<div align="center">

`rclonetree` Package Vignette

</div>

1   Introduction

This package is designed to enable users to map clonal somatic mutations to a specified tree. The method requires the mutant read counts and depth for each sample/loci in the form of matrices.

2   Simulated data

Generate tree

```
>   tree=generate_random_tree(50)
```
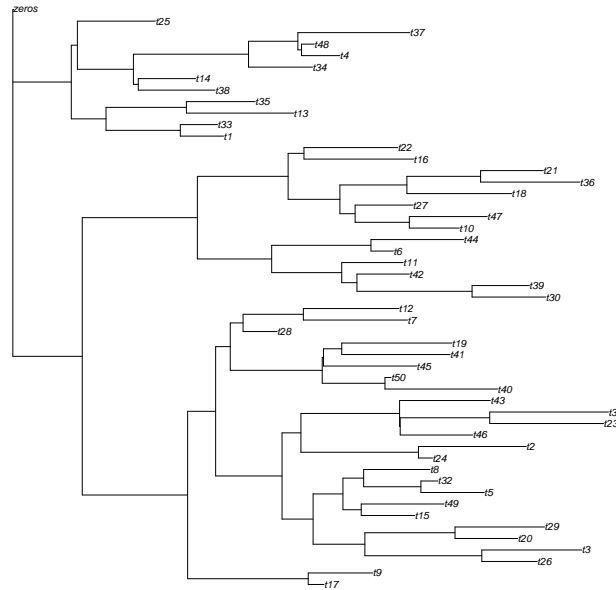
```
> plot(tree)
```


Figure 1: Randomly generated tree

Note that the tree includes an outgroup "zeros". This can be added to a tree obtained from other sources by

```
> tree=bind.tree(tree,read.tree(text="(zeros:0);"))
```

Now we create a summary object that summarises each branch of the tree as a genotype. This just requires the APE tree as input:

```
>   df=reconstruct_genotype_summary(tree)
> cat(df$samples,"\n")

t17 t9 t26 t3 t20 t29 t15 t49 t5 t32 t8 t24 t2 t46 t23 t31 t43 t40 t50 t45 t41 t19 t28 t7 t1

> head(df$df[, 1:3])

                                                profile edge_length mut_count
1 111111111111111111111111111111111111111111000000000000         114        39
2 111111111111111111111110000000000000000000000000000000         173        25
3 110000000000000000000000000000000000000000000000000000         198         2
4 100000000000000000000000000000000000000000000000000000          26         1
5 010000000000000000000000000000000000000000000000000000         106         1
6 001111111111111111111111110000000000000000000000000000          46        23
```

Where

1. "profile" column indicates the genotype (in the same order as df sample)

2. "edge_length" gives the branch edge length i.e. number of mutations assigned to branch (perfectly known here)

3. "mut_count" Indicates the number of mutant samples implied by the genotype (i.e. the number of 1s in "profile")

