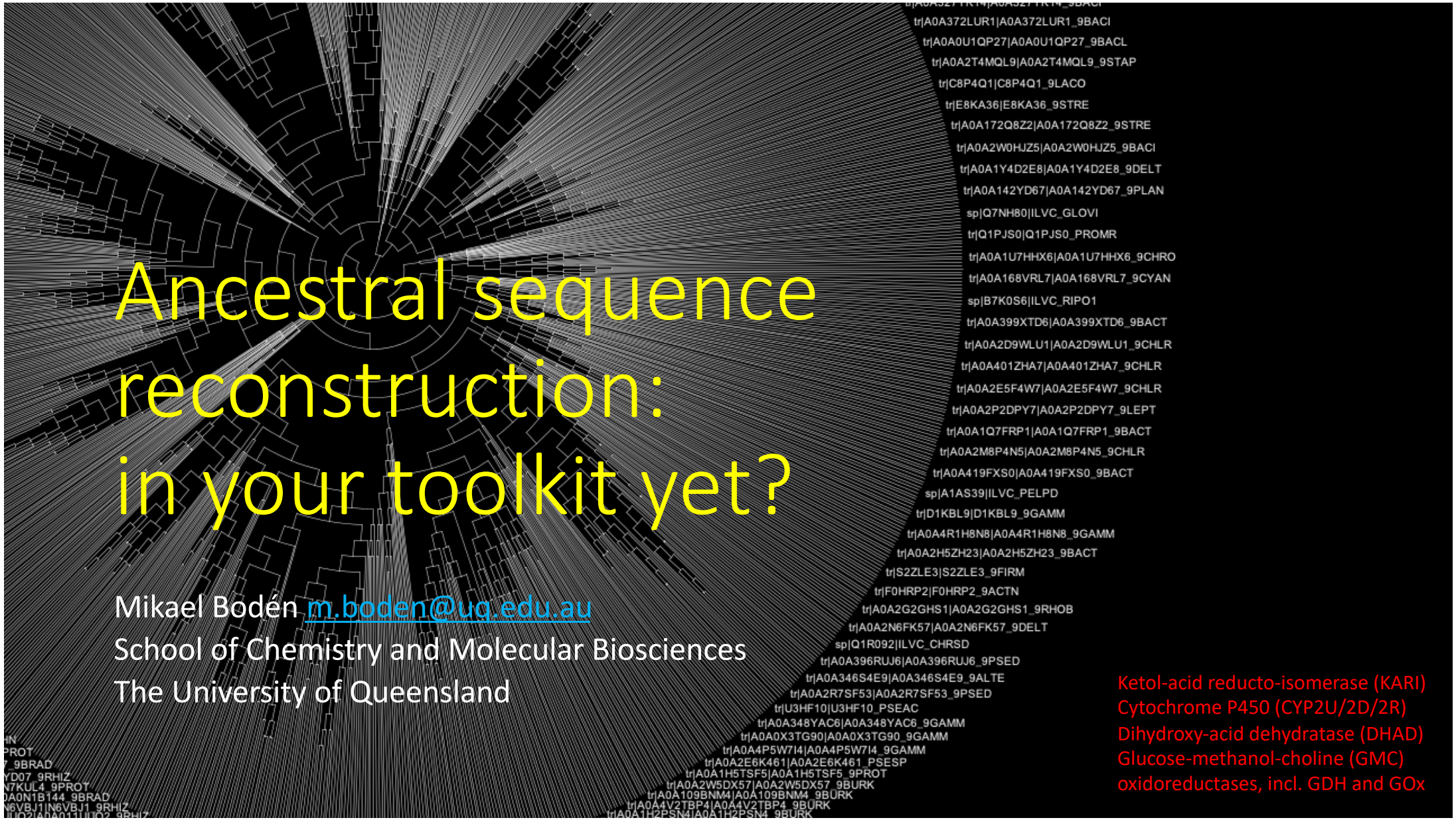


Ancestral sequence reconstruction: in your toolkit yet?

