

# Ancestral sequence reconstruction: in your toolkit yet?

Mikael Bodén [m.boden@uq.edu.au](mailto:m.boden@uq.edu.au)

School of Chemistry and Molecular Biosciences

The University of Queensland

tr|A0A372LUR1|A0A372LUR1\_9BACI  
tr|A0A00U1QP27|A0A00U1QP27\_9BACL  
tr|A0A2T4MQL9|A0A2T4MQL9\_9STAP  
tr|C8P4Q1|C8P4Q1\_9LACO  
tr|E8KA36|E8KA36\_9STRE  
tr|A0A172Q8Z2|A0A172Q8Z2\_9STRE  
tr|A0A2W0HJZ5|A0A2W0HJZ5\_9BACI  
tr|A0A1Y4D2E8|A0A1Y4D2E8\_9DELT  
tr|A0A142YD67|A0A142YD67\_9PLAN  
sp|Q7NH80|ILVC\_GLOVI  
tr|Q1PJS0|Q1PJS0\_PROMR  
tr|A0A1U7HHX6|A0A1U7HHX6\_9CHRO  
tr|A0A168VRL7|A0A168VRL7\_9CYAN  
sp|B7K0S6|ILVC\_RIPO1  
tr|A0A399XTD6|A0A399XTD6\_9BACT  
tr|A0A2D9WLU1|A0A2D9WLU1\_9CHLR  
tr|A0A401ZHA7|A0A401ZHA7\_9CHLR  
tr|A0A2E5F4W7|A0A2E5F4W7\_9CHLR  
tr|A0A2P2DPY7|A0A2P2DPY7\_9LEPT  
tr|A0A1Q7FRP1|A0A1Q7FRP1\_9BACT  
tr|A0A2M8P4N5|A0A2M8P4N5\_9CHLR  
tr|A0A419FXS0|A0A419FXS0\_9BACT  
sp|A1AS39|ILVC\_PELPD  
tr|D1KBL9|D1KBL9\_9GAMM  
tr|A0A4R1H8N8|A0A4R1H8N8\_9GAMM  
tr|A0A2H5ZH23|A0A2H5ZH23\_9BACT  
tr|S2ZLE3|S2ZLE3\_9FIRM  
tr|F0HRP2|F0HRP2\_9ACTN  
tr|A0A2G2GHS1|A0A2G2GHS1\_9RHOB  
tr|A0A2N6FK57|A0A2N6FK57\_9DELT  
sp|Q1R092|ILVC\_CHRSD  
tr|A0A396RUJ6|A0A396RUJ6\_9PSED  
tr|A0A346S4E9|A0A346S4E9\_9ALTE  
tr|A0A2R7SF53|A0A2R7SF53\_9PSED  
tr|U3HF10|U3HF10\_PSEAC  
tr|A0A348YAC6|A0A348YAC6\_9GAMM  
tr|A0A0X3TG90|A0A0X3TG90\_9GAMM  
tr|A0A4P5W7I4|A0A4P5W7I4\_9GAMM  
tr|A0A2E6K461|A0A2E6K461\_PSESP  
tr|A0A1H5TSF5|A0A1H5TSF5\_9PROT  
tr|A0A2W5DX57|A0A2W5DX57\_9BURK  
tr|A0A109BNM4|A0A109BNM4\_9BURK  
tr|A0A4V2TBP4|A0A4V2TBP4\_9BURK  
tr|A0A1H2PSN4|A0A1H2PSN4\_9BURK

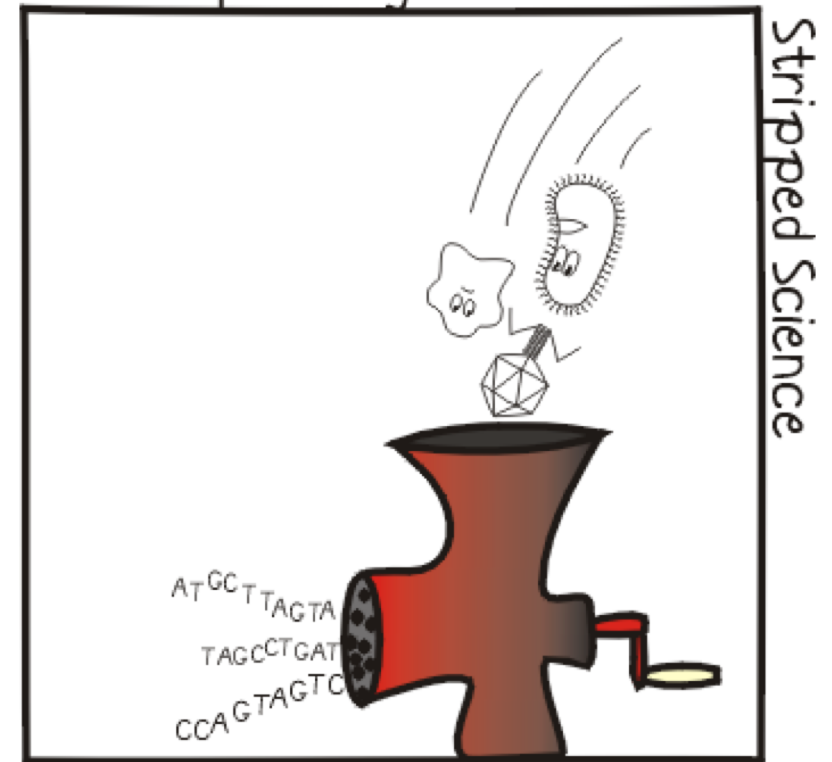
Ketol-acid reducto-isomerase (KARI)  
Cytochrome P450 (CYP2U/2D/2R)  
Dihydroxy-acid dehydratase (DHAD)  
Glucose-methanol-choline (GMC)  
oxidoreductases, incl. GDH and GOx

- What in my protein's sequence give it the structure or function it has?
- How do I change the sequence to alter structural or functional qualities of my protein?

