Engineering cytochromes P450 from ancestral predictions using the novel tool GRASP

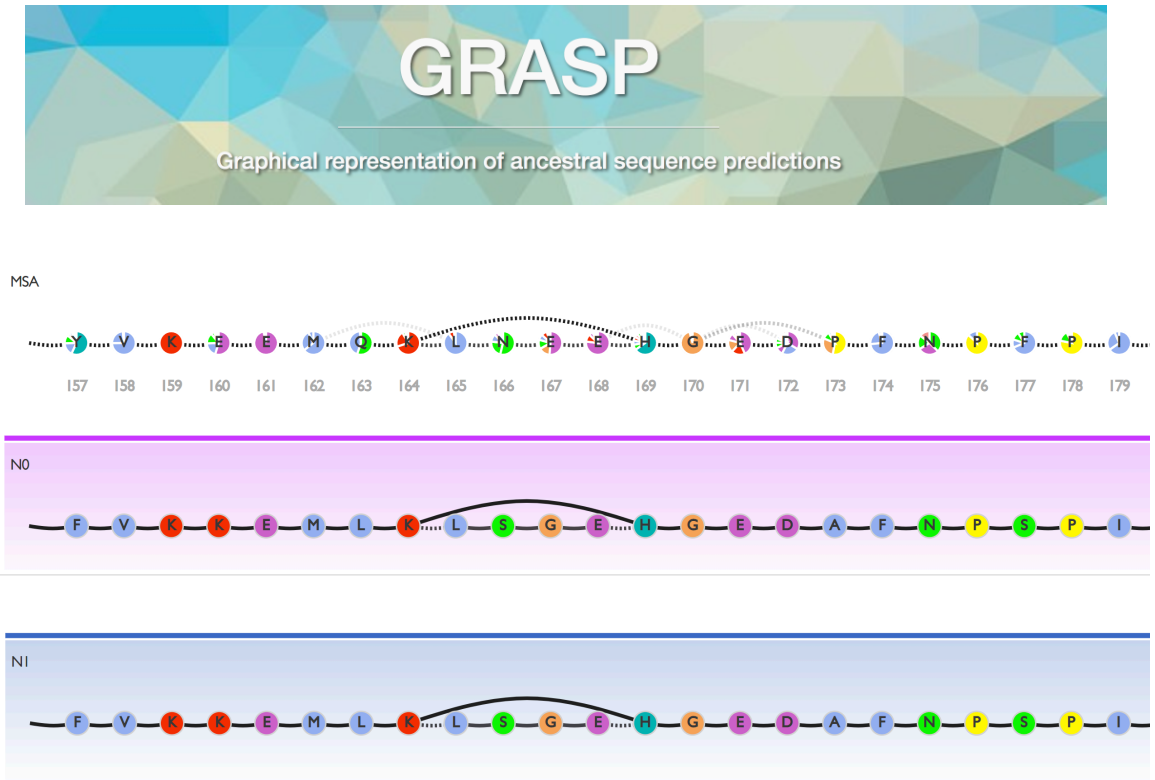
Gabe Foley

School of Chemistry and Molecular Biosciences

The University of Queensland

27/06/2019

Graphical representation of ancestral sequence predictions (GRASP)



- Ancestral sequence reconstruction (ASR) tool designed for large data sets
- Successful reconstruction of cytochromes P450 in collaboration with Liz Gillam at UQ

http://grasp.scmb.uq.edu.au

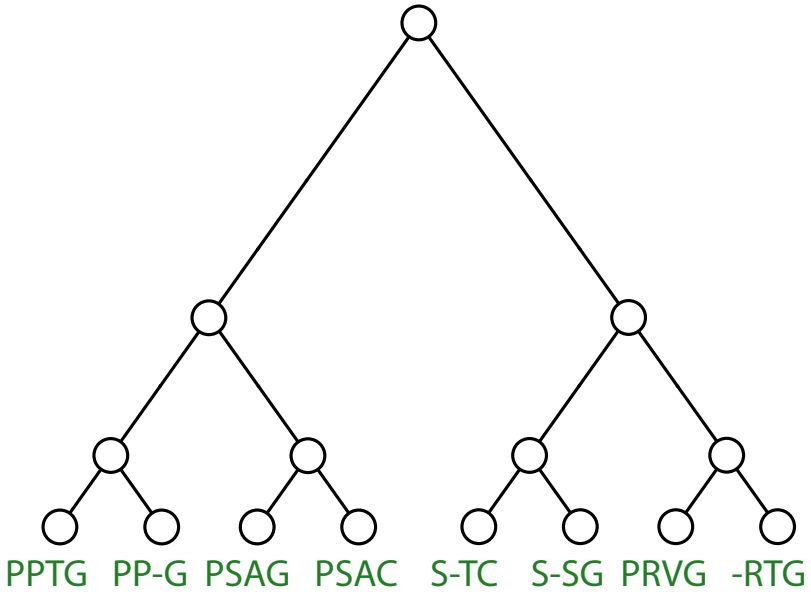
Overview

- *What* is ancestral sequence reconstruction (ASR) ?
- *Why* use it?
- ASR on big data

- How **GRASP** enables big data and extends the reach of ASR

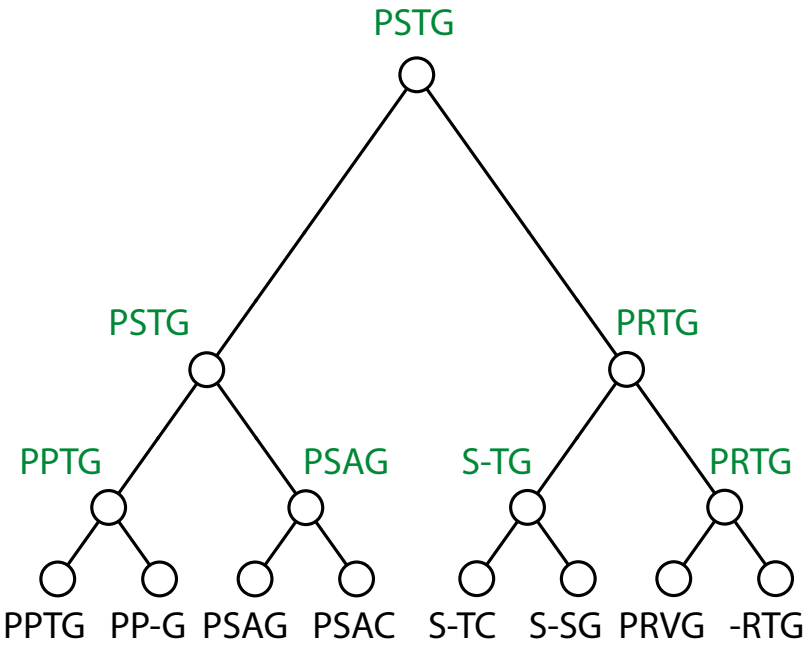
What is ancestral sequence reconstruction?

- Using the information in modern day biological sequences to infer what their ancestors looked like



What is ancestral sequence reconstruction?

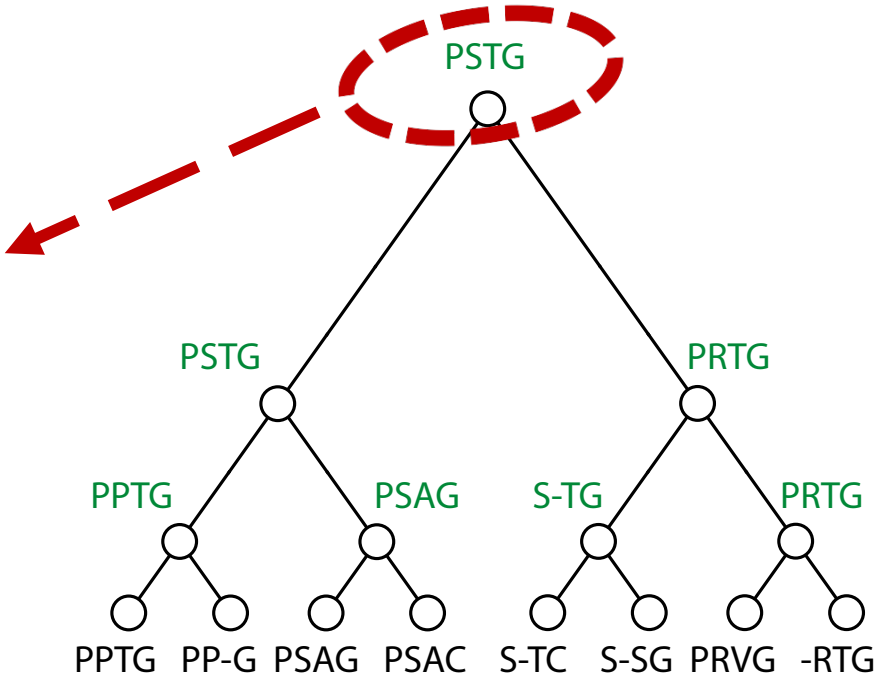
- Using the information in modern day biological sequences to infer what their ancestors looked like



What is ancestral sequence reconstruction?