Mikael Bodén m.boden@uq.edu.au
School of Chemistry and Molecular Biosciences
The University of Queensland

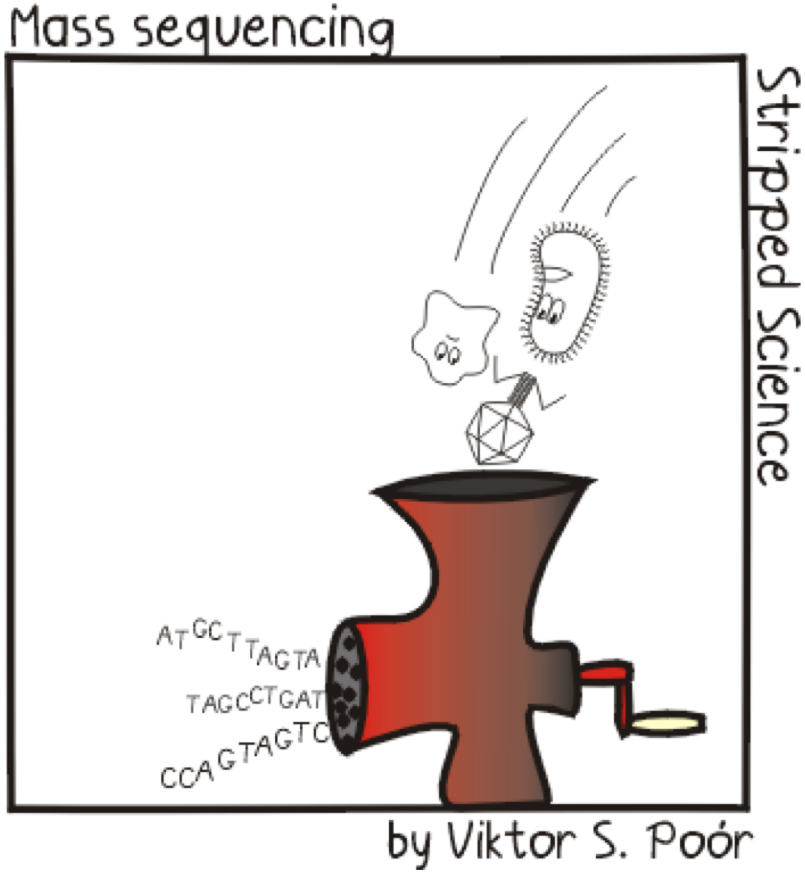
Ketol-acid reducto-isomerase (KARI)
Cytochrome P450 (CYP2U/2D/2R)
Dihydroxy-acid dehydratase (DHAD)
Glucose-methanol-choline (GMC) oxidoreductases, incl. GDH and GOx

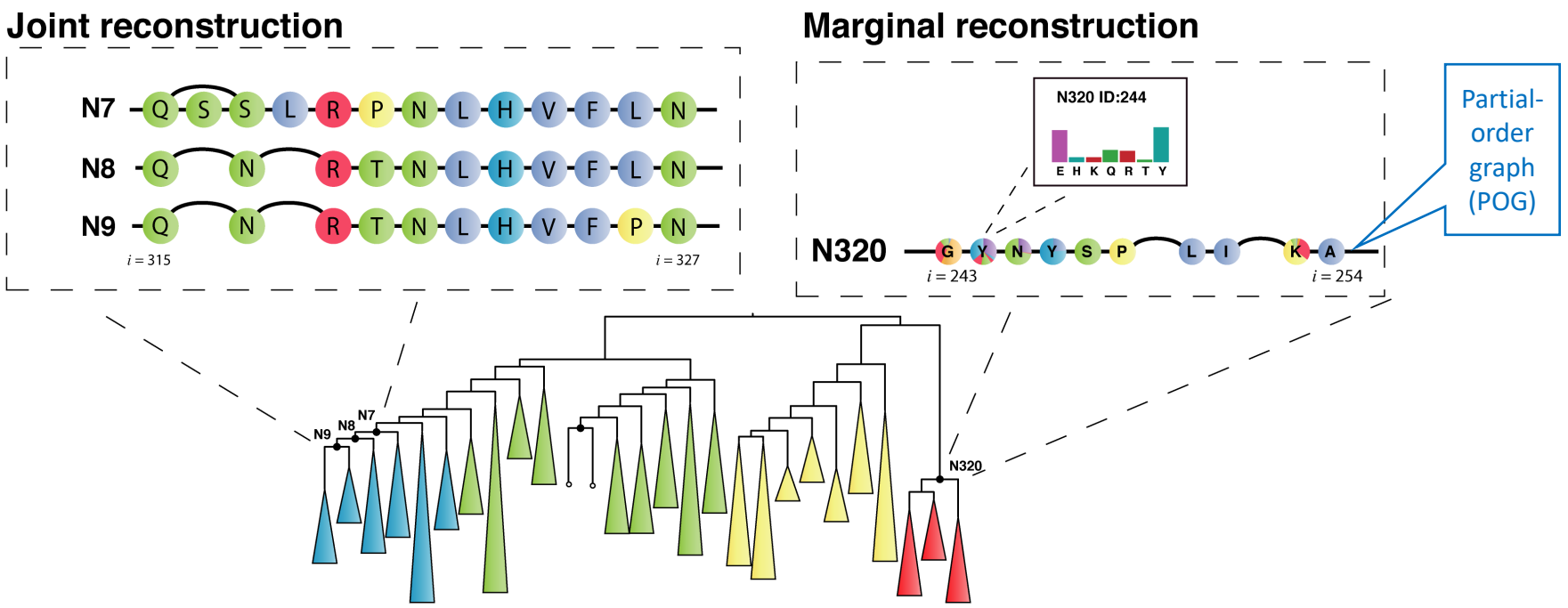
- What in my protein’s sequence give it the structure or function it has?

