

Compatibility, Cliques and Clonal Frames

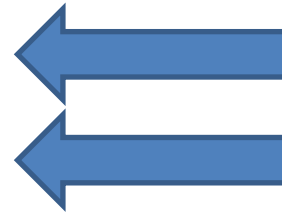
Barbara Holland
University of Tasmania



Unravelling the processes of bacterial evolution

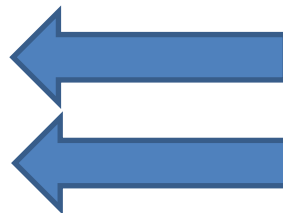
- Processes

- Mutation
- Homologous recombination
- HGT



- Data is available at multiple levels of resolution

- Gene presence / absence
- Allele profile
- Sequence data

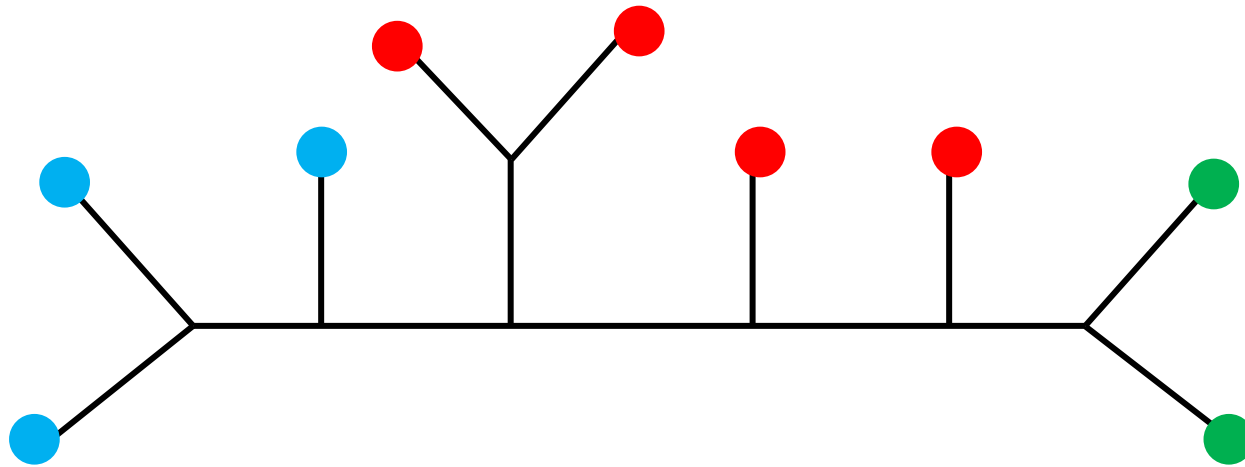


Talk structure

- Compatibility concepts
- Allele profile data
- Range of recombination models
- Fit of *Campylobacter* data to models
- To do list

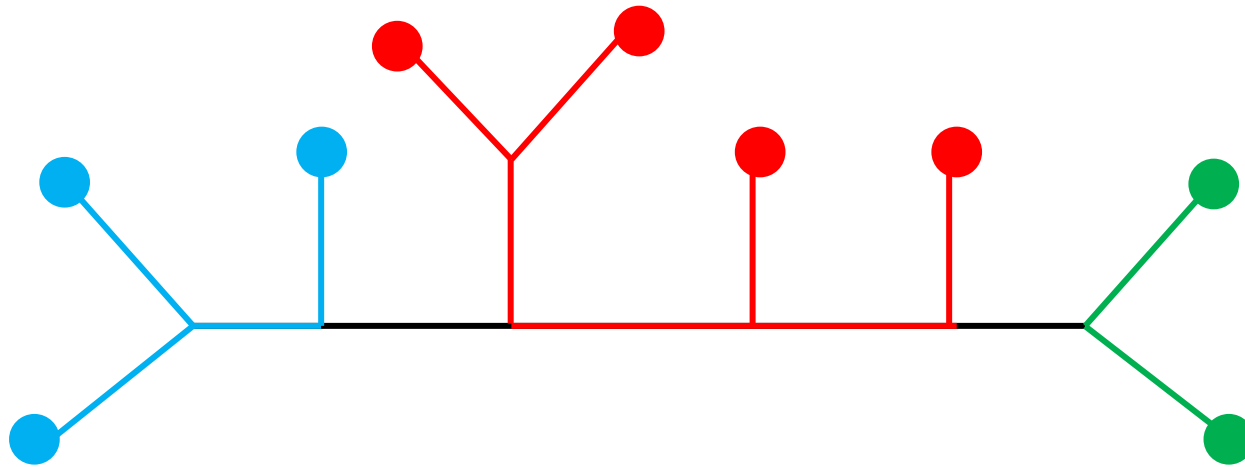
Compatibility

Given a character C and a tree T we can ask if the character is compatible with the tree.



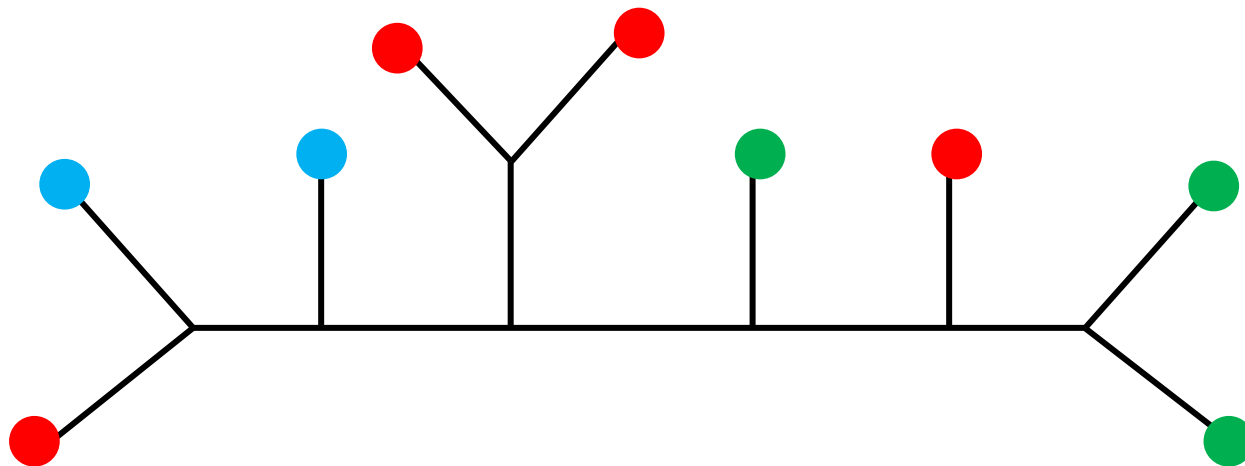
Compatibility

Given a character C and a tree T we can ask if the character is compatible with the tree.