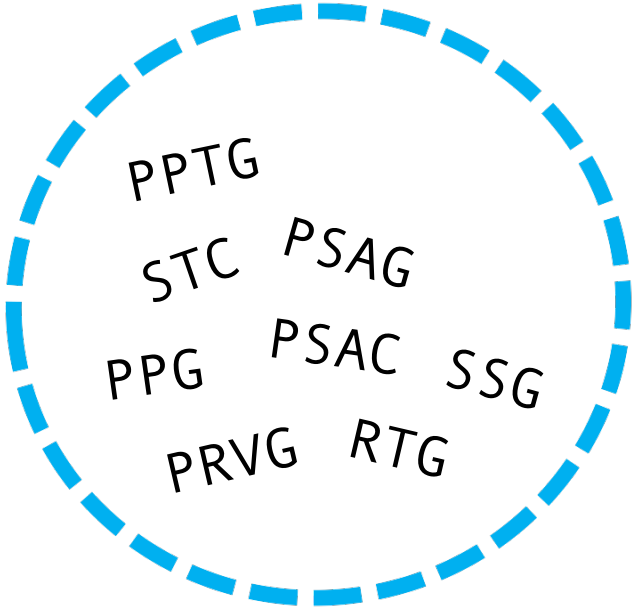
- Using the information in modern day biological sequences to infer what their ancestors looked like

- Ancestral sequences can be ‘resurrected’ – synthesised and studied alongside modern day proteins