# Wealth of biological sequences: access to diversity

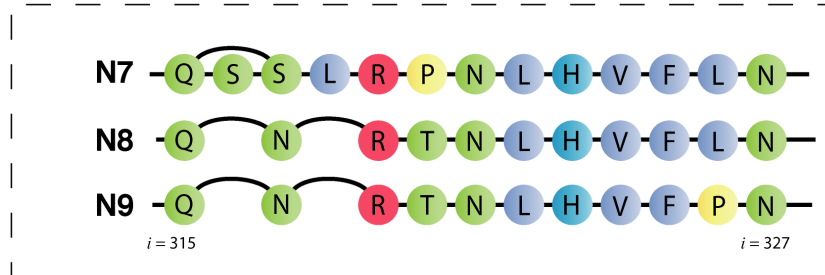
- Homology: a shared evolutionary origin explains similar sequence, structure and function
- Models of evolution: substitution, insertions and deletions
- **How** did we end up with these sequences
  - infer a *history* of evolutionary events
  - correlating changes in sequence with those in function (and structure)

Mass sequencing

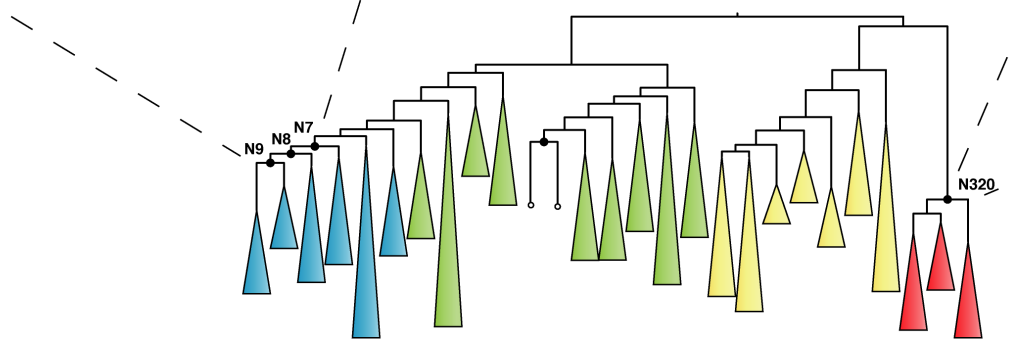
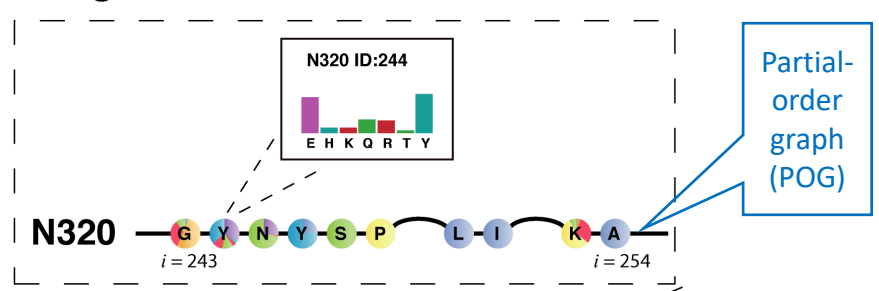


by Viktor S. Poór

## Joint reconstruction







## Marginal reconstruction



**GRASP**  
 Graphical representation of ancestral sequence predictions

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

 GDH I	 GDH III
 GDH II	 GOx I

399 sequences



# Opportunities and hypotheses

Tracking indels offers a new source of sequence variation for protein design: “hybrid ancestors”

1. within-family insertions and deletions can be used as building blocks to support the engineering of biologically active ancestors and novel catalysts

Greater volume of data implies access to more natural diversity (and further evolutionary reach)

2. Including more sequences while accounting for indels improves robustness of ancestor inference  
also enables a new approach to survey activity of members “by-ancestor proxy”