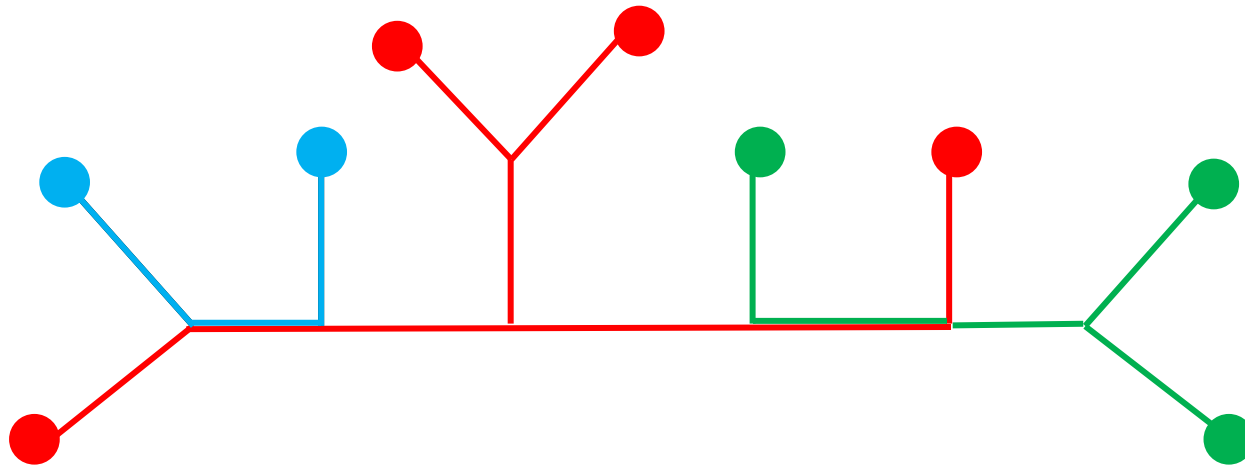


A compatible character

Incompatibility



Incompatibility



An incompatible character

Compatible pairs of characters

- Two characters are said to be compatible with each other if there exists a tree which they are both compatible with.

Cliques

- Two possible definitions
 - A set of characters that are all pairwise compatible
 - A set of characters for which there exists a tree that they are all compatible with
- These two are equivalent for binary characters but not for characters with 3 or more states.

Allele profile data

- Multi-level data
 - Strain type
 - Allele profile
 - Sequence

e.g. MLST data

locus	L1	L2	L3	L4	L5	L6	L7
ST1	1	1	1	1	1	1	1
ST2	1	1	2	1	1	1	1
....							



L3

1	CCCTTGTTTAGTCCAAATTCACACCAATTTCA
2	CCCTT A TTTAGTCCAAATTCACACCAATTTCA
...	...

Allele profile data

- Multi-level data
 - Strain type
 - Allele profile
 - Sequence

e.g. MLST data

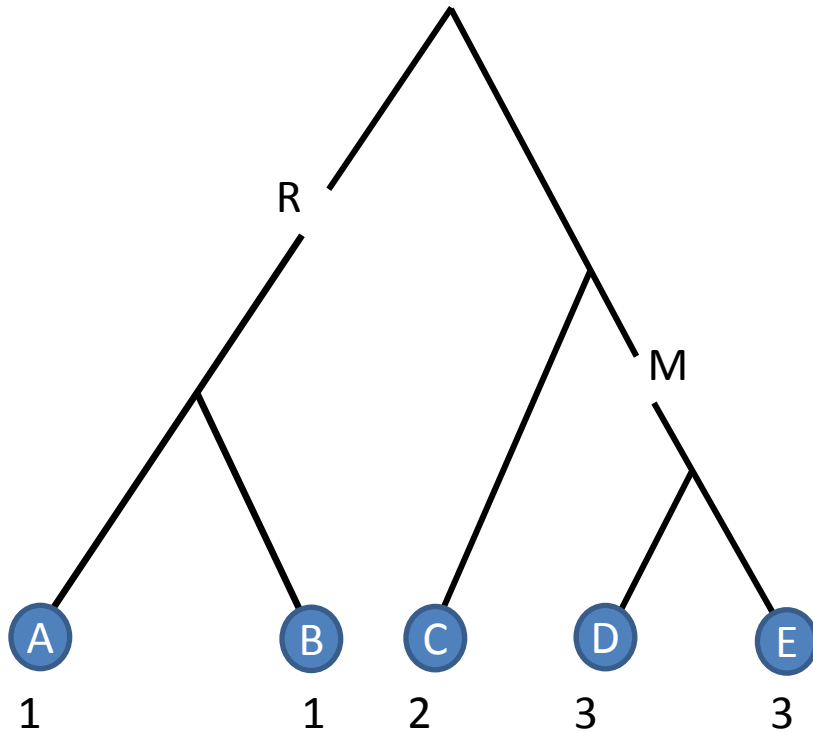
locus	L1	L2	L3	L4	L5	L6	L7
ST1	1	1	1	1	1	1	1
ST2	1	1	2	1	1	1	1
....							



L3

1	CCCTTGTTTAGTCCAAATTCACACCAATTTCA
2	CCCTT A T C T G G C TCAAATTCACACCAATTTCA
...	...

Clonal Frame



Evolution of a single **locus** along a clonal frame by mutation (M) and recombination (R) events. A locus is a contiguous stretch of DNA – it will be represented by one column in an allele profile.

Allele types

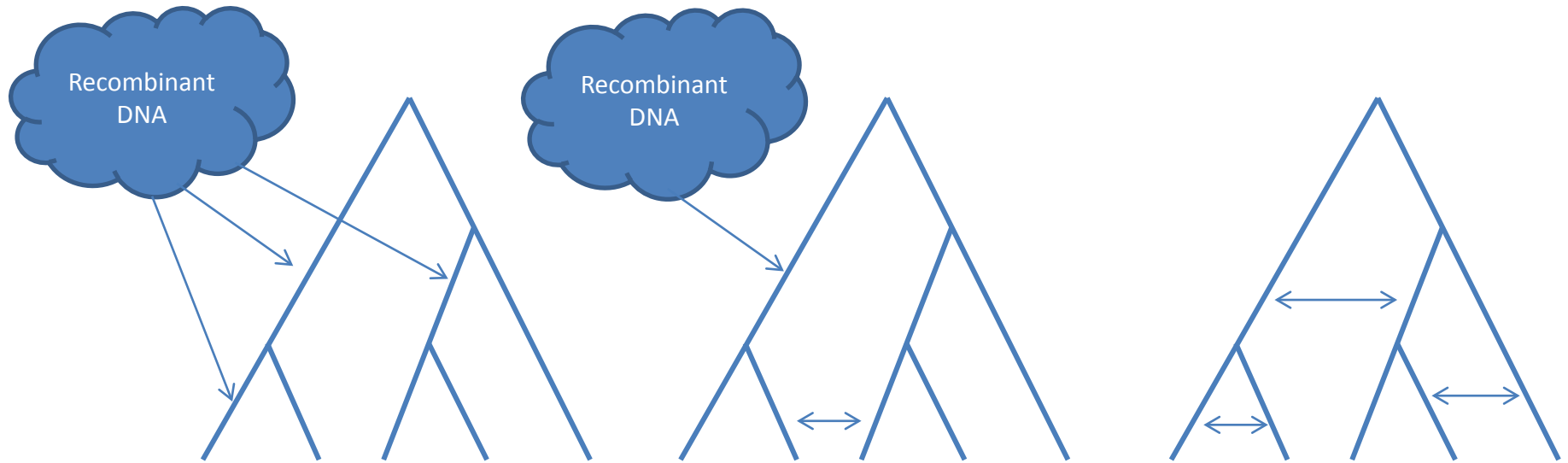
1 ACCG**ATAT**AGGAT**TCGTTCGT**CA
 2 ACCGTTGCAGGACTGCTAGCCA
 3 ACCGTTGCAGG**T**CTGCTAGCCA

Allele type 2 and 3 differ from each other in a single position due to a mutation event. Allele type 1 and 2 differ from each other in many positions due to a recombination event. This locus makes up a single column (bold) of the allele profile below.