We now simulate some data based on the tree:

```
>  simdat=simulate_reads_from_tree(df,12)
> head(simdat$mtr)

      t17 t9 t26 t3 t20 t29 t15 t49 t5 t32 t8 t24 t2 t46 t23 t31 t43 t40 t50 t45
[1,]    9 10   3  2   5   7   2   4  7   5  2  11  8   9   7   5   5   5  11  10
[2,]    8  7   5 11   7  11   7  13 11   2  4   5  7   7   4   6   9   7   5   8
[3,]    4  3  13  3  10   5   8   4  6   5  8   9  3   6   1   8   6   6   9   3
[4,]    5  7   2  9   3   7   5   4  5   4  4   4 10   6   5   6   5   4   6   7
[5,]    4  3   6  7   5   6   2   5  4   3  4   5  9   4   7   3   3   4   3   4
[6,]   11  4   7  8   6  10   3   3  8   3  9   9  6   6   9   3   9   6   9   8
     t41 t19 t28 t7 t12 t30 t39 t42 t11 t6 t44 t10 t47 t27 t18 t36 t21 t16 t22
[1,]   4   3   6  4   8   3   4  12  14  6   5  10   3   8   8   7   2   5   8
[2,]   6   7   7  8   7   9   5   4   9  8   5   4   6   6   7   4   4   4   6
[3,]   4   7   4  6   7   9   6   4   2 11   5   7   2   8   9   3   4   4   5
[4,]   4   5   5  3   4   5   7   8   1 11   3   3   7   9   4   3   3   7   7
[5,]  11   6   6  5   7   5   4   9   5  3   7   7   3   2   6   3   4   4   6
[6,]  11  10   6  6   6   3   3   4   4  7   4   8   8   3   2   7   2   6   7
```

```
      t1 t33 t13 t35 t38 t14 t34 t4 t48 t37 t25 zeros
[1,]   0   0   0   0   1   0   0  0   0   0   0     0
[2,]   0   0   0   0   0   0   0  1   0   0   0     0
[3,]   0   0   0   0   0   0   0  0   0   0   0     0
[4,]   0   0   1   0   0   0   0  0   0   0   0     0
[5,]   0   0   0   0   0   0   0  0   0   1   0     0
[6,]   0   0   1   0   0   0   0  0   0   0   1     0

> head(simdat$depth)

      t17 t9 t26 t3 t20 t29 t15 t49 t5 t32 t8 t24 t2 t46 t23 t31 t43 t40 t50 t45
[1,]   19 17   5  8   9  11  10   8  9  10  9  20 12  19  14   9  13  11  14  12
[2,]   13 13  12 17  16  18  14  15 19  10  8  12 17  14   8   8  19  16  11  14
[3,]   11  8  16  6  19  16  12  14 13  16 13  20 13   8   9  10  12  13  14  11
[4,]    7 13   6 14   8  15   8  10  7  16 10  15 18   7  10  13   8   9  14  14
[5,]    8  8  11 14  12  11   6   7 11   9  7   8 14   9  12   9   7  10  12  10
[6,]   17  8   8 13  13  14   9   6 12  15 16  17 13  13  12   8  13  13  17  11
      t41 t19 t28 t7 t12 t30 t39 t42 t11 t6 t44 t10 t47 t27 t18 t36 t21 t16 t22
[1,]   10   8   8 12  13  11  10  14  20 10  13  16   7  16  18  10   9  16  12
[2,]   12  12  16 12  15  12  12  13  15 12   8   5  14  11  14   9  14  12  10
[3,]   13  12   9 15  14  17   9  10   9 16   8  12   8  14  19   7   8  14  15
[4,]   10   9  12  9   7   8  12  13   4 19   8   7  14  14  12   9  16  11  12
[5,]   17  12  13 13  14  14   7  15  12  4  12  10  15   9  13  11   9   8  15
[6,]   22  13  17  9  15   6   7  15   9 15  10  20  12   9   8  11   7  12  13
      t1 t33 t13 t35 t38 t14 t34 t4 t48 t37 t25 zeros
[1,]  12  13  13  10  18  18  13  7   9  14  11    10
[2,]  16  19  10   8   8  16  20 10  15   9  10    10
[3,]  13  19  11  15  11  13  11 14  10   8   9    10
[4,]  12  15  11  14  10  13  14  9  15  11   7    10
[5,]  15   7   8  13   8  11  12  4  13  11  11    10
[6,]  13  13  20  12  14  10  16  9  11   7  12    10

> print(simdat$p.error)

 [1] 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02
[13] 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02
[25] 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02
[37] 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02 1e-02
[49] 1e-02 1e-02 1e-06
```