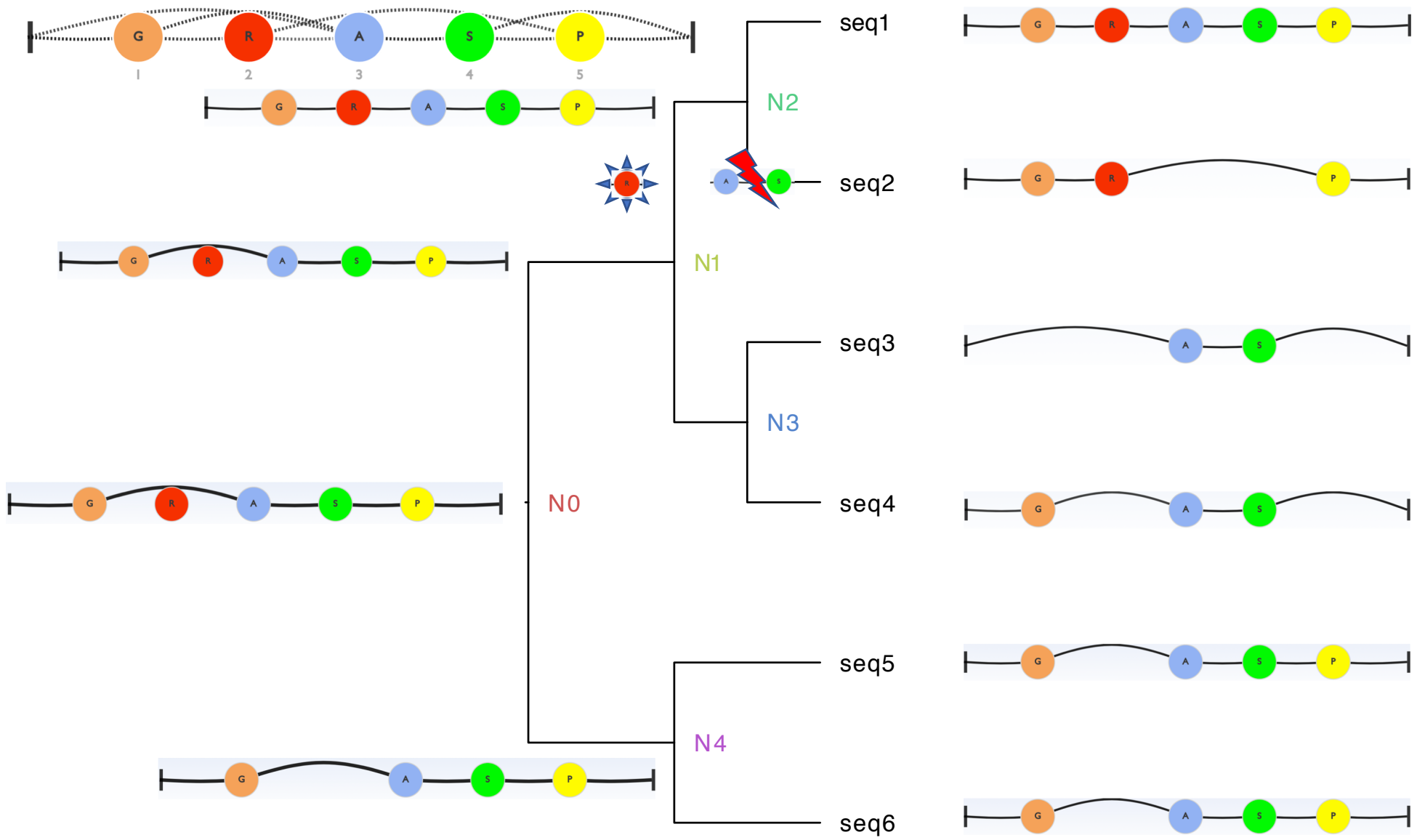
# Capture traces of change

- GRASP accounts for (and tracks)
  - substitutions and
  - insertions/deletions = indels

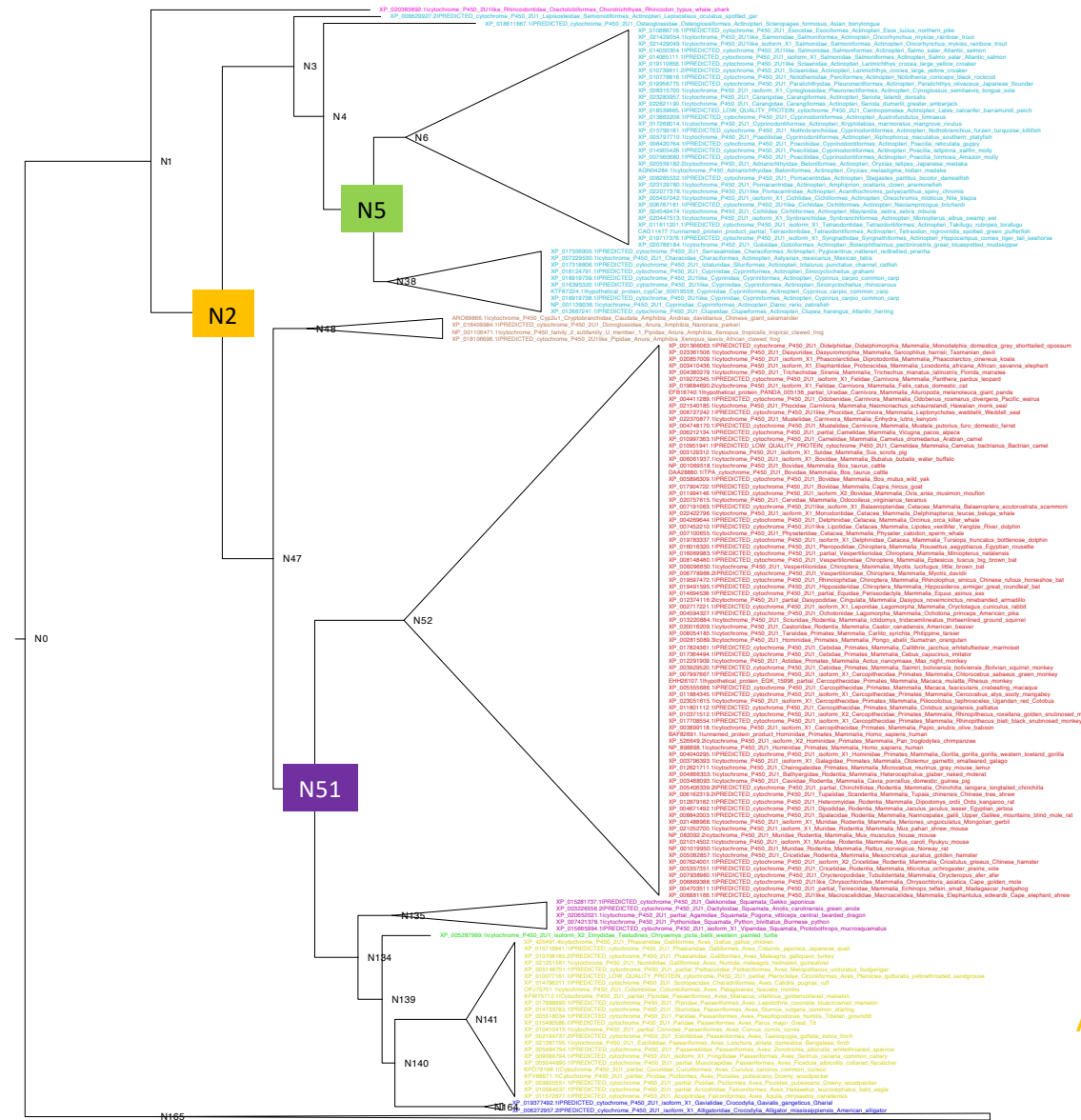
Inferred substitutions and ambivalent amino acid states have been used successfully to engineer and explore ancestor variants