Allele Profile

A 1**1**111...
 B 1**1**212...
 C 1**2**113...
 D 2**3**114...
 E 2**3**114...

A range of recombination models



(A) ClonalFrame model:
Recombination always
introduces novel genetic
material.

(B) Intermediate model



(C) ClonalOrigin model:
Recombination always
occurs within a closed
population.

Open system



Closed system

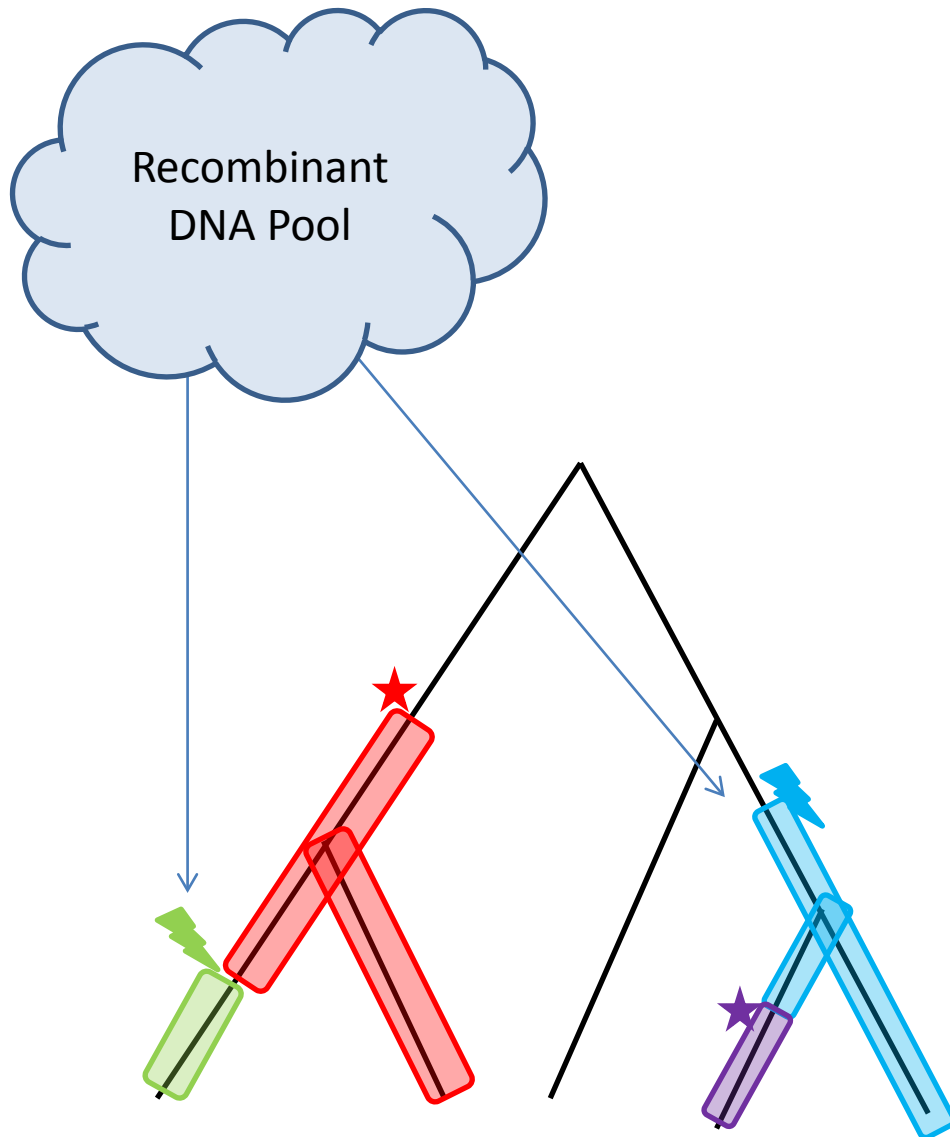
Clonal Frame model – Infinite Alleles Model?

A particular locus can undergo two types of events
mutation 
recombination 

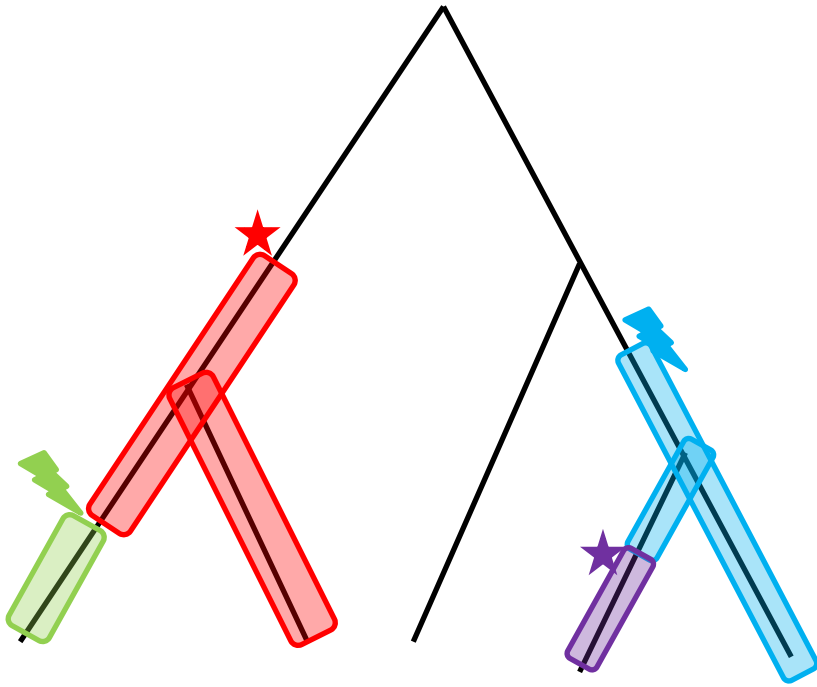
Parallel mutation should be infrequent as it requires

- 1) that the next mutation in the sequence for that locus occurs at the same site, i.e. without any other mutations occurring in the meantime $p \propto \frac{1}{L}$
- 2) And it further requires that the mutation is back to the initial state

Parallel recombination might be more likely, especially in a closed system. In an open system – as per the ClonalFrame model – parallel recombination should be even less likely than parallel mutation.

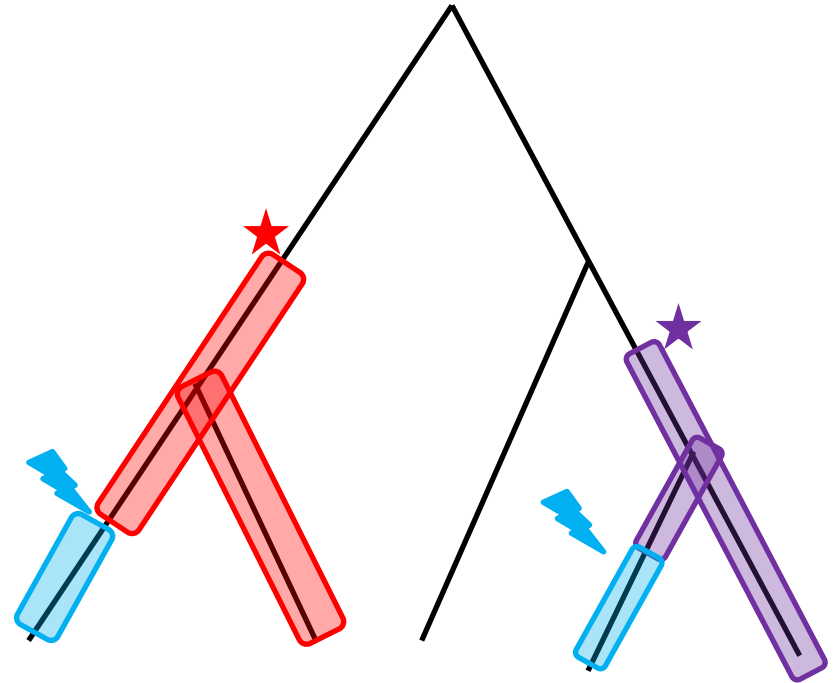


A compatible character



Loci that haven't undergone parallel recombination will produce a character (i.e. a column in the allele profile) that is compatible with the clonal frame.