We now have required information: mutant read matrix "mtr", depth matrix
"depth", genotype summary "df",and base calling error rate "p.error" - note how
we've set the last entry corresponding to "zeros" outgroup very low.

```
>  res=assign_to_tree(tree,simdat$mtr,simdat$depth,error_rate=simdat$p.error)

delta edge length= 0.9906
Loglik= -602728
```

```
delta edge length= 0.01205
Loglik= -552757
delta edge length= 0.0007682
Loglik= -552750
delta edge length= 6.847e-05
Loglik= -552749
delta edge length= 7.886e-06
Loglik= -552749
Finished assigning mutations
calculating pvalues
On 1000  of  10341
On 2000  of  10341
On 3000  of  10341
On 4000  of  10341
On 5000  of  10341
On 6000  of  10341
On 7000  of  10341
On 8000  of  10341
On 9000  of  10341
On 10000  of  10341

> tree_estimated=res$tree
> par(mfcol=c(1,2))
> plot(ladderize(tree,right=TRUE),cex=0.5)
> plot(ladderize(tree_estimated,right=TRUE),cex=0.5)


> sim=list(edge_length_orig=df$df$edge_length,
          edge_length_inferred=res$df$df$edge_length,
          expected_edge_length_inferred=res$df$df$expected_edge_length,
          edge_idx_orig=simdat$edge,
          edge_idx_ml=res$summary$edge_ml)
> plot_sim_result(sim)
```
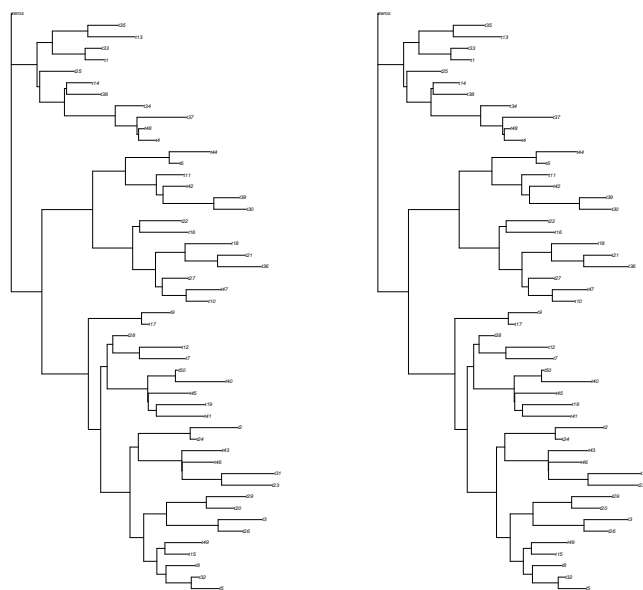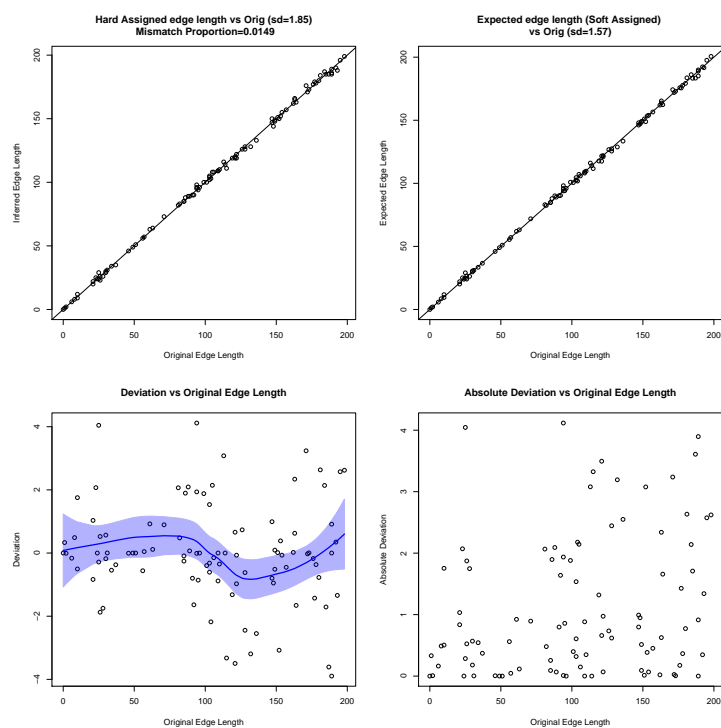
Figure 2: Tree comparison

Figure 3: Edge length comparison