edge.color = edge.col,
                  col.edge.label = "blue", cex=.75)
```


REDB_MOHD_01
REDB_MOHD_02
REDB_EDSN_01
REDB_SHPN_01
REDB_EDSN_04
REDB_SHPN_02
KRNB_KKWC_01
KRNB_CHGC_01
LKGT_UWLC_02
LKGT_SILV_01
LKGT_UWLC_01
LKGT_SILV_02
STLH_LYBA_02
HRNB_COLE_01
HRNB_COLE_02
RAIN_EGLK_01
RAIN_EGLK_02
STLH_LYBA_01
CAGT_VCLS_01
CAGT_STLC_01
CAGT_VCLS_02
REDB_DISM_02
REDB_CTWD_01
REDB_CTWD_02
REDB_DISM_01
REDB_SHLD_02
REDB_SHLD_01