Hypothesis 1:

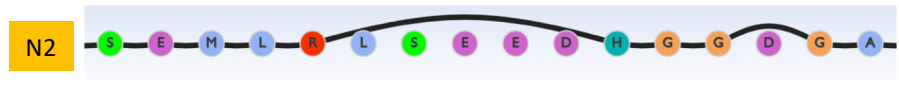
*within-family insertions and deletions can be used as building blocks to support the engineering of biologically active ancestors and novel catalysts*



# Cytochrome P450 2U1 subfamily



Fish



Mammalia



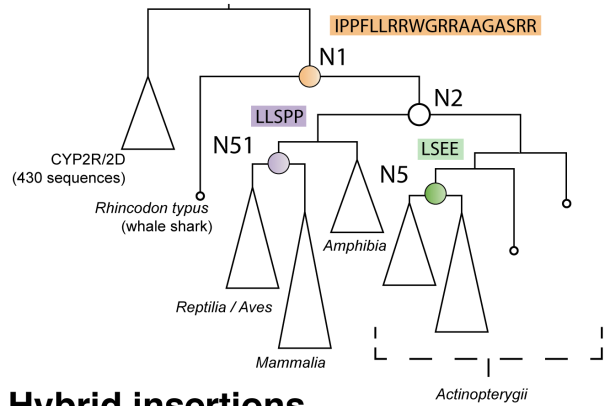
Aves

595 sequences

0.3



### a) CYP2U/2R/2D phylogenetic tree



### b) Hybrid insertions

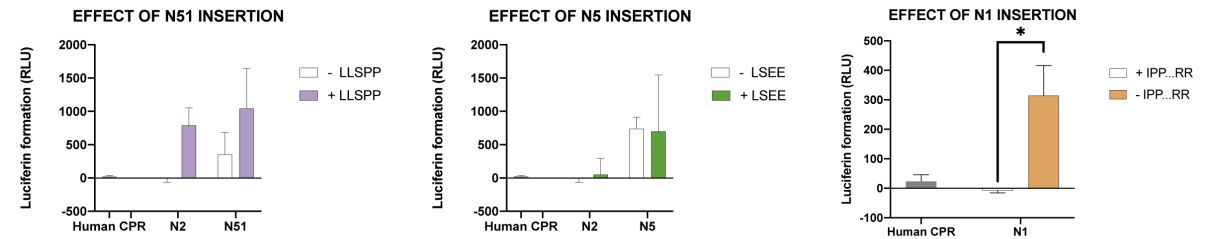
```

N51      ...VLLPPFLRRRW L L S P P L R R A A G A G R R S A L ...
N51_27dLLSPP ...VLLPPFLRRRW - - - - L R R A A G A G R R S A L ...
N2       ...L L I P P F L L R R W - - - - G R R A A G A S R R S A L ...
N2_27iLLSPP ...L L I P P F L L R R W L L S P P G R R A A G A S R R S A L ...
          16                               44

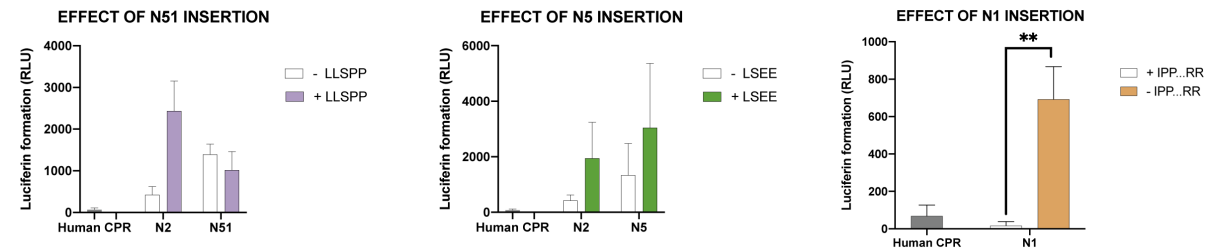
N5       ...G L A I V K S E L L R L S E E S G G S G V D L T P L I S N ...
N5_153dLSEE ...G L A I V K S E L L R - - - S G G S G V D L T P L I S N ...
N2       ...E L K F V K S E M L R - - - H G G G A F N P S P I I N N ...
N2_152iLSEE ...E L K F V K S E M L R L S E E H G G G A F N P S P I I N N ...
          142/143                          170/171

N1       ...L L S L L I P P F L L R R W G R R A A G A S R R S A L L S ...
N1_19dIP..RR ...L L S L L - - - - - - - - - - - - - - - S A L L S ...
          14                               42
    
```

### c) Activity with luciferin CEE



### d) Activity with luciferin ME-EGE

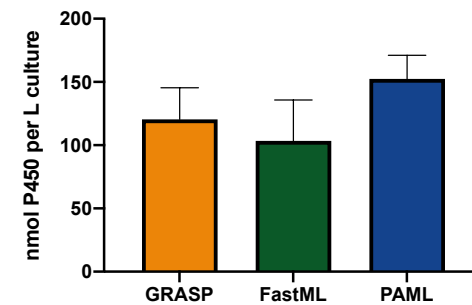


# On smaller datasets, GRASP is consistent with FastML and PAML

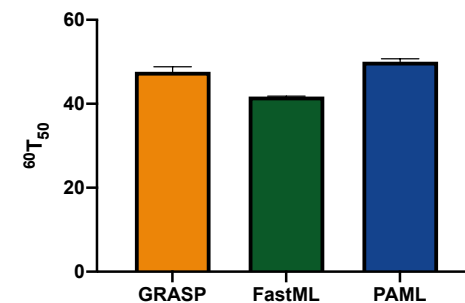
On a reconstruction with 359 sequences:

- Above 95% sequence identity
- Same/similar expression
- Same/similar thermal stability
- GRASP faster

Expression level of cytochrome P450 CYP2U1 ancestor



Thermal stability of cytochrome P450 CYP2U1 ancestor



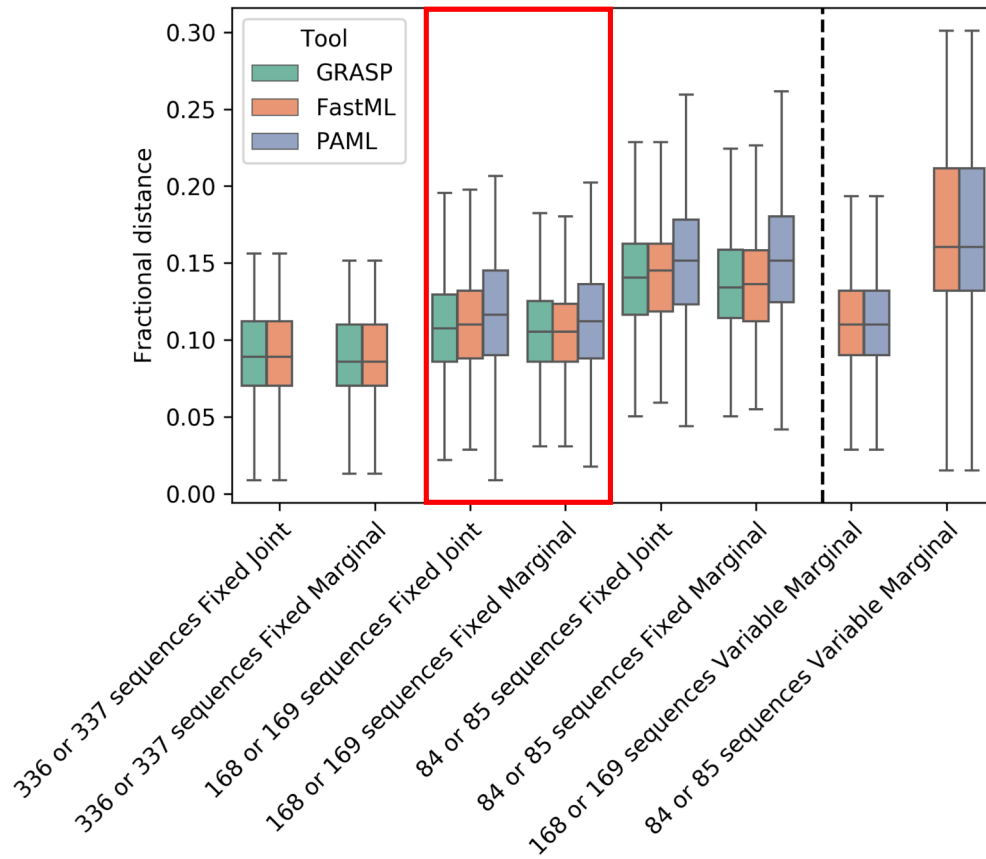
**Methods' consensus standard:**

**Generate ancestors that are similar to those of other methods**

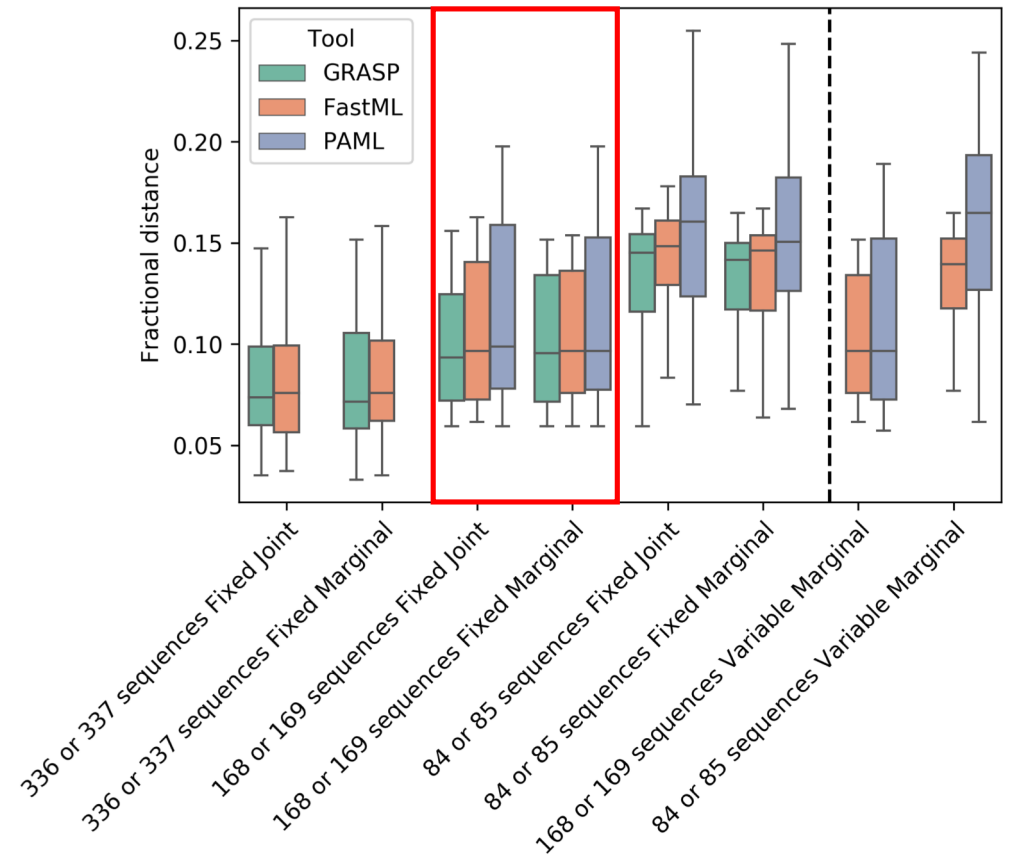
**Data consensus standard:**

**Generate ancestors close to that of the superset**

Distance between members across methods



Distance between members and superset ancestor



# Capture quantity of data, tap into diversity (part 1)

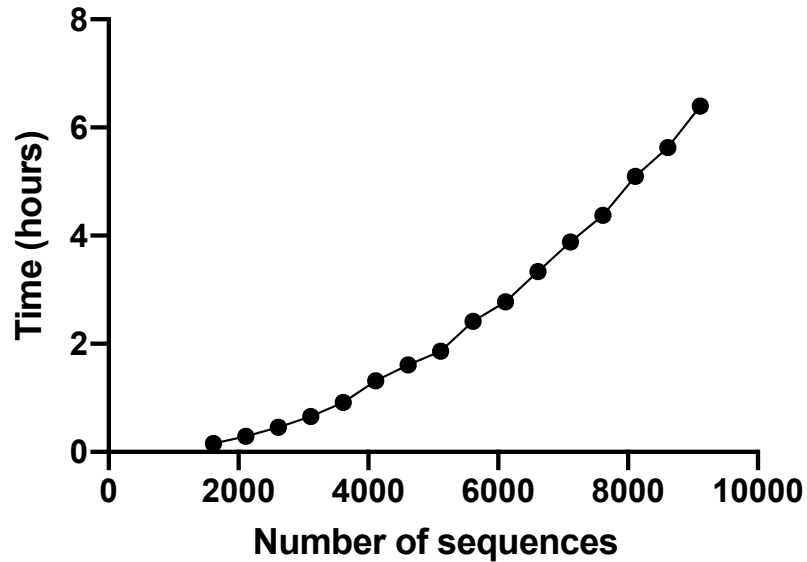
- Large volumes of data
  - mean more insertion and deletions, and
  - require attention to the computational expense of phylogenetic inference
- Inference engine in GRASP is efficient
  - uses any standard evolutionary rate matrix