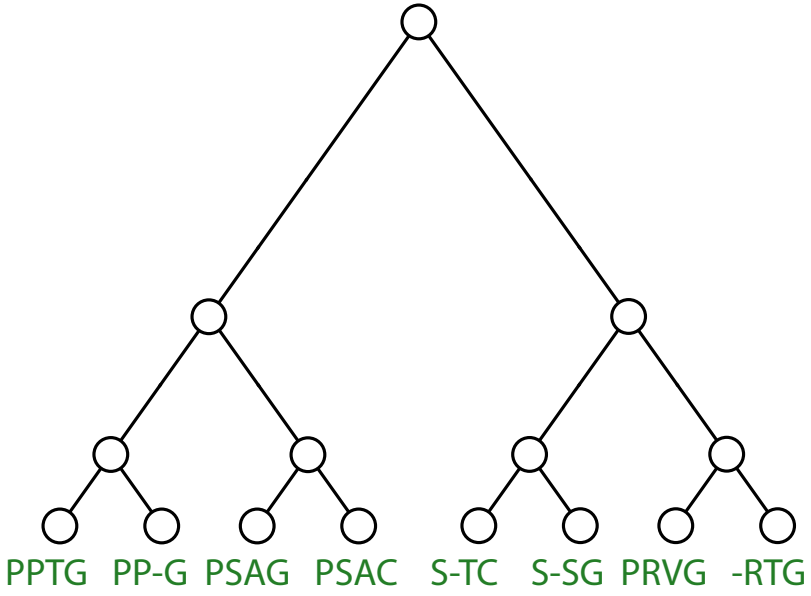


Ancestral sequence reconstruction steps

1. Collect sequences



PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

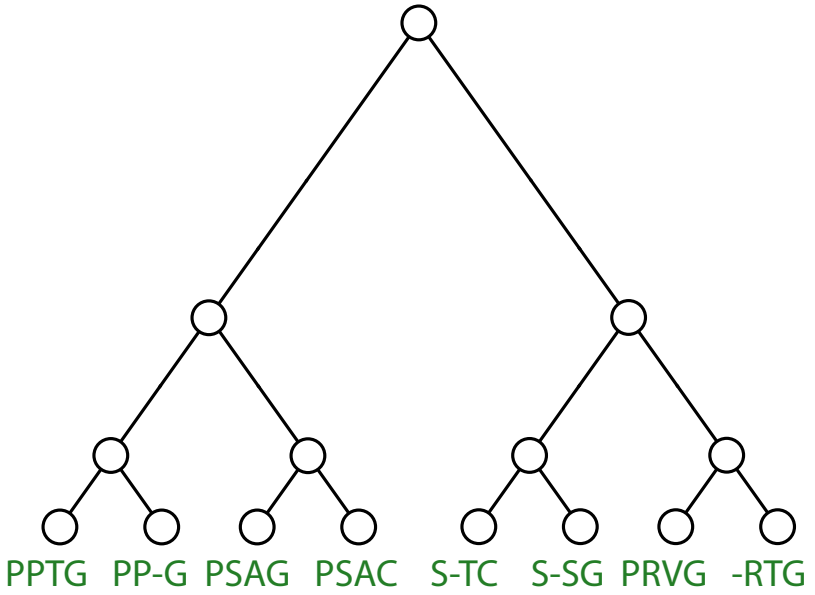


Ancestral sequence reconstruction steps

2. Align sequences

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

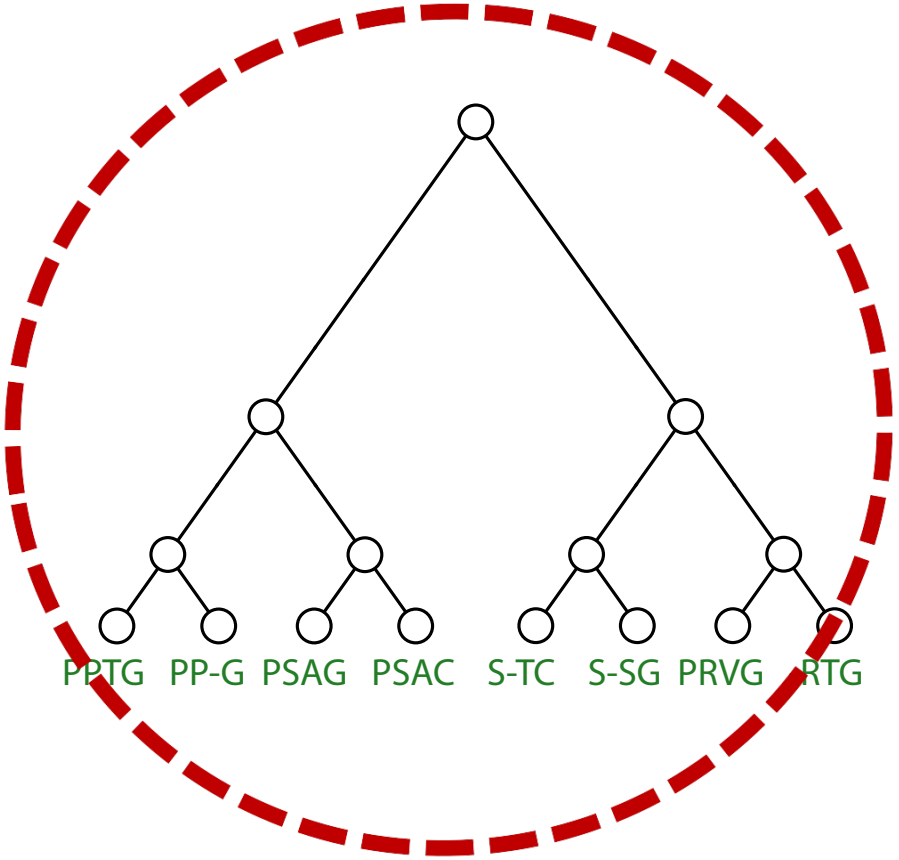


Ancestral sequence reconstruction steps

3. Infer phylogenetic tree

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

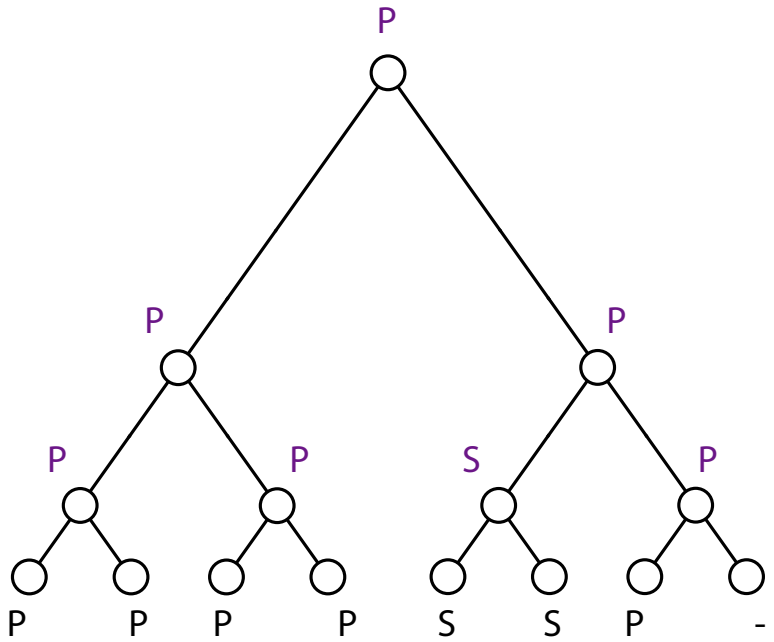


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

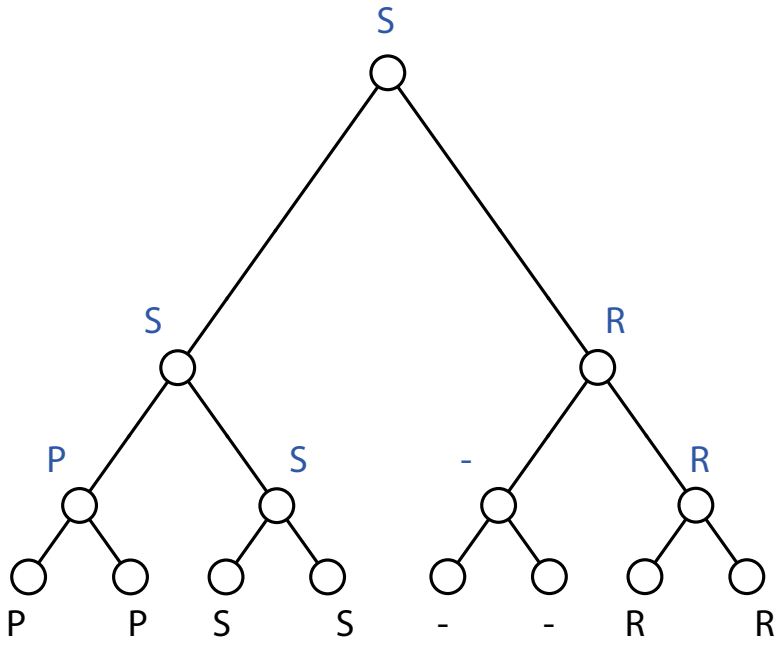


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

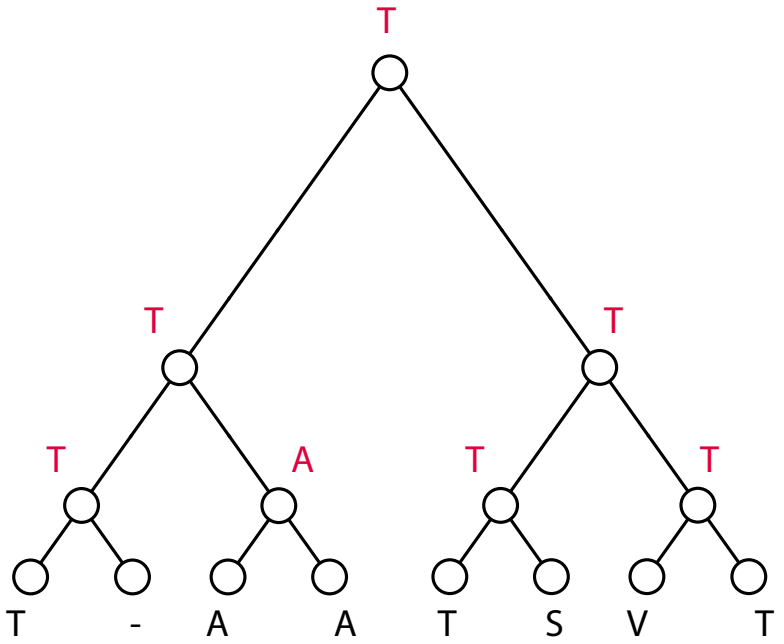


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG



Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

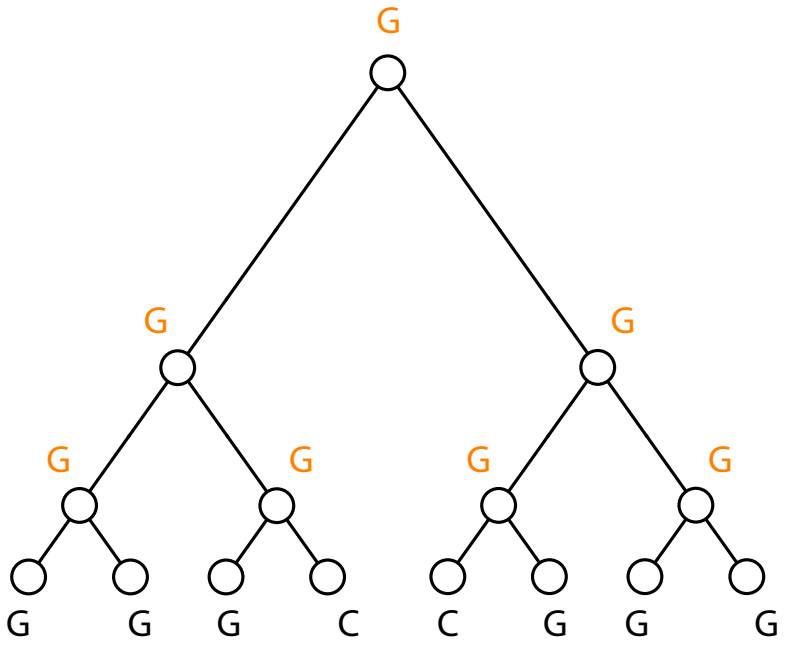
PPTG

STC PSAG

PPG PSAC SSG

PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

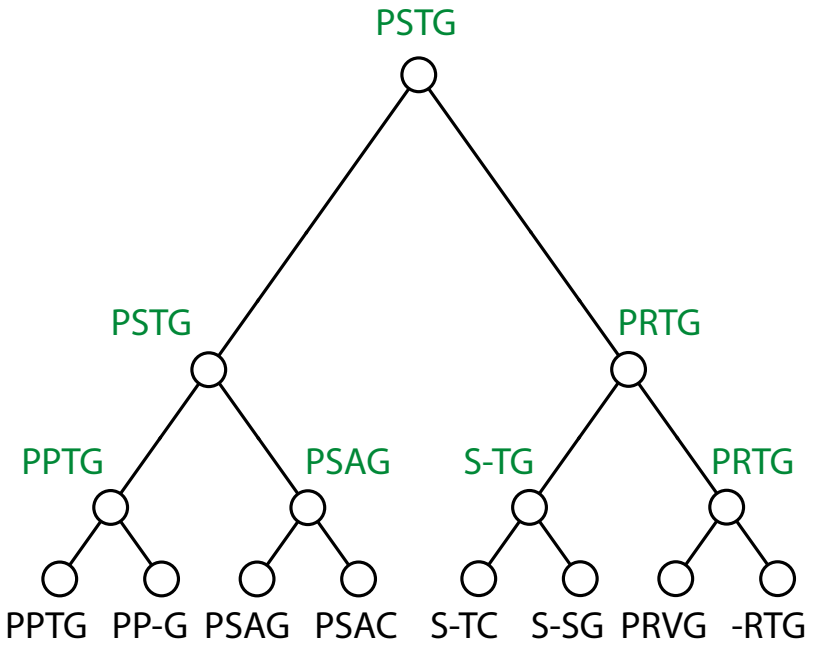


Ancestral sequence reconstruction steps

5. Concatenate predictions into a complete sequence

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

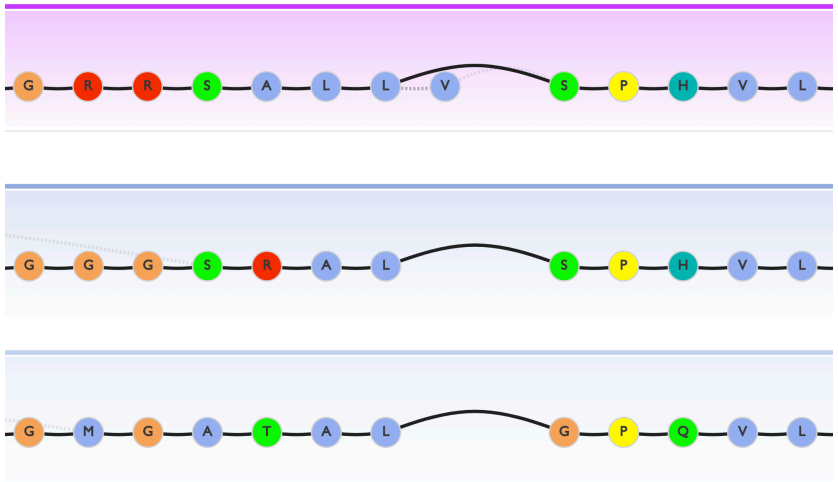
PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