- 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)





GRASP

Graphical representation of ancestral sequence predictions

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

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”

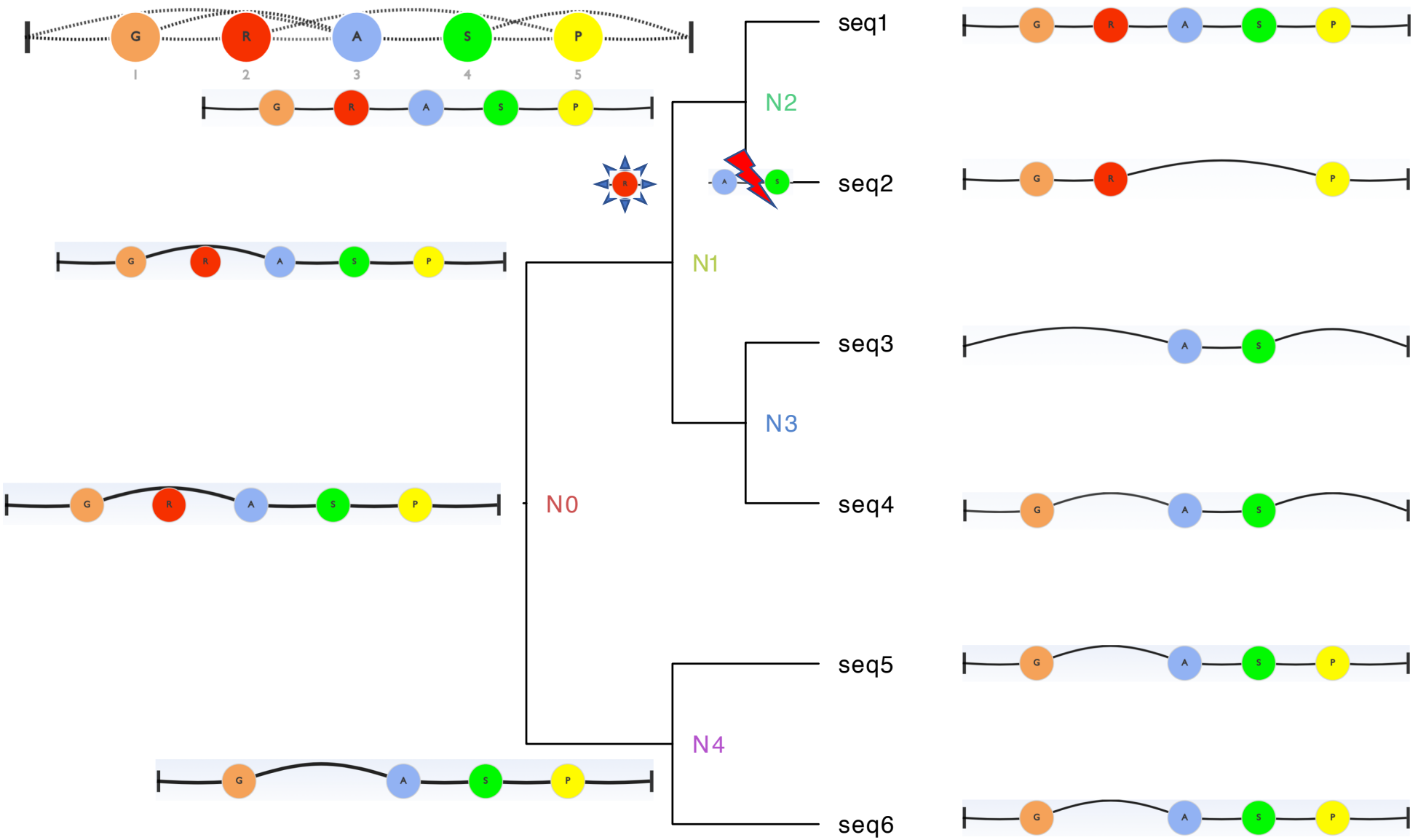
Capture traces of change

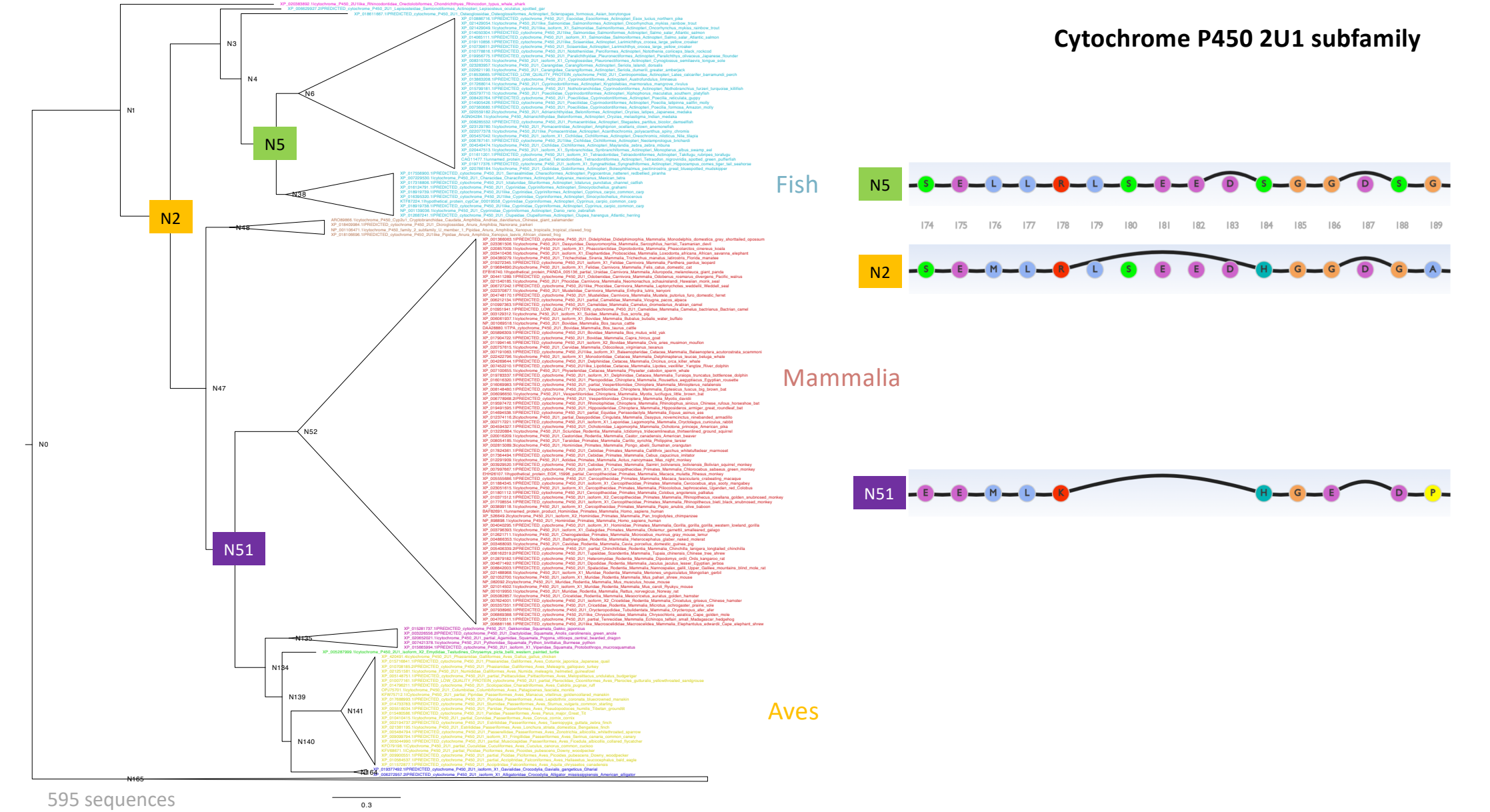
- GRASP accounts for (and tracks)
  - substitutions and
  - insertions/deletions = indels