  - implements an algorithm that decomposes the problem into the smallest number of operations, in the order with the least computational complexity

Capacity of popular tools is typically limited to less than a thousand sequences. What are the limits?

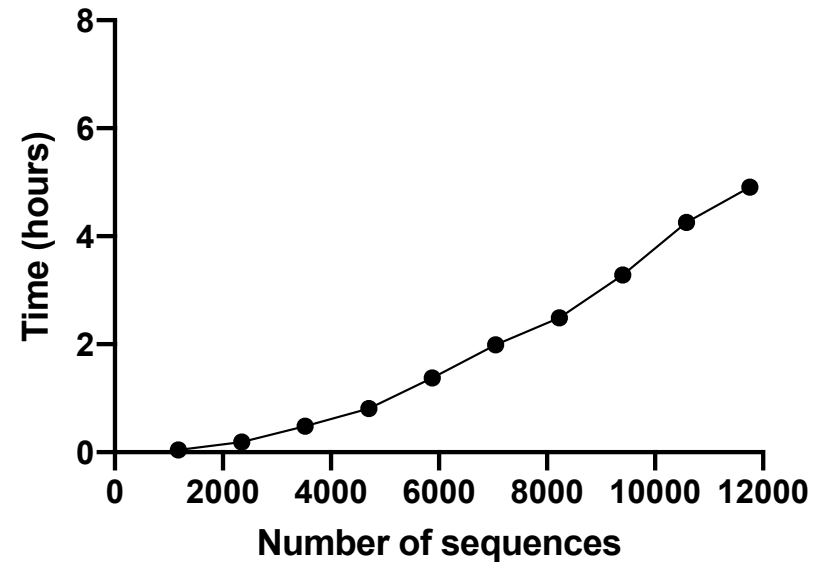


Run times (64 GB RAM, 5 threads on 2x 2.6 GHz 14C Xeon VM)

**DHAD run time**



**KARI run time**



# Capture quantity of data, tap into diversity (part 2)

- Greater volume of data implies access to more natural diversity
- Ancestor POGs track indels over time and across clades
  - to predict substitutions between homologous sections *only*
  - to present ambivalent indels to user  
(visual exploration offered in web service)

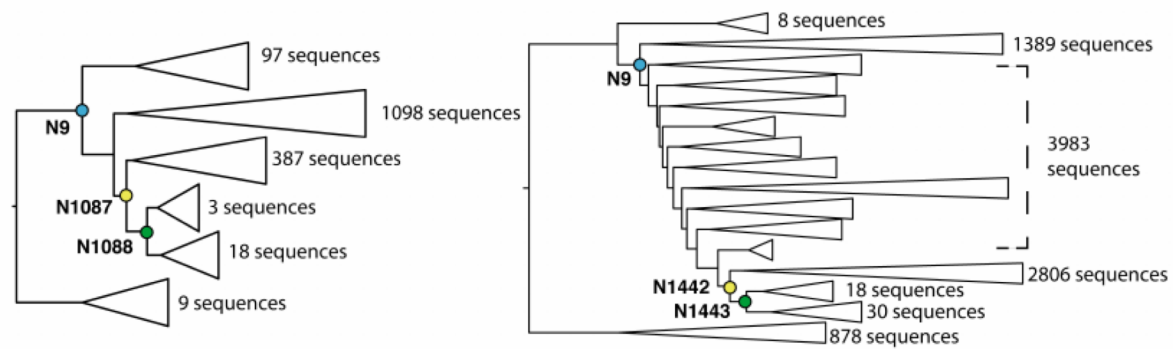
Columns in a classical alignment increases with sequence numbers, indicating complex indel histories