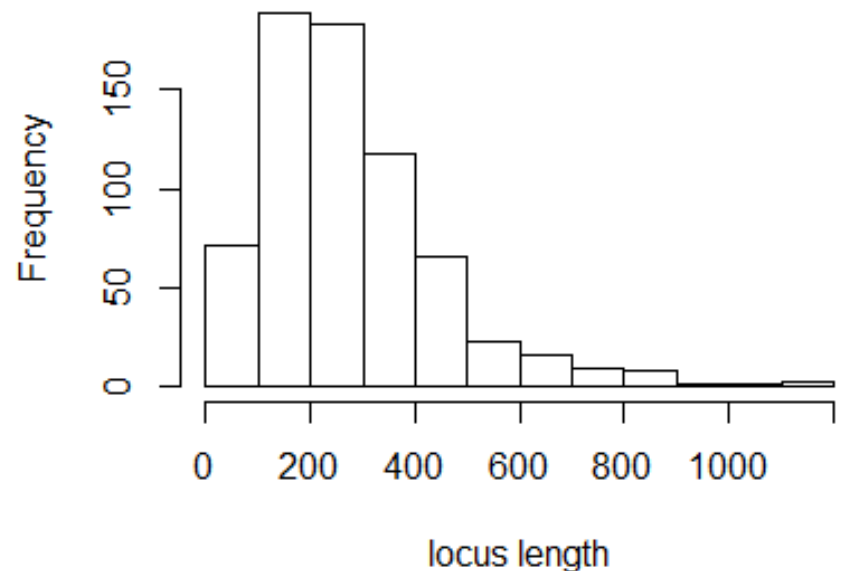
An incompatible character



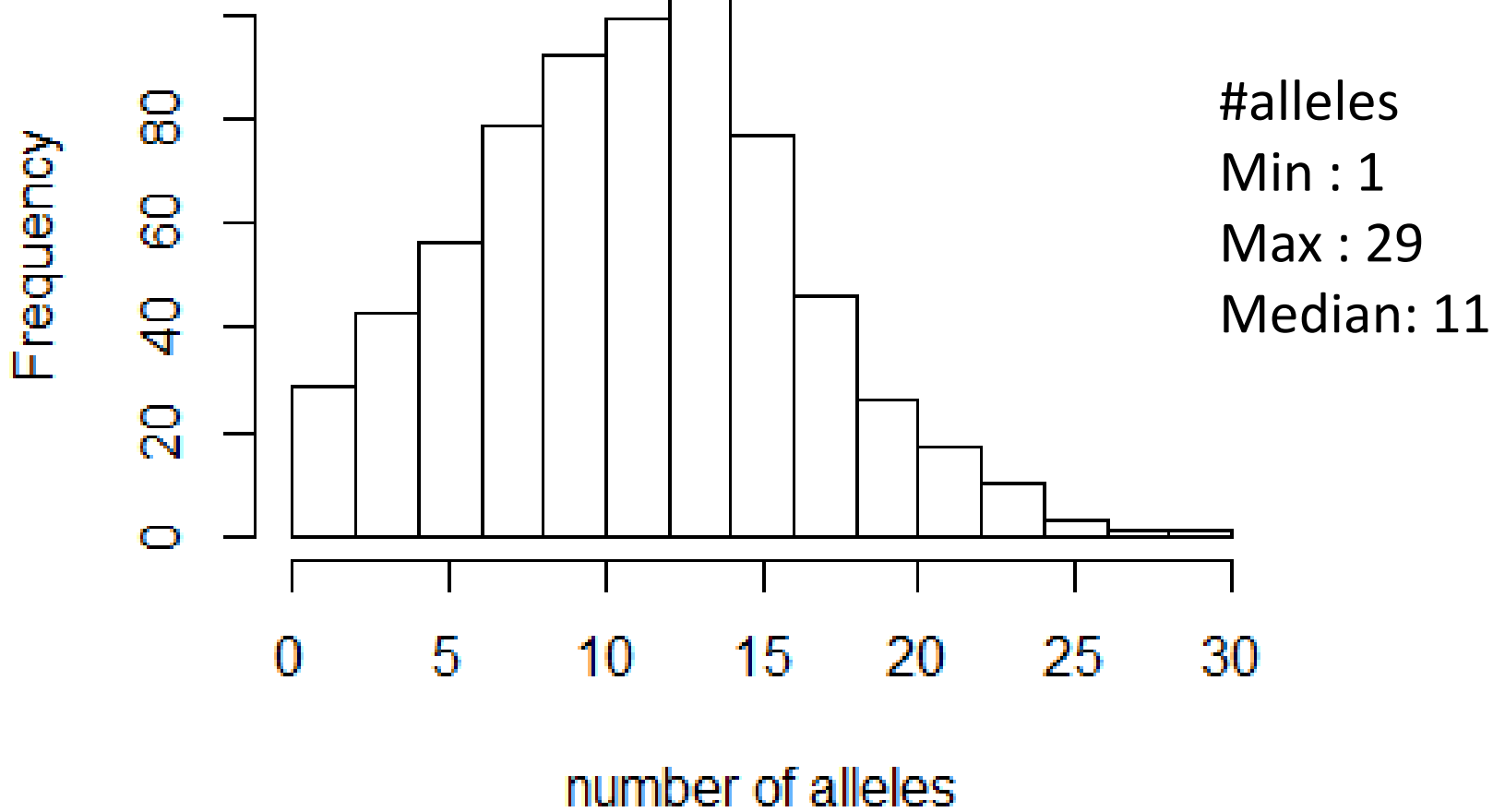
Blocks that have undergone parallel recombination (or parallel mutation) may produce characters that are not compatible with the clonal frame.