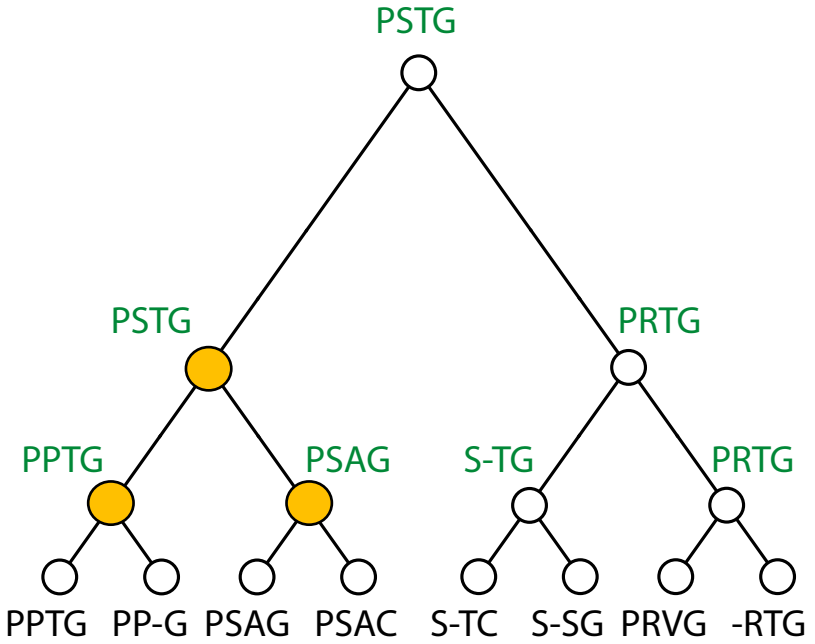


Ancestral sequence reconstruction steps

Joint reconstruction

Infer predictions for all ancestors



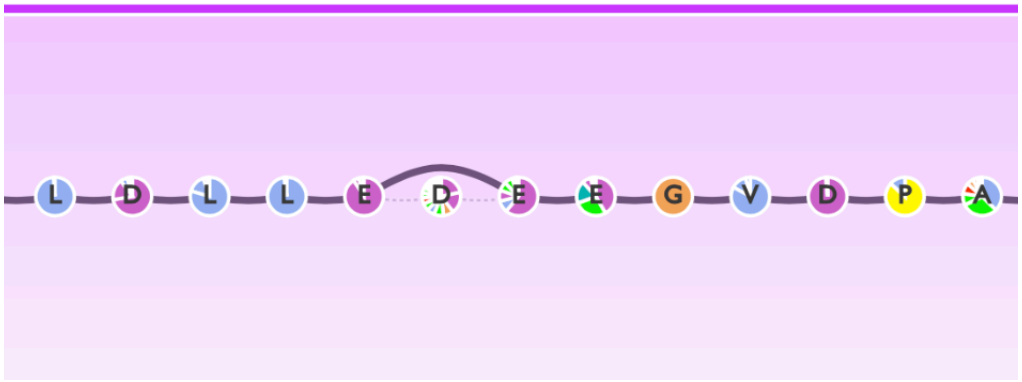


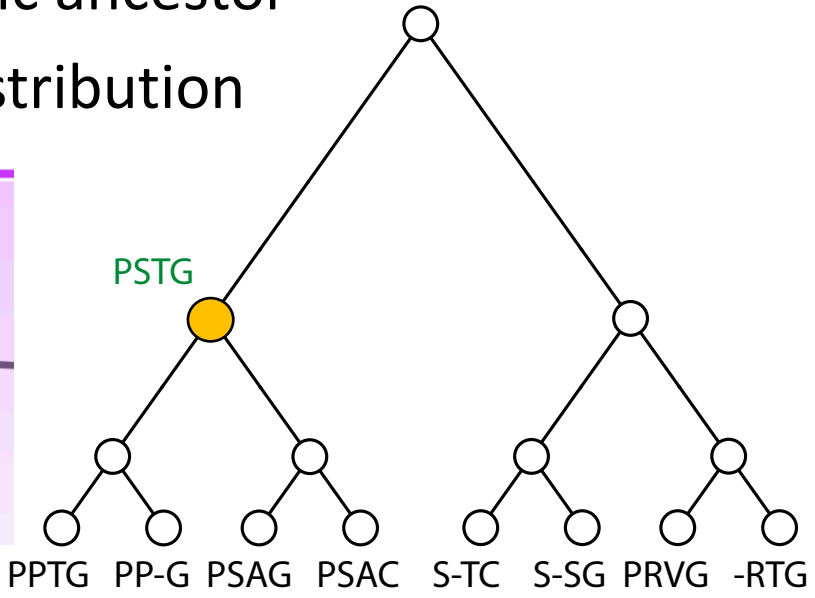
Ancestral sequence reconstruction steps

Marginal reconstruction

Center prediction around a specific ancestor

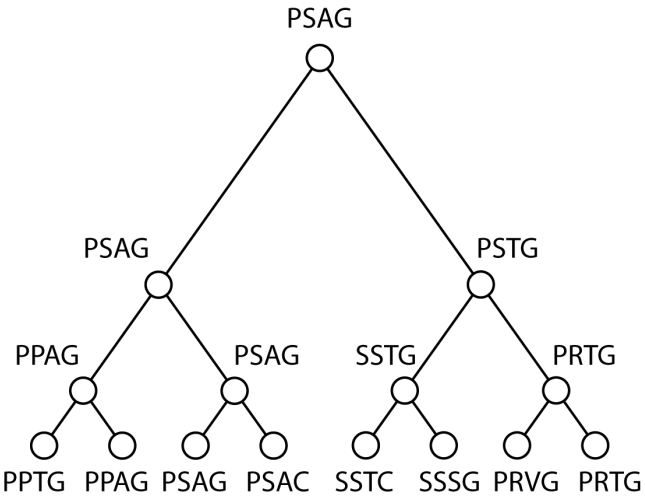
Each position has a probability distribution





Why use ancestral sequence reconstruction?

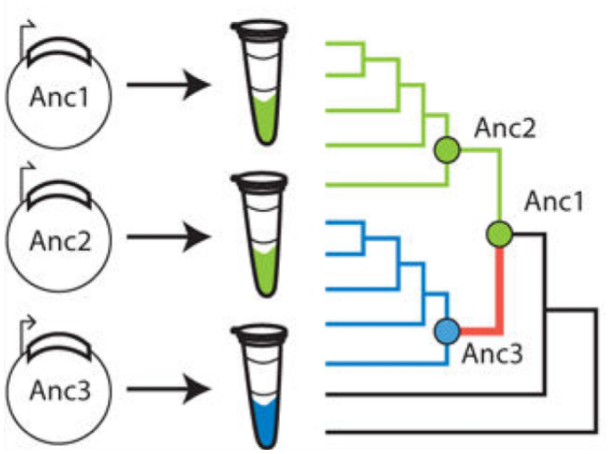
- **Studying evolutionary histories**
- Determining important functional residues
- Engineering ancestors from templates
- Constructing novel sequences



A reconstructed history of character states

Why use ancestral sequence reconstruction?

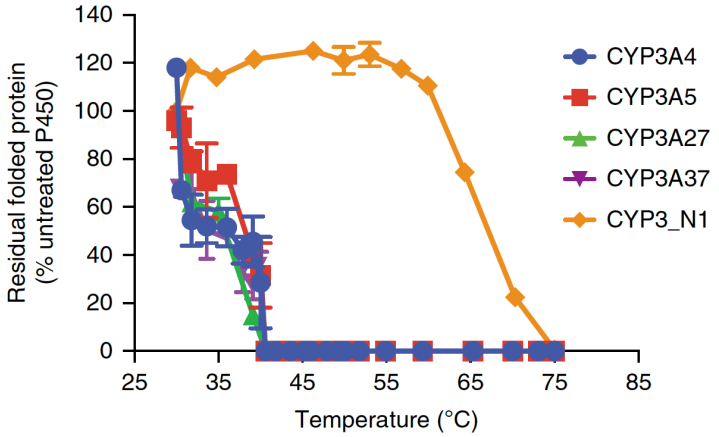
- Studying evolutionary histories
- **Determining important functional residues**
- Engineering ancestors from templates
- Constructing novel sequences



Adapted from Hochberg & Thornton**,** *Annu Rev Biophys* **46**, 247–269 (2017)

Why use ancestral sequence reconstruction?

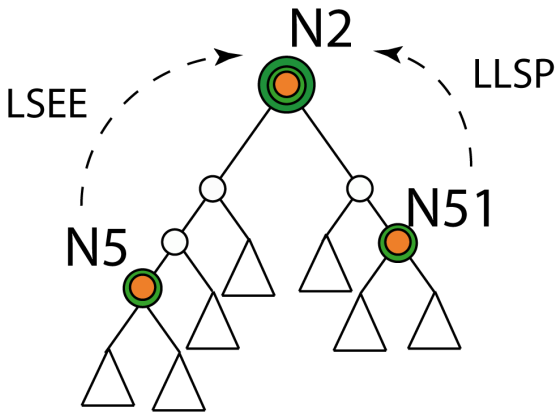
- Studying evolutionary histories
- Determining important functional residues
- **Engineering ancestors from templates**
- Constructing novel sequences



Adapted from Gumulya et al., *Nature Catalysis* **1**, 878 (2018).