## Hypothesis 2:

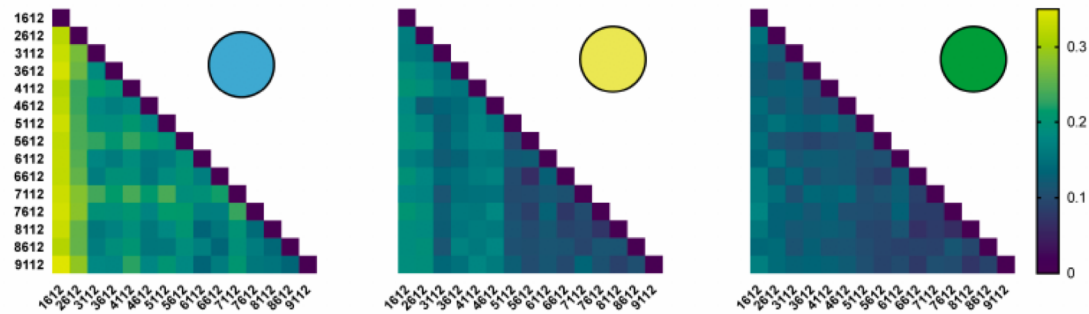
*More sequences, accounting for indels improves robustness of ancestor inference; also enables a new approach to survey activity of members “by-ancestor proxy”*

# Increasing data volume constrains ancestral predictions

a) DHAD phylogenetic trees of 1612 vs 9112 sequences

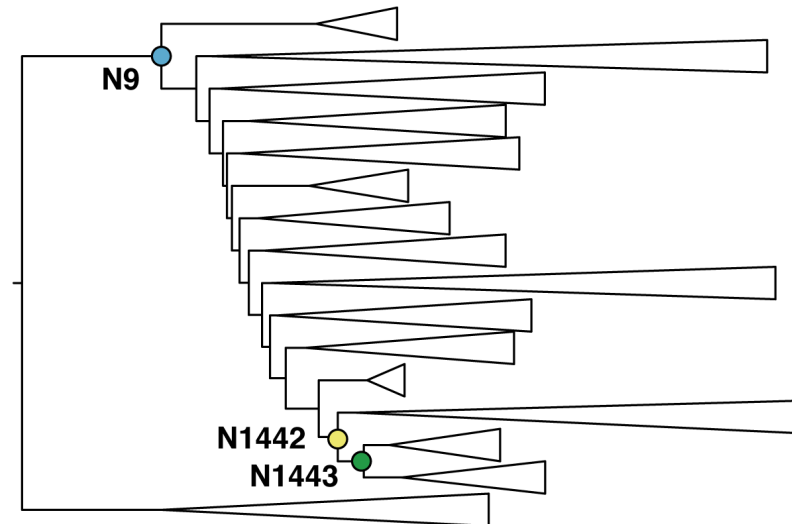


b) DHAD distance maps

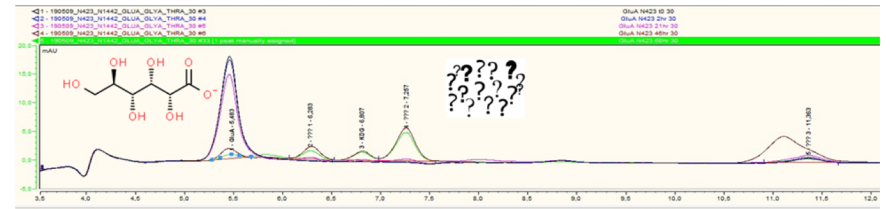


# Surveying function across clades?

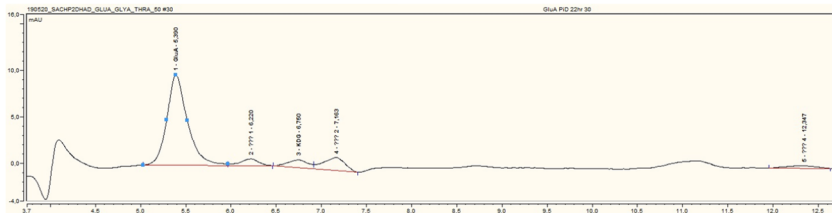
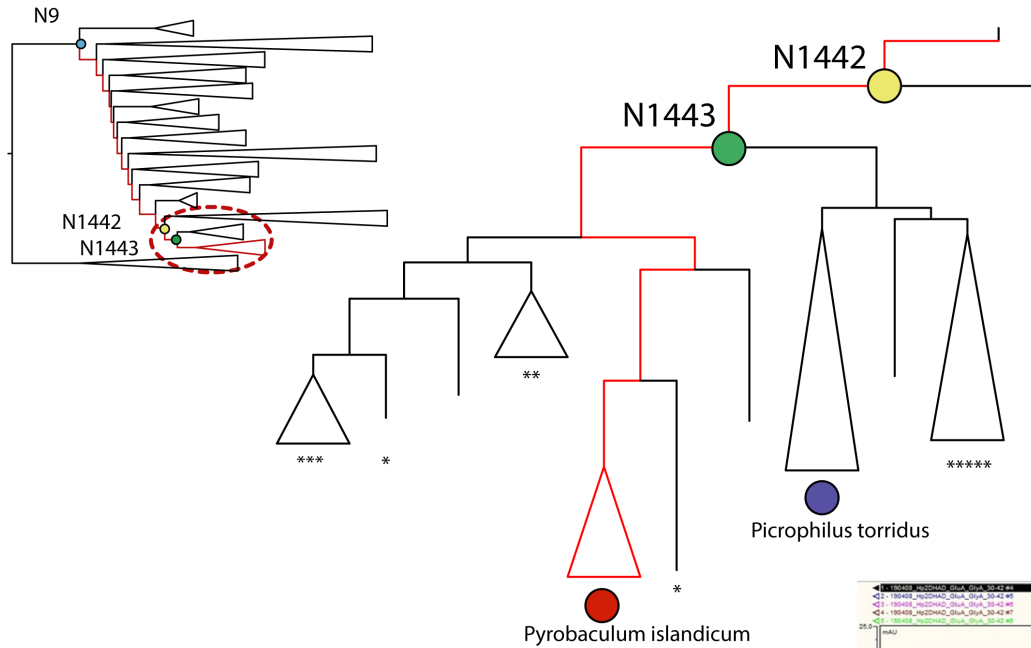
- Two reconstructions at the extremes (585 v 9112 sequences)
  - Characterised three ancestors in each, “middle” ancestor had 72.6% sequence identity
- Of 9112 sequences only
  - ~5% have high quality annotations (SwissProt)
  - 22 organisms are experimentally characterised (BRENDA database)