The *Campylobacter jejuni* data

- 46 *C. jejuni* genomes
- 686 genes in common across all 46 genomes
 - Total length over 686 alignments: 190595bp
 - min length: 44bp
 - max length: 1200bp
 - median length: 243.5 bp



Allele profile



Initial analysis

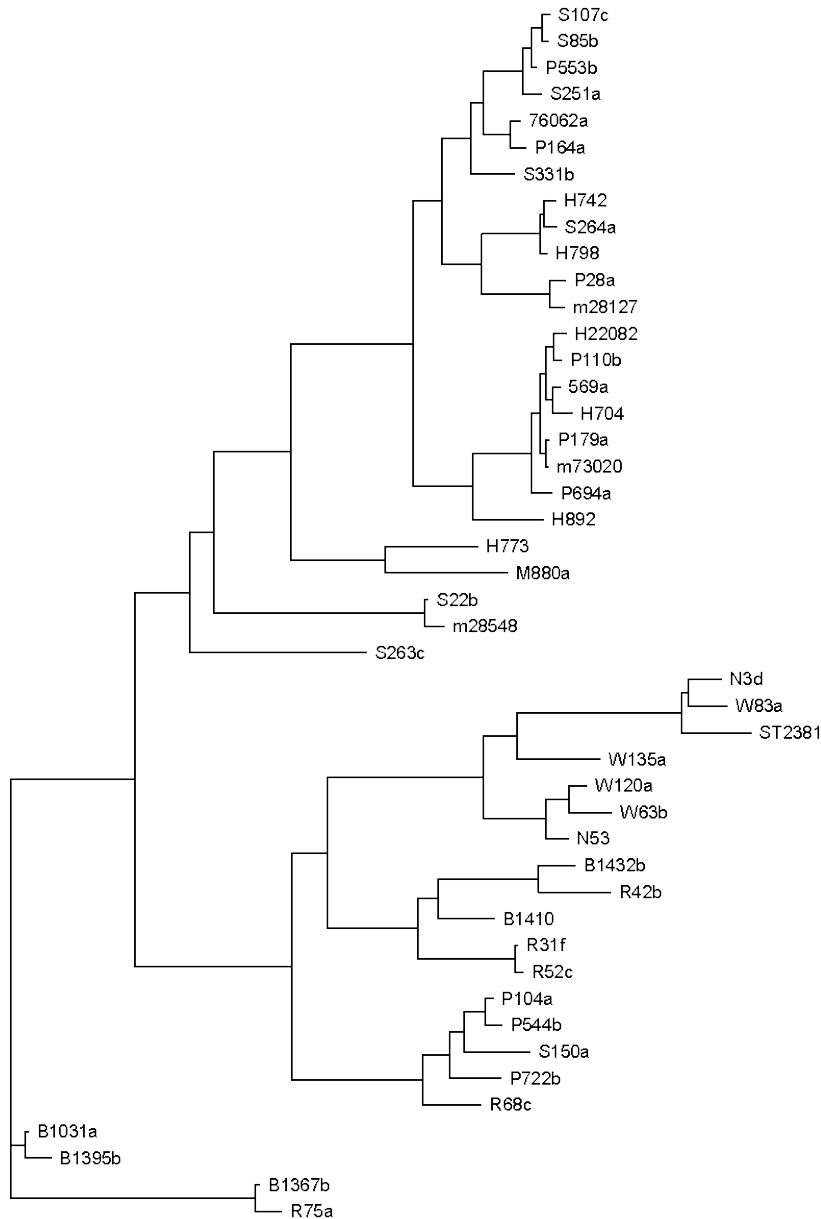
- 686 characters
- 9 constant, 2 parsimony uninformative
- Theoretical best parsimony score 7083

$$\sum_{l=1}^{686} (r_l - 1)$$

Where r_l is the number of alleles at locus l

- Parsimony finds 3 equally parsimonious trees with score 8274
- Consistency index 0.856

One of the Most Parsimonious Trees



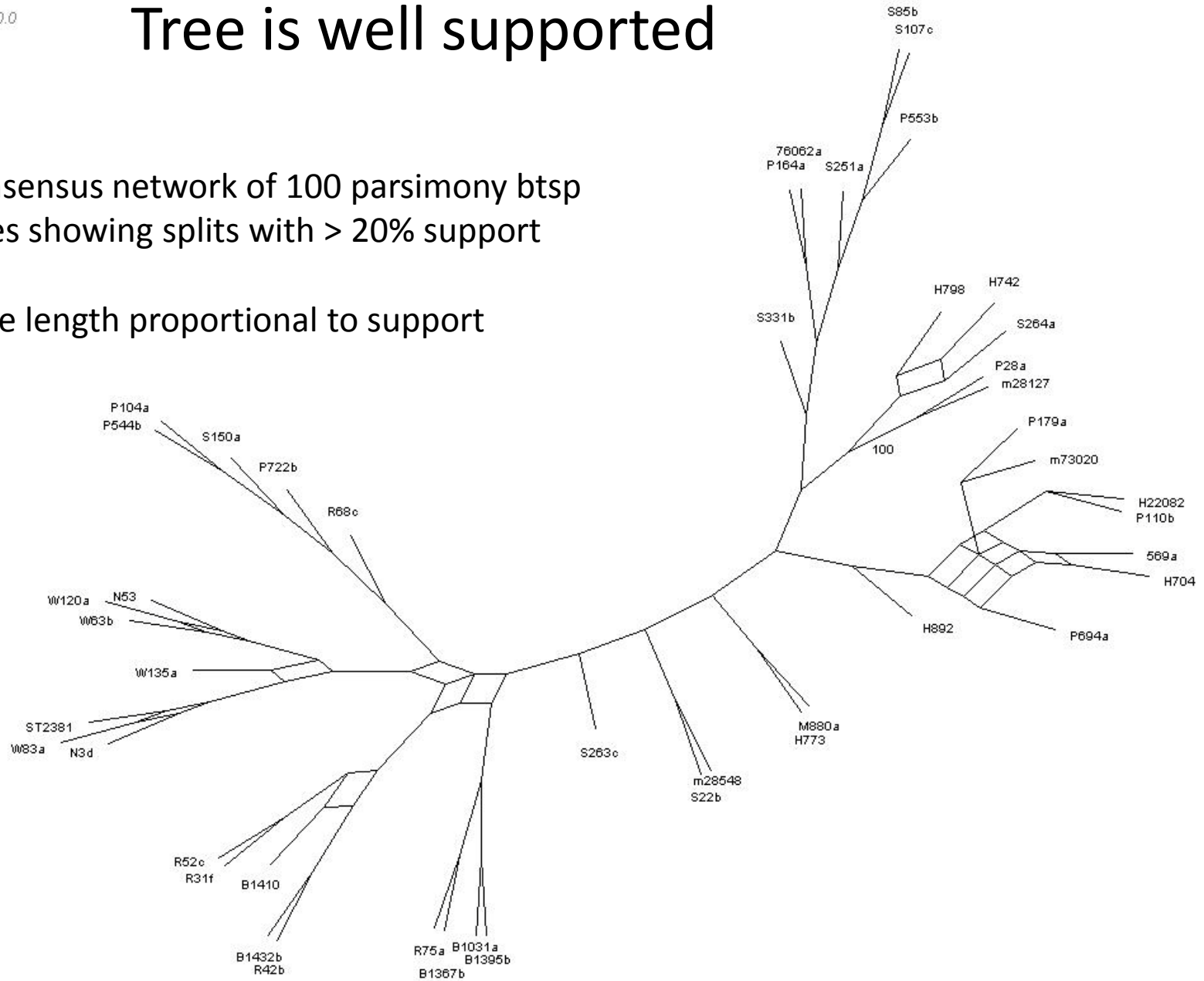
Why parsimony?