Why use ancestral sequence reconstruction?

- Studying evolutionary histories
- Determining important functional residues
- Engineering ancestors from templates
- **Constructing novel sequences**

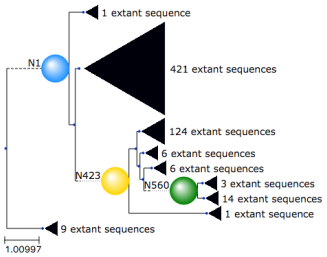


Successfully reconstructed CYP2U1 variants

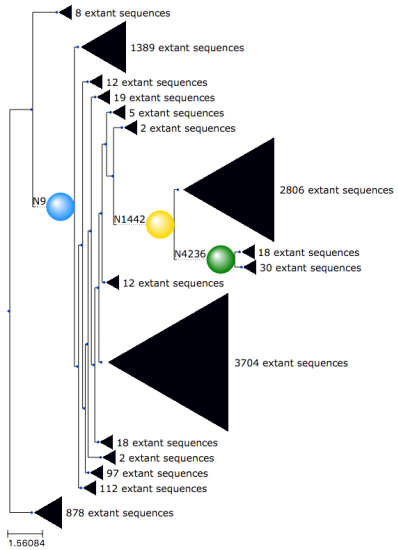
ASR in the era of big data

- Better coverage increases robustness of predictions
  - Enables us to classify allowable variation
- Incorporation of distant homologs can allow us to infer further back in evolutionary time
- Ancestral data sets become rich sources of information which can be mined and studied

Large data sets approach a canonical form of ancestor

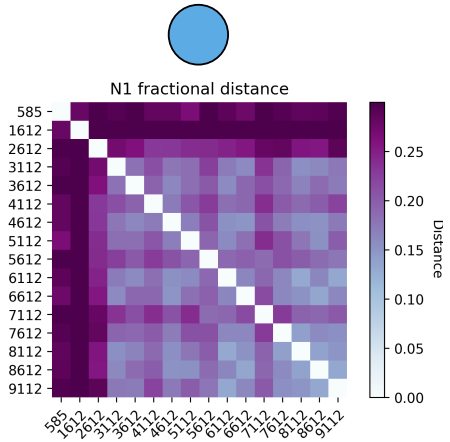


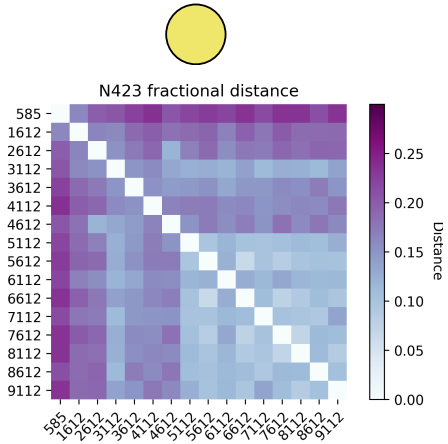
585 sequences

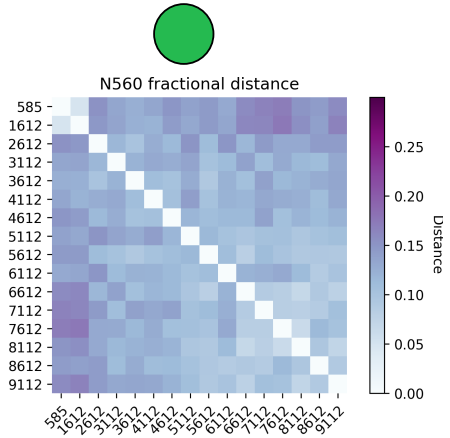


9112 sequences

Dihydroxy-acid dehydratase data set
Fractional distances between different ancestors

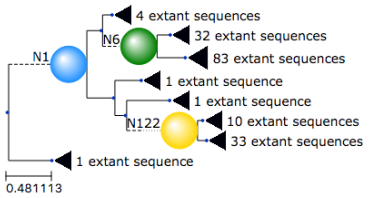




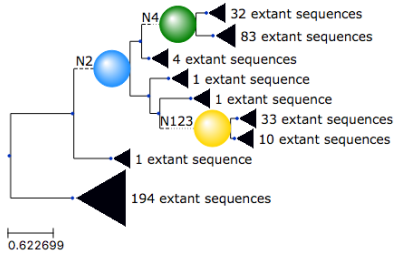


- Increasing sequence count mean predictions approach canonical forms
- Ancestors closer to extants are less affected

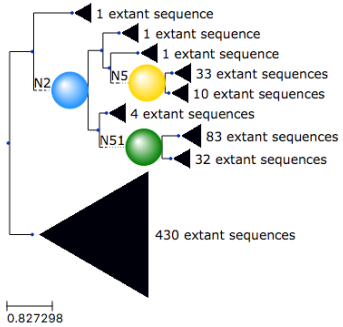
Large data sets approach a canonical form of ancestor



CPY2U1: 165 sequences

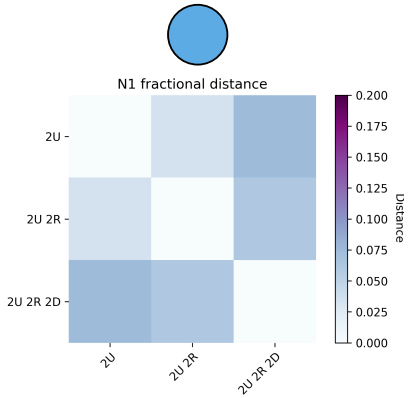


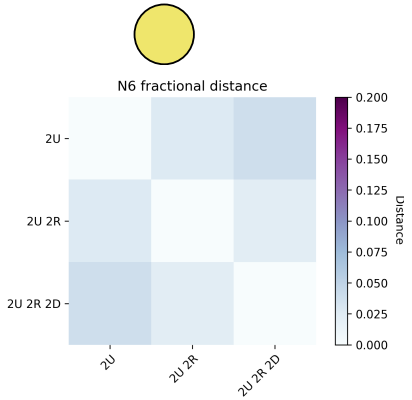
CYP2U1 / CYP2R1 : 359 sequences

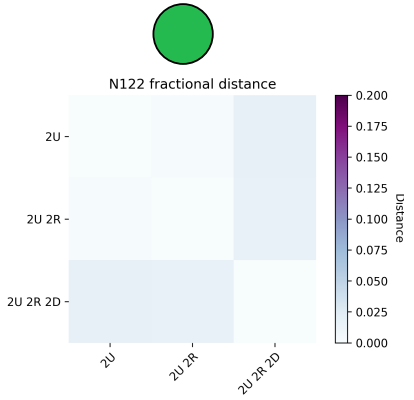


CYP2U1 / CYP2R1 / CYP2D: 595 sequences

CYP2U1 / CYP2R1 / CYP2D data set
Fractional distances between different ancestors







- Increasing sequence count mean predictions approach canonical forms
- Ancestors closer to extants are less affected

ASR – challenges with big data

- Processing large data sets takes a long time or is impossible
  - Current tools, FastML and PAML capable of ~500 - 600 sequences
- Increased presence of insertions and deletions
  - Increases alignment length and must be dealt with in order to predict sensible ancestors
- Extracting information is a much harder tasker
  - More alternatives exist, scale of data is harder to examine

GRASP – solutions for big data

- Processing large data sets takes a long time or is impossible
  - GRASP is capable of inferring data set sizes of ~9000
- Increased presence of insertions and deletions
  - GRASP uses partial order graphs to discretely model insertion and deletion events
- Extracting information is a much harder tasker
  - GRASP is an interactive tool built for exploration, with annotations, mutant suggestions, and motif searching

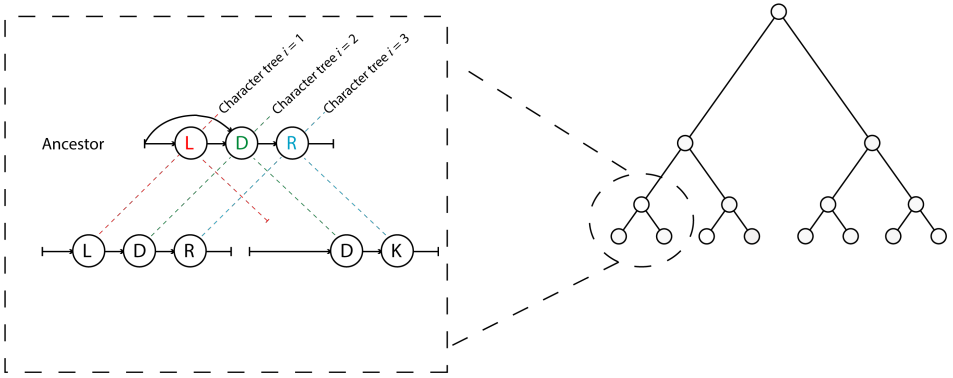
Processing large data sets

- **Data structure** is a Bayesian network and we use variable elimination for efficient inference
- **Inference algorithm** is equivalent to FastML or PAML