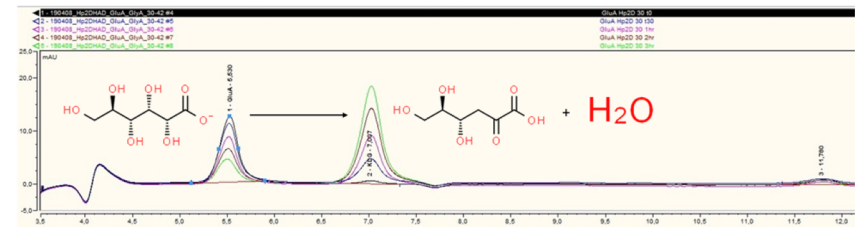




N1442 DHAD incubation



*Pyrobaculum islandicum* DSM 4184 DHAD incubation



*Hydrogenophaga pseudoflava* DSM 1084 DHAD incubation  
(with D-Gluconate in HEPES buffer pH 7.0)

# Conclusion

- Ancestral sequence reconstruction is a valuable resource to understand, explore and use evolution; GRASP is new tool which compares well with existing methods
- We report on experimentally resurrected ancestors with variants for three enzyme families but GRASP has been used on several more
- Ability to infer indels onto a phylogenetic tree...
  - predicts events that define whole clades
  - POGs allow ambiguous calls to be tracked and disentangled across time
  - hybrid ancestors are a novel class of variant, identified by partitioning of sequence by indel events, with re-purposing at alternate branch-points
- Ability to process in excess of 10,000 sequences...
  - extends evolutionary reach,
  - taps into natural diversity, and
  - supports annotation of function in extants “by-ancestor proxy”

# Acknowledgements

FUNDED BY



**Australian Government**  
**Australian Research Council**

RESEARCH in the national interest - enabling the future



## UQ Gillam lab

Elizabeth Gillam

Connie Ross

Raine Thomson

## Boden lab

Gabriel Foley

Ariane Mora

Marnie Lamprecht

Julian Zaugg

Luke Guddat

Alexandra Essebier

Gary Schenk

Brad Balderson

Bostjan Kobe

Rhys Newell

Ross Barnard

## TUM Sieber lab

Volker Sieber (TUM)

Scott Bottoms (TUM)

Jörg Carsten (TUM)

Burkhard Rost (TUM)

## BOKU Haltrich lab

Dietmar Haltrich (BOKU)

Leander Sützl (BOKU)



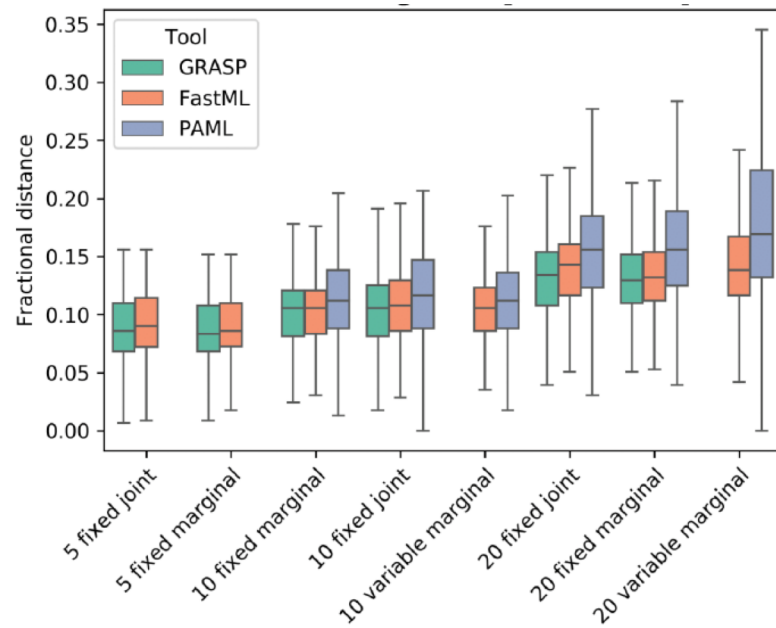
<http://grasp.scmb.uq.edu.au>



SCAN ME

Foley *et al.* Identifying and engineering ancient variants of enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP) **to be posted on bioRxiv by the end of 2019**

# GRASP infers ancestors consistent with existing methods



Percent Identity Matrix - created by Clustal2.1

1: PAML_Joint_554	100.00	95.46	95.05
2: CYP2U1_CYP2R1_Realigned_N1	95.46	100.00	96.08
3: FastML_Joint_N2_Gaps_Removed	95.05	96.08	100.00



# Cytochrome P450 2U1 subfamily

Cytochrome P450 enzymes are members of a superfamily of monooxygenases that play a **critical role in metabolism**

**CYP2U1** – cytochrome P450 subfamily found across, amphibians, reptiles, mammals, birds, and fish.



**CYP2U1 is interesting because –**

- No exact established function and substrate specificity
- Previous cytochrome P450 ancestors showed increased stability and promiscuity



# Dihydroxy-acid dehydratase (DHADs)

- Enzyme family found across bacteria, archaea, fungi, algae, and some plants
- Attractive target for protein engineering because it is **used in dehydration reactions that are important for biofuel production**