(Steel & Penny 2004)

Tree is well supported

Consensus network of 100 parsimony btsp trees showing splits with > 20% support

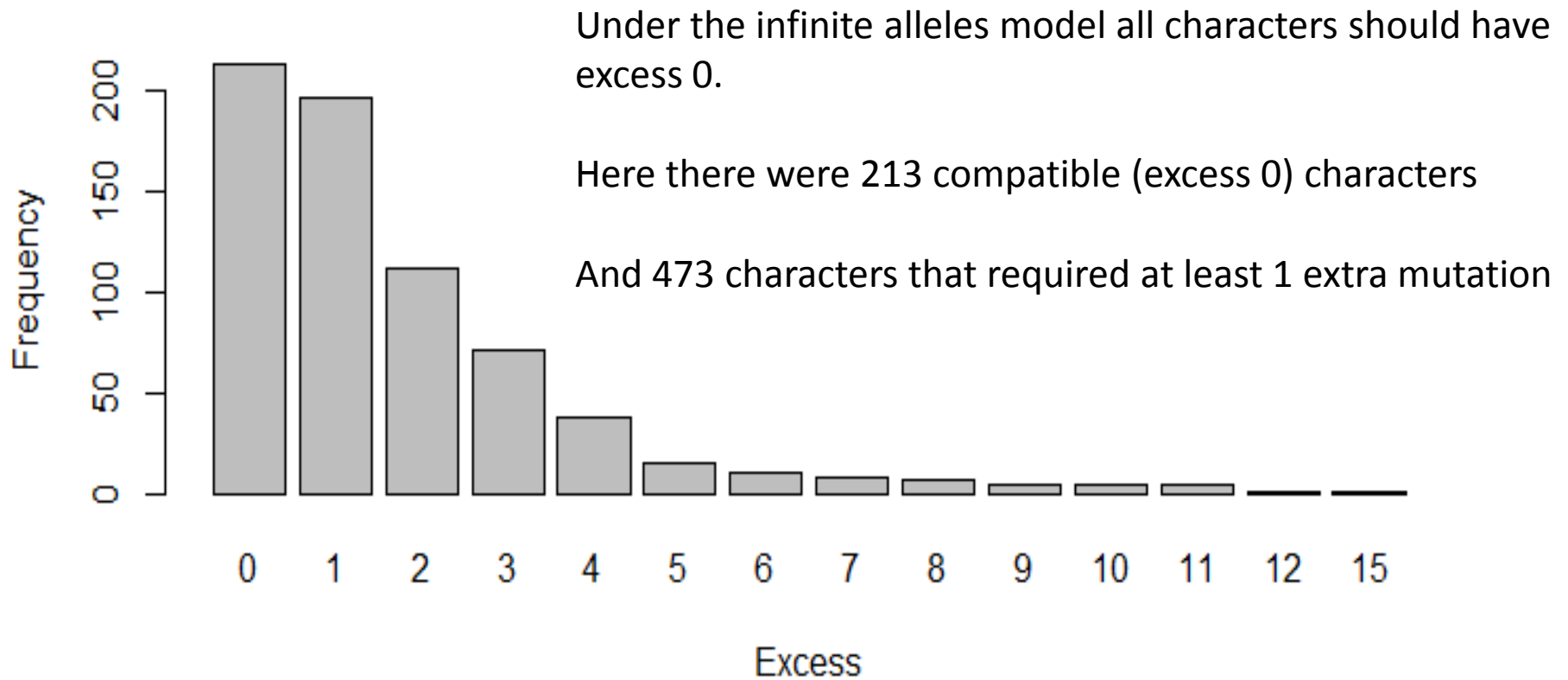
Edge length proportional to support



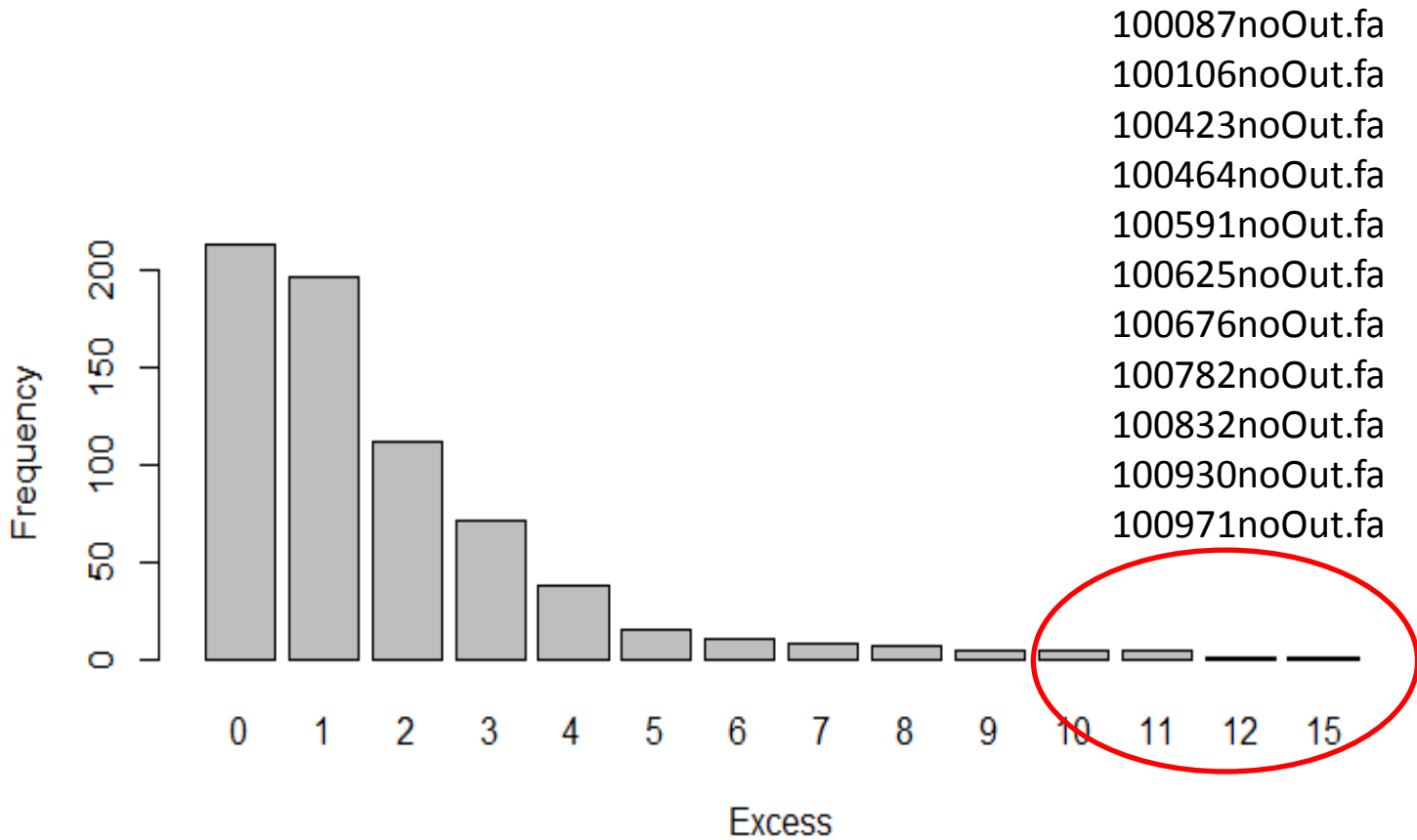
Excess

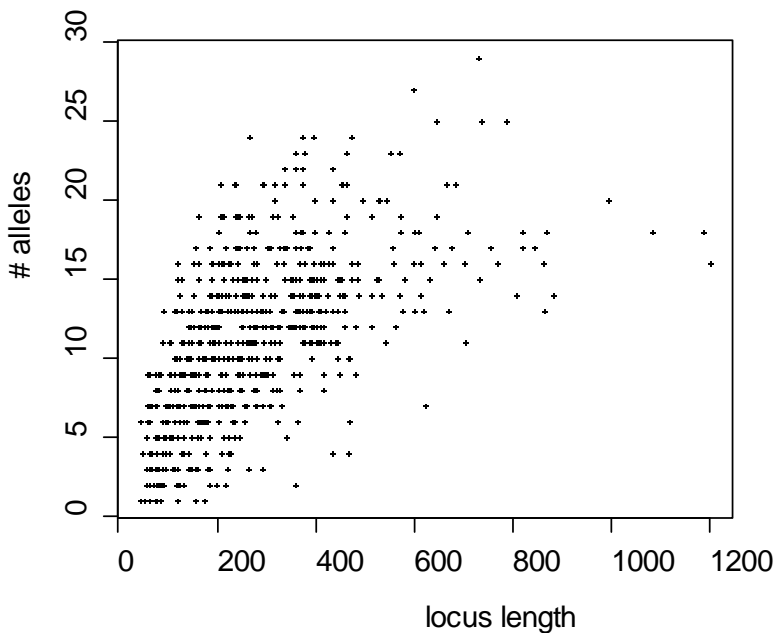
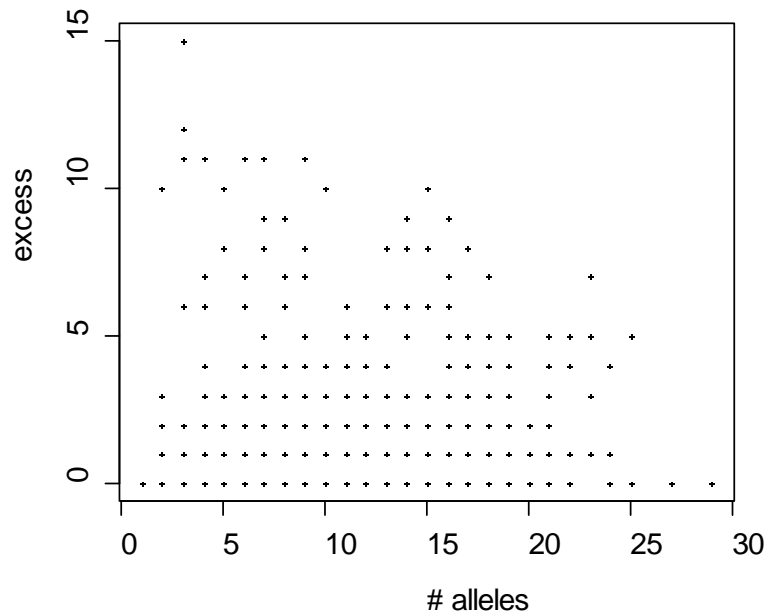
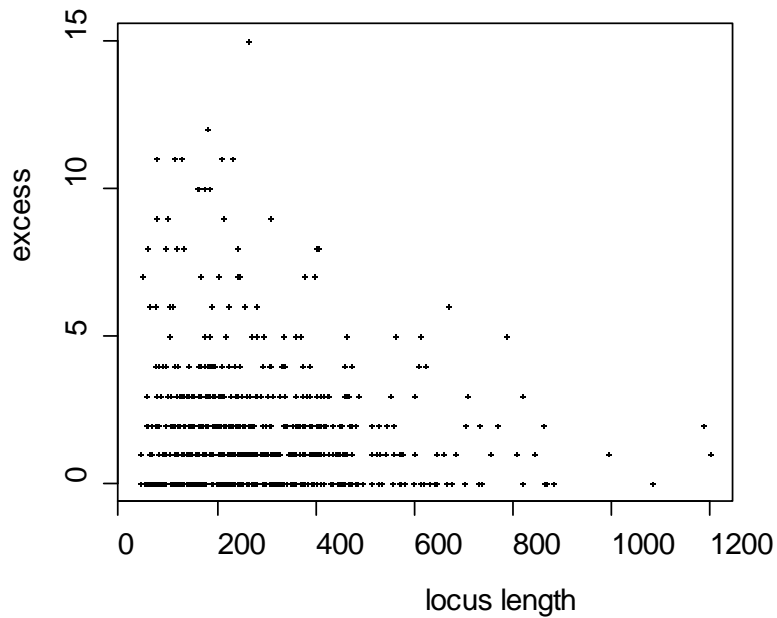
- The *Excess* of a character is its cost on the tree minus its minimum possible cost, or in other words it is the number of extra mutations required to explain the character.
- $\text{Excess} = \text{parsimony cost} - (r_l - 1)$
- A character is compatible with a tree if and only if it has excess zero on the tree.

Fit of characters in the Allele Profile to the Most Parsimonious Tree