INFERENCE STEPS

1. Calculate all possible state
2. Calculate a consensus path

- Importantly, we can dynamically process these on demand



359 sequences

Tool	Run time (full output)	Run time (selected output)
GRASP	3 min	1 min 30 seconds
FastML	8 hours	Not possible
PAML	13 hours	Not possible

1529 sequences

Tool	Run time (full output)	Run time (selected output)
GRASP	1 hour 5 mins	9 min

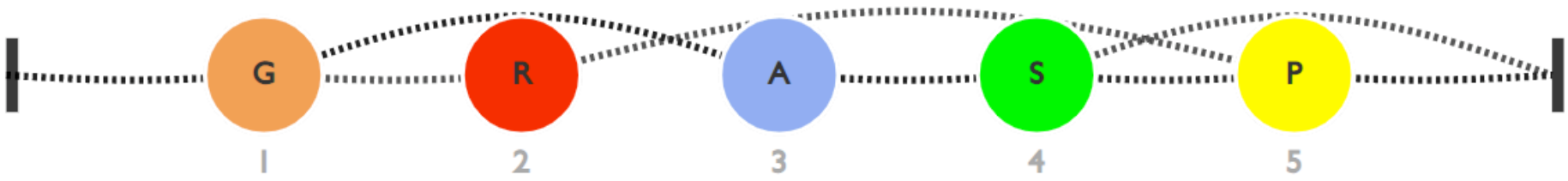
9112 sequences

Tool	Run time (full output)	Run time (selected output)
GRASP	~ 7 days	~ 1 day

Modelling indels with partial order graphs

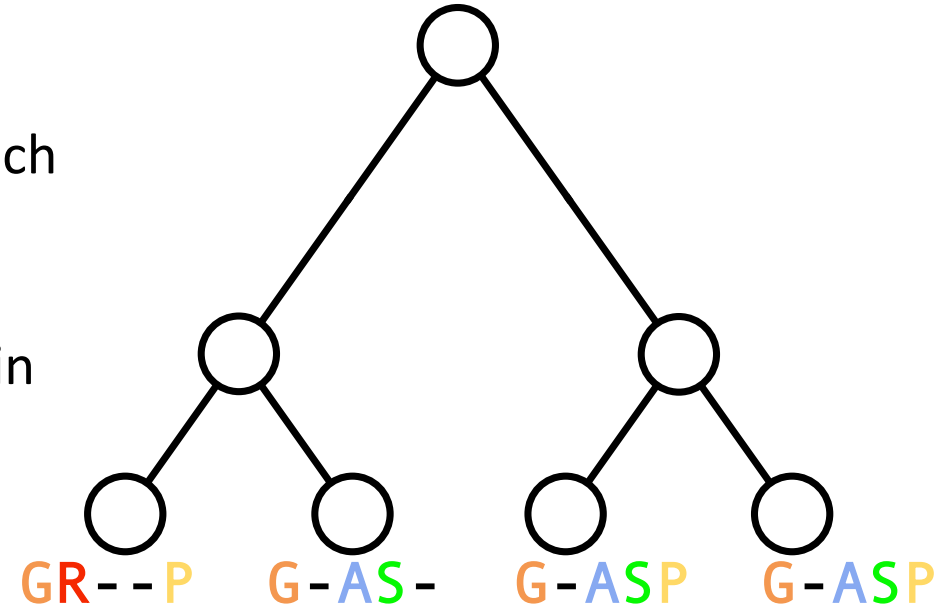
- Represents ambiguity
- Summarises indel events as edges on a graph

GR--P
G-AS-
G-ASP
G-ASP

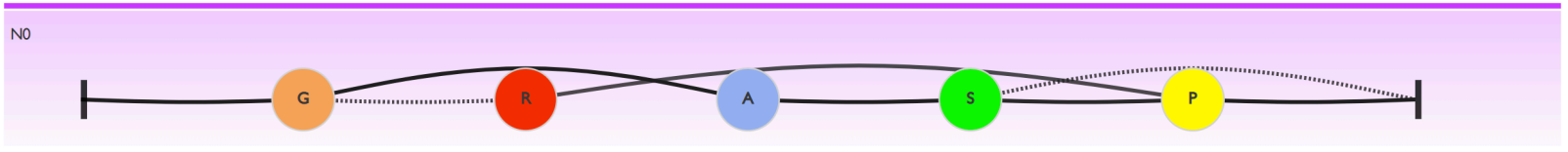


Inferring a consensus path

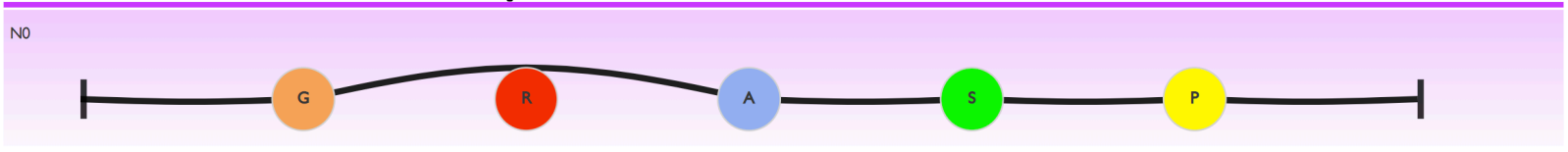
- Parsimony is used to score each out edge *and* each in edge
- Edges that are parsimonious in both directions are preferred



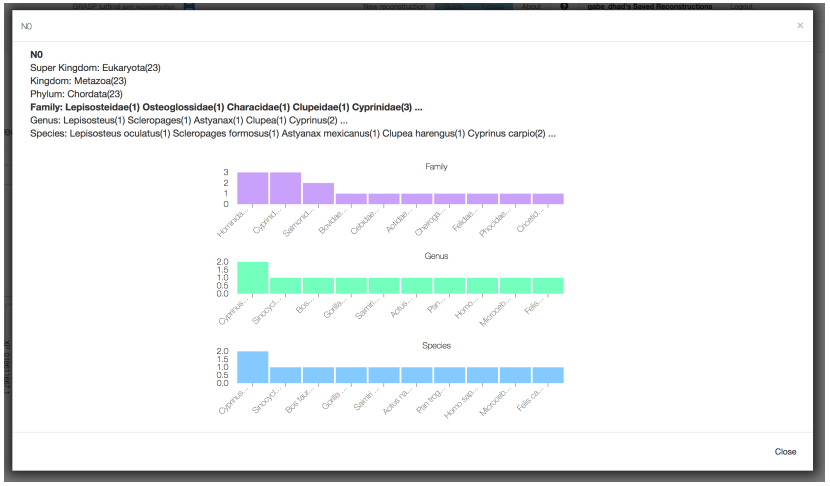
Ancestor with alternative pathways



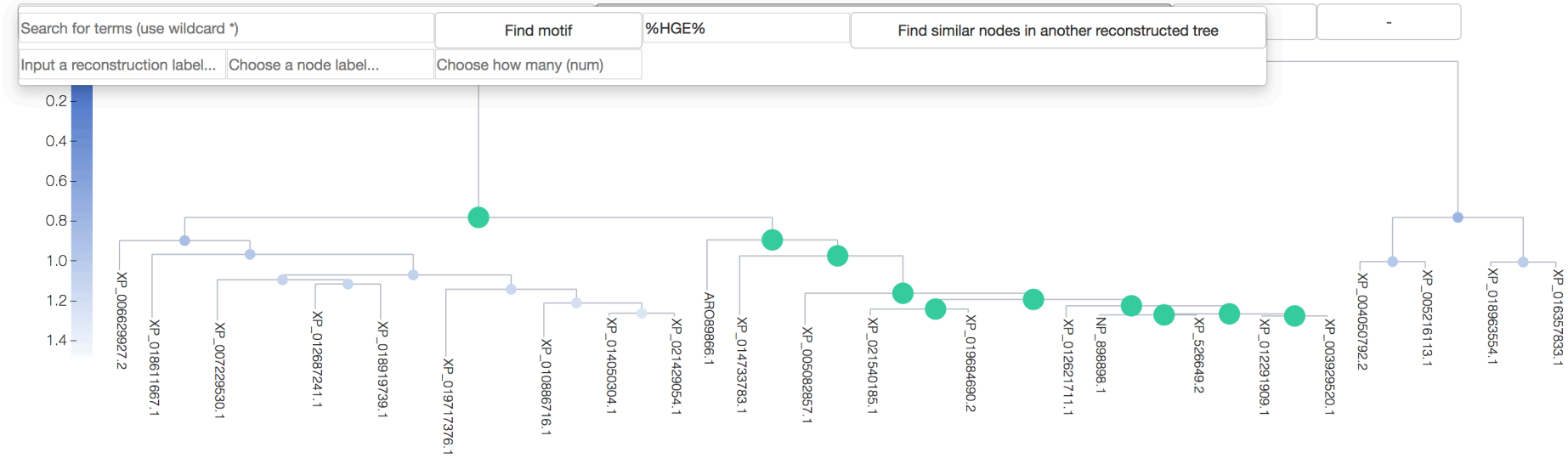
Ancestor with consensus path



GRASP annotations and searching



- Taxonomic annotation from UniProt / NCBI
- Searching ancestors for -
  - Annotations
  - Sequence motifs