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

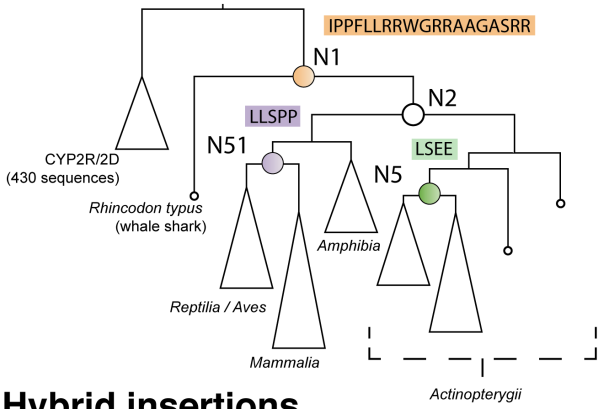
Hypothesis 1:

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

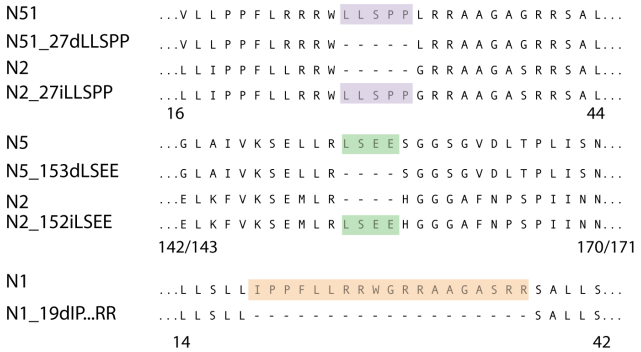

G
R
A
S
P
1
2
3
4
5
seq1
seq2
seq3
seq4
seq5
seq6
N0
N1
N2
N3
N4


Cytochrome P450 2U1 subfamily
Fish
Mammalia
Aves
N0
N1
N2
N3
N4
N5
N6
N47
N48
N51
N52
N134
N135
N139
N140
N141
N165
595 sequences
0.3