Anything interesting about the ones with highest excess?





Call:

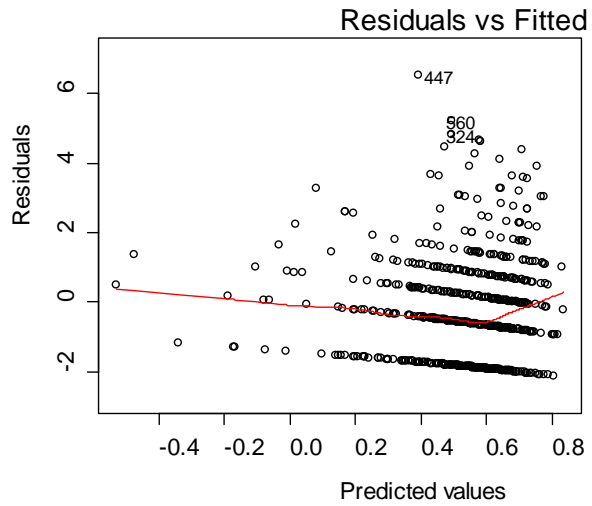
`glm(formula = excess ~ length + alleles, family = poisson())`

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.6637431	0.0710336	9.344	< 2e-16 ***
length	-0.0012571	0.0002396	-5.248	1.54e-07 ***
alleles	0.0197115	0.0069552	2.834	0.0046 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Diagnostics



Big cliques of compatible characters?

- Largest clique comprises the 213 characters that fit on the MPT
- Are there any larger than expected cliques of (pairwise) compatible characters amongst the rest of the characters?
- Have a look with a greedy algorithm
- [1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9, 10, 11, 12, 13, 14, 14, 14, 18, 22, 29, 32]