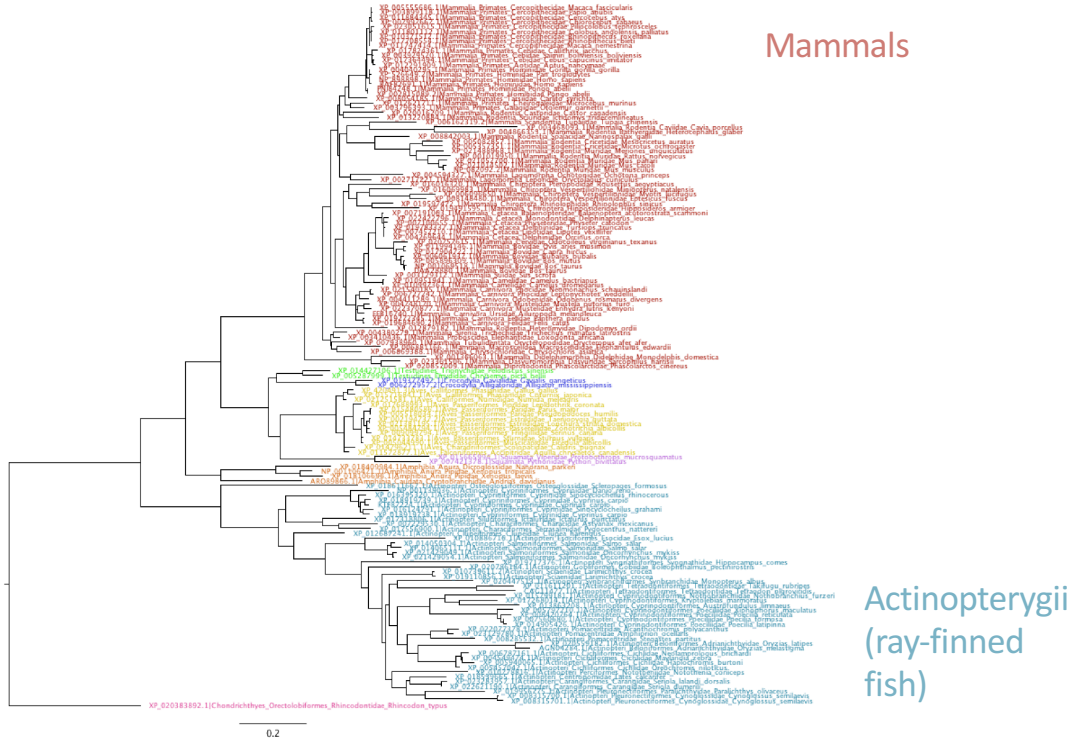
SeqScrub for curation

- Annotates
- Cleans
- Checks for obsolete sequences
- Checks for given characters

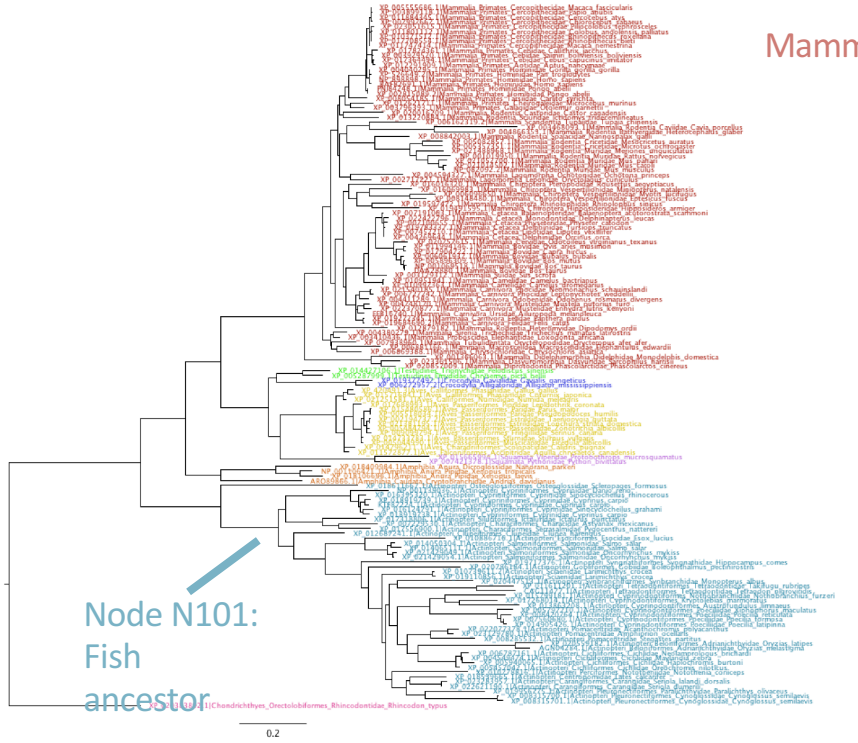
- **Communicates with NCBI / UniProt**
- **Completely in-browser application**

Foley, Sützl, D'Cunha, Gillam, Bodén, *BioTechniques* (2019) doi:10.2144/btn-2018-0188

GRASP enables inspection of indel histories



GRASP enables inspection of indel histories

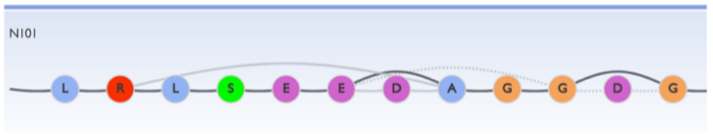


Mammals

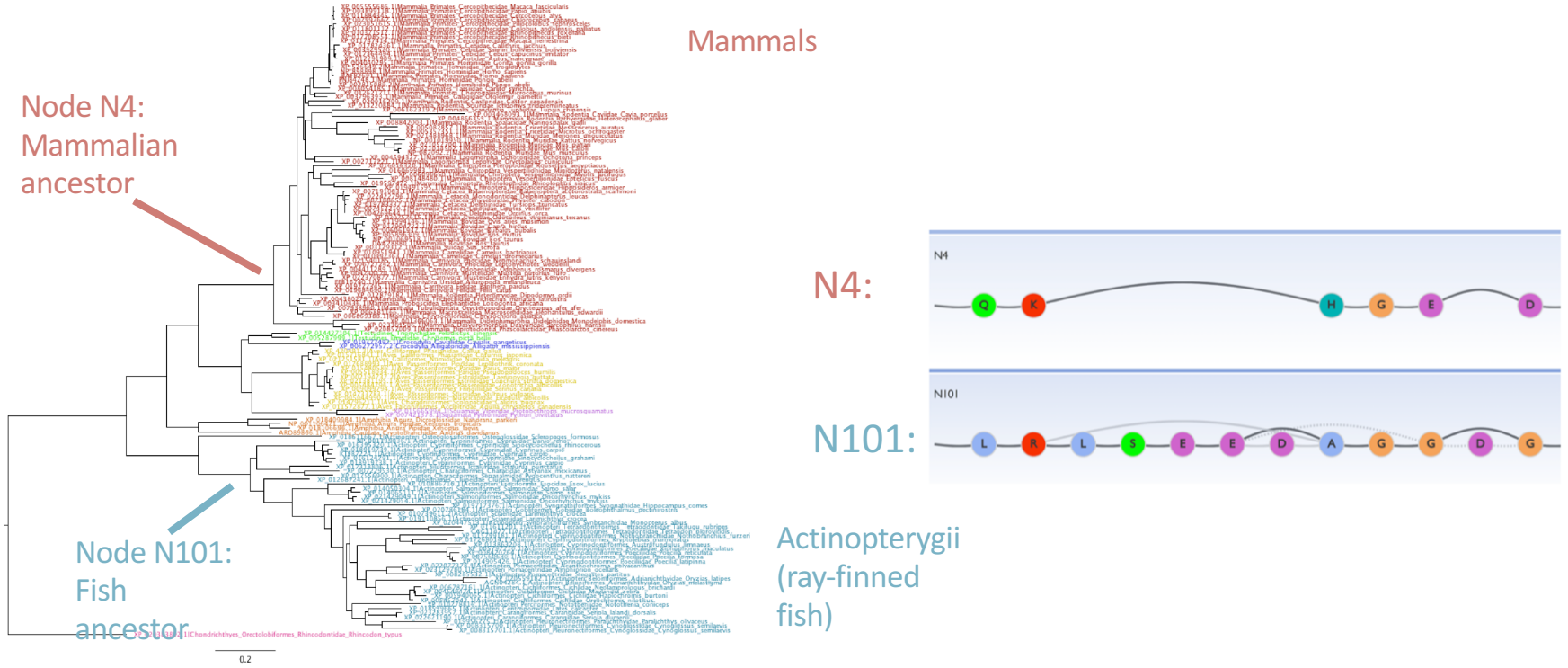
Node N101: Fish ancestor

Actinopterygii (ray-finned fish)