Work in progress

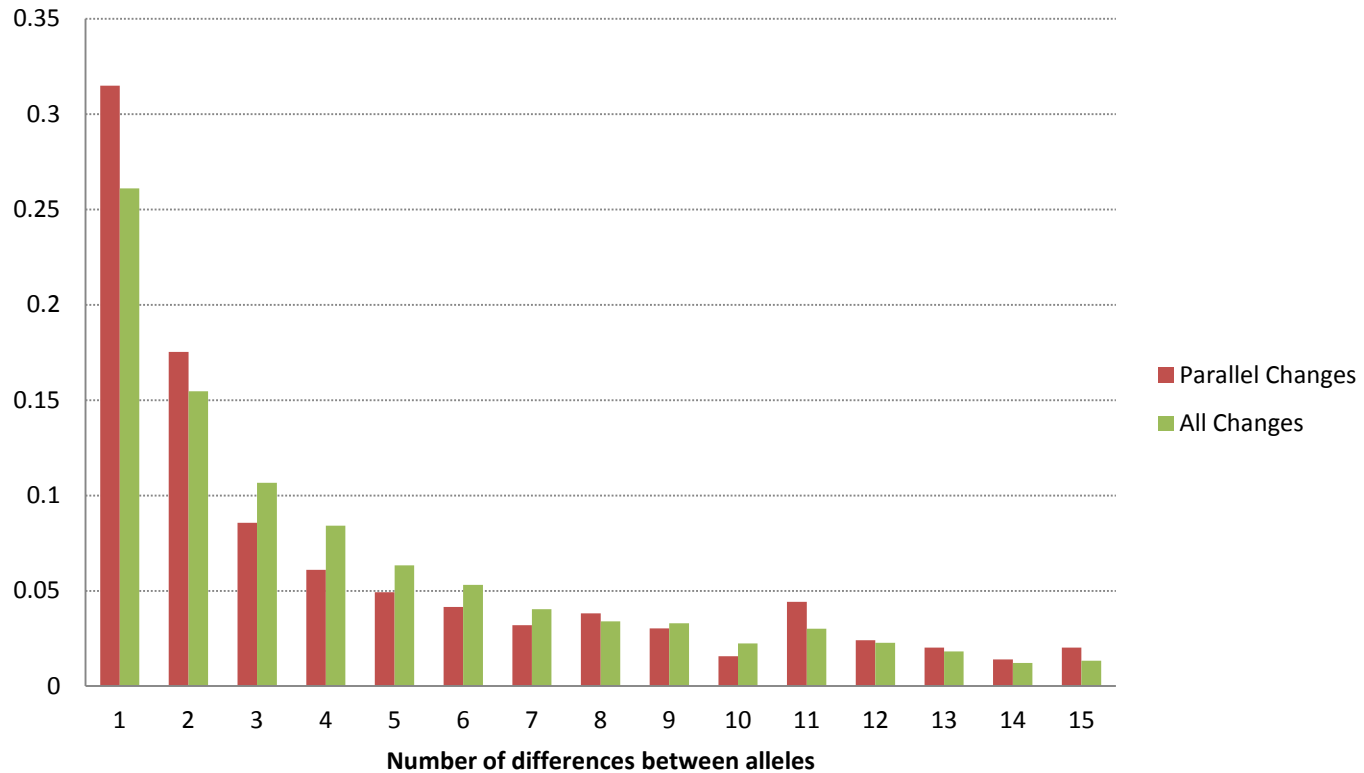
- In Holland et al 2010 (cormorants and shags) we developed a method to test if a clique of compatible characters is larger than expected by chance.
- We developed a “shuffle” that gave a random character with the same parsimony score as an unshuffled character.
- On further consideration, it seems better to create a random character with the same excess

Focus on transitions from allele to allele

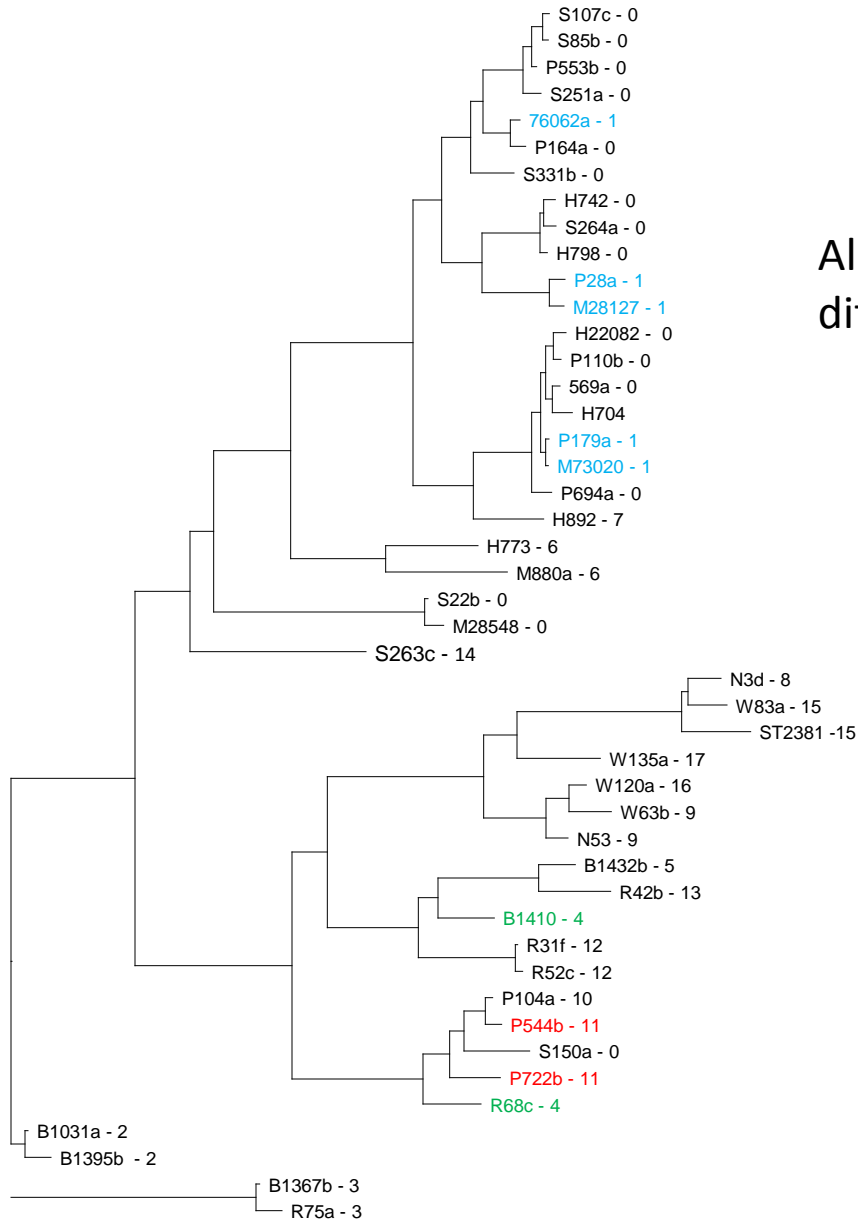
- Find the clonal frame by looking for the biggest clique of compatible characters.
- Use parsimony to work out all the transitions from one allele to another – look at the distribution of differences between pairs of alleles.
- For all the characters that are incompatible with the clonal frame – reconstruct their history on the clonal frame. Identify all the parallel transitions – look at the distribution of differences between pairs of alleles.

Parallel changes versus All changes

Relative frequencies of allele differences



Can identify clear cases of parallel recombination



Allele 0 and 1
differ at 20 sites

100185noOut.fa
18 alleles
Excess of 3

Conclusions

- Overall AP data is very consistent, i.e. highly compatible, consistency index > 0.85
- Some evidence that parallel mutation is more common than parallel recombination but both do occur.
- Clonal Frame wastes a lot of computational effort on finding the clonal frame but its model predicts (close to) perfect phylogenies.

Future work

- Use simMLST to create data under a mutation only model and compare levels of compatibility.
- Improve statistical test of whether or not cliques of compatible characters are bigger than expected by chance.
- Better way to do ancestral state reconstruction than parsimony? i.e. something that accounts for branch lengths.

Acknowledgements

- Nigel French, Patrick Biggs, Shoukai Yu