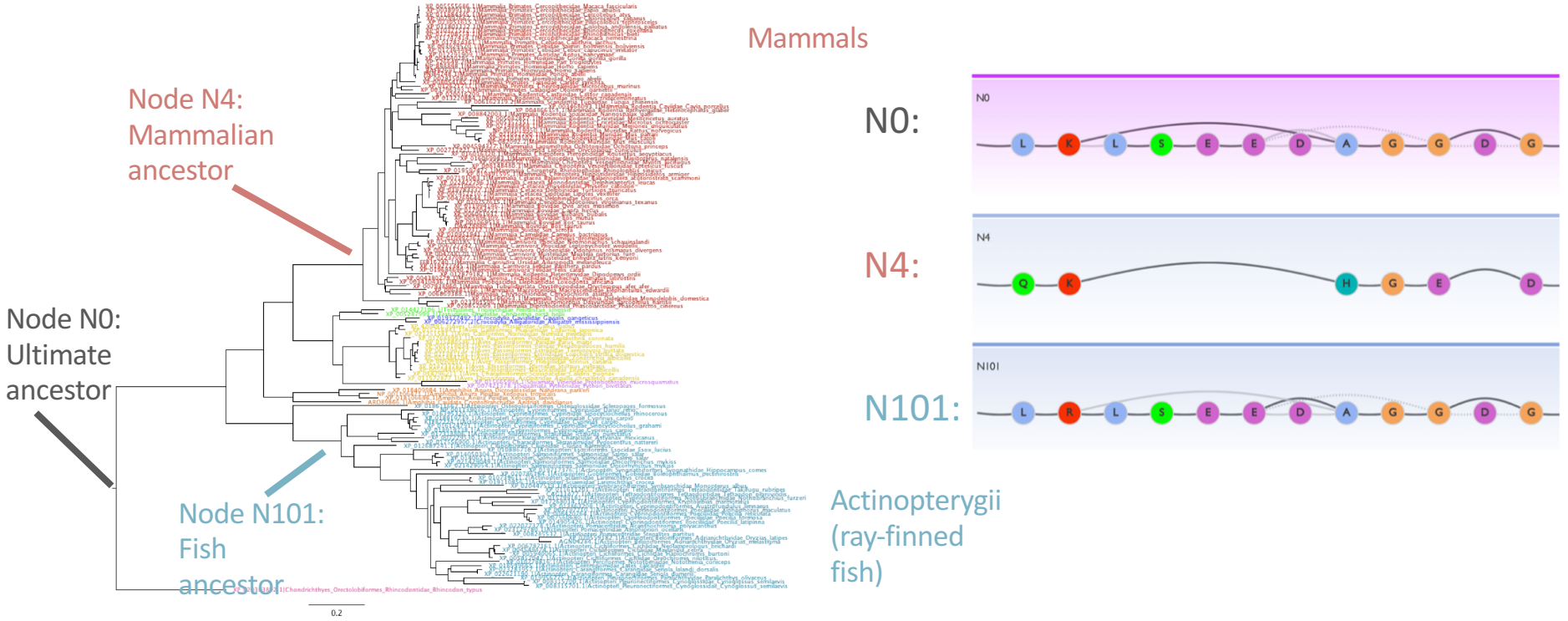
N101:



GRASP enables inspection of indel histories



GRASP enables inspection of indel histories



Conclusion

- Ancestral sequence reconstruction is a valuable resource to understand, explore, and utilise evolution
- Large data sets allow us to extend the reach of ASR
- GRASP enables novel experiments on previously unobtainable data set sizes

Acknowledgements

PhD supervisors

Elizabeth Gillam

Mikael Bodén

Ross Barnard

GRASP

Graphical representation of ancestral sequence predictions

http://grasp.scmb.uq.edu.au

The Gillam and Bodén groups

Groups of Volker Sieber and Dietmar Haltrich

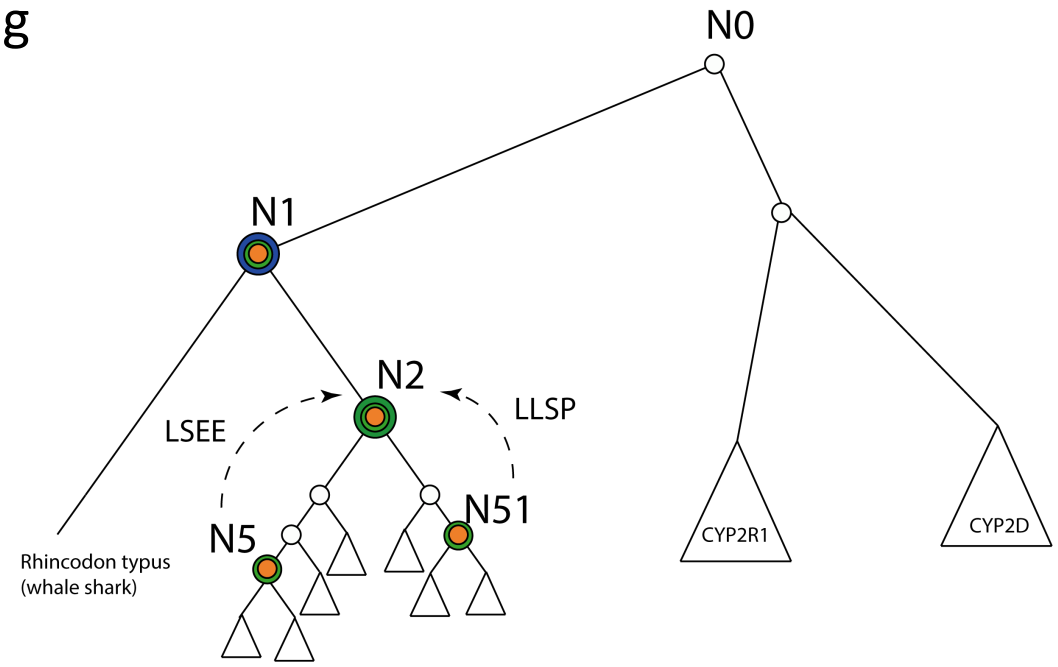
Connie Ross, Ariane Mora, Marnie Lamprecht,
Raine Thomson, Yosephine Gumulya,
Kurt Harris, Stephlina D'Cunha

Additional slides

Constructing novel indel variants

From this tree we **reconstructed 10 CYP2U1 ancestors**, including **six ancestors** that either reverted or pre-empted insertions and deletions.

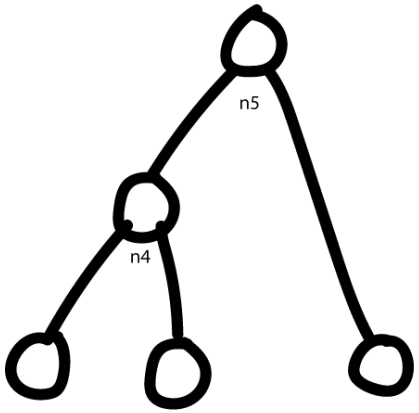
All ancestors were able to express and show a characteristic P450 spectrum.



CYP2U1 / CYP2R1 / CYP2D tree

Experimental work performed by Connie Ross

Marginal & joint differences



P(n4 = A, n5 = A) = 0.4
P(n4 = A, n5 = C) = 0.3
P(n4 = C, n5 = A) = 0.05
P(n4 = C, n5 = C) = 0.25

Joint reconstruction of node n4 and node n5
Find the highest probability
P(n4 = A, n5 = A) = 0.4

Character at n5 is assigned A

Marginal reconstruction of node n5
Sum up all the ways we could get n5=A
P(n4 = A, n5 = A) + P(n4= C, n5 =A)
= 0.4 + 0.05
= 0.45

Sum up all the ways we could get n5=C
P(n4 = A, n5 = C) + P(n4 = C, n5 = C)
= 0.3 + 0.25
= 0.55

Character at n5 is assigned C

Marginal & joint differences

Posterior probability distributions from the CYP2U1 CYP2R1 Realigned marginal reconstruction at positions where the marginal and joint reconstructions differ