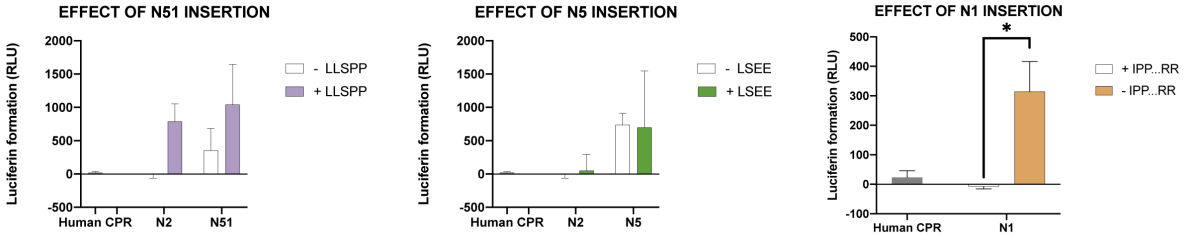
a) CYP2U/2R/2D phylogenetic tree



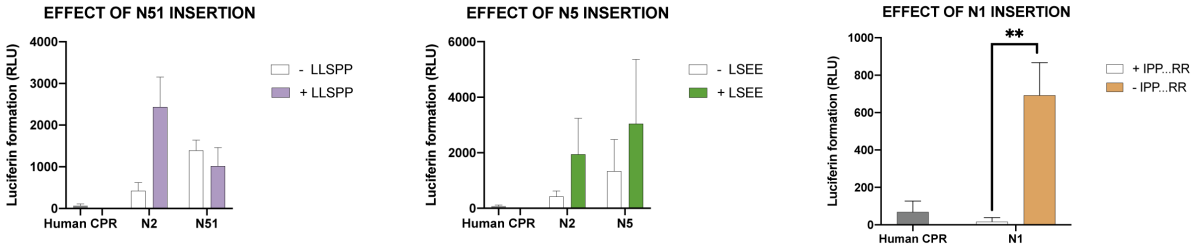
b) Hybrid insertions



c) Activity with luciferin CEE



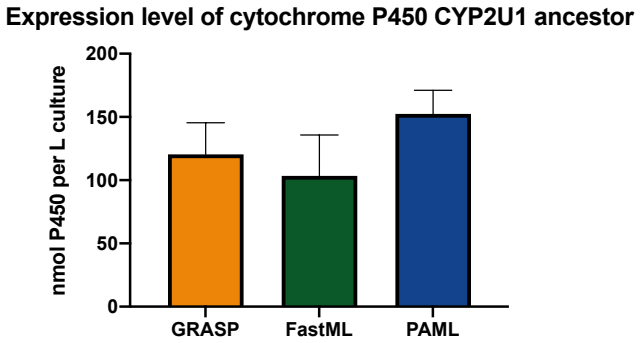
d) Activity with luciferin ME-EGE

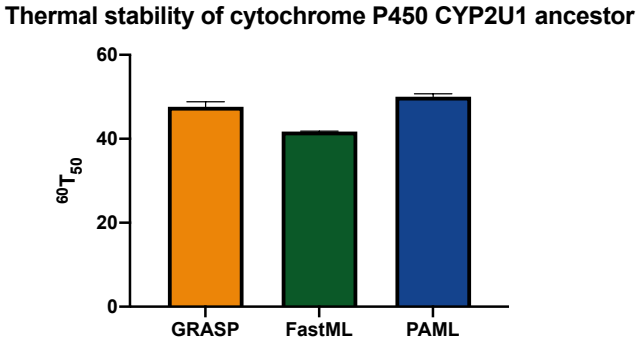


On smaller datasets, GRASP is consistent with FastML and PAML

On a reconstruction with 359 sequences:

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

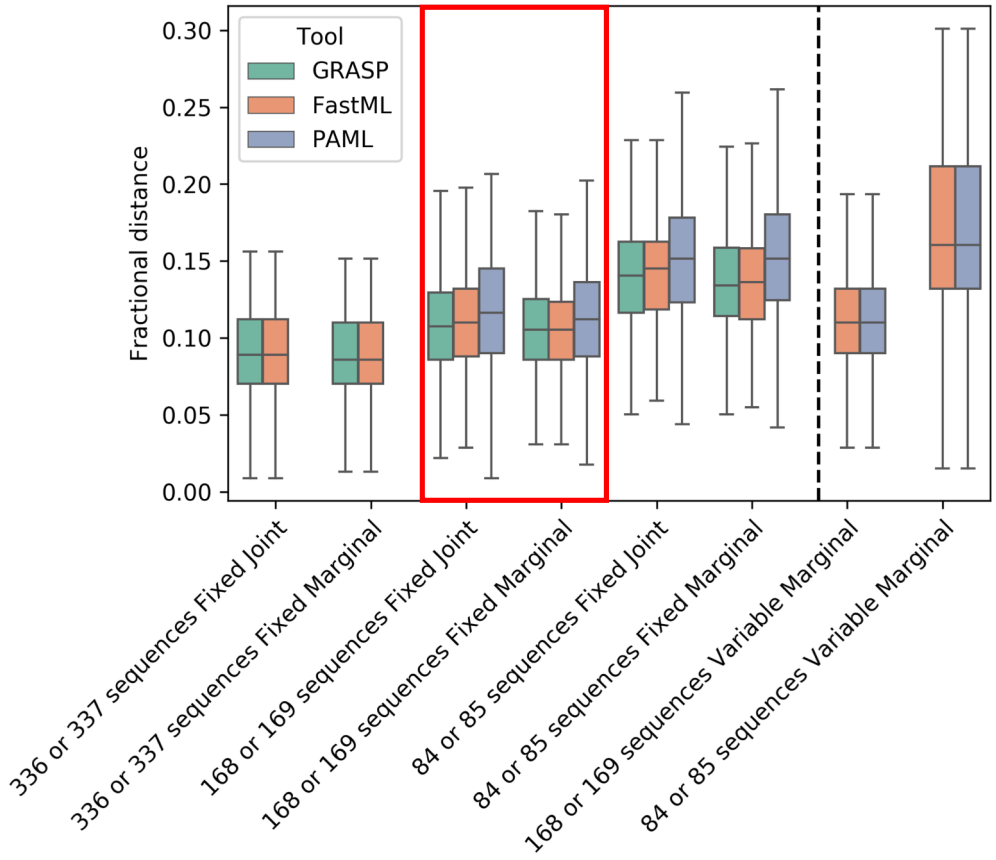




Methods' consensus standard:

Generate ancestors that are similar to those of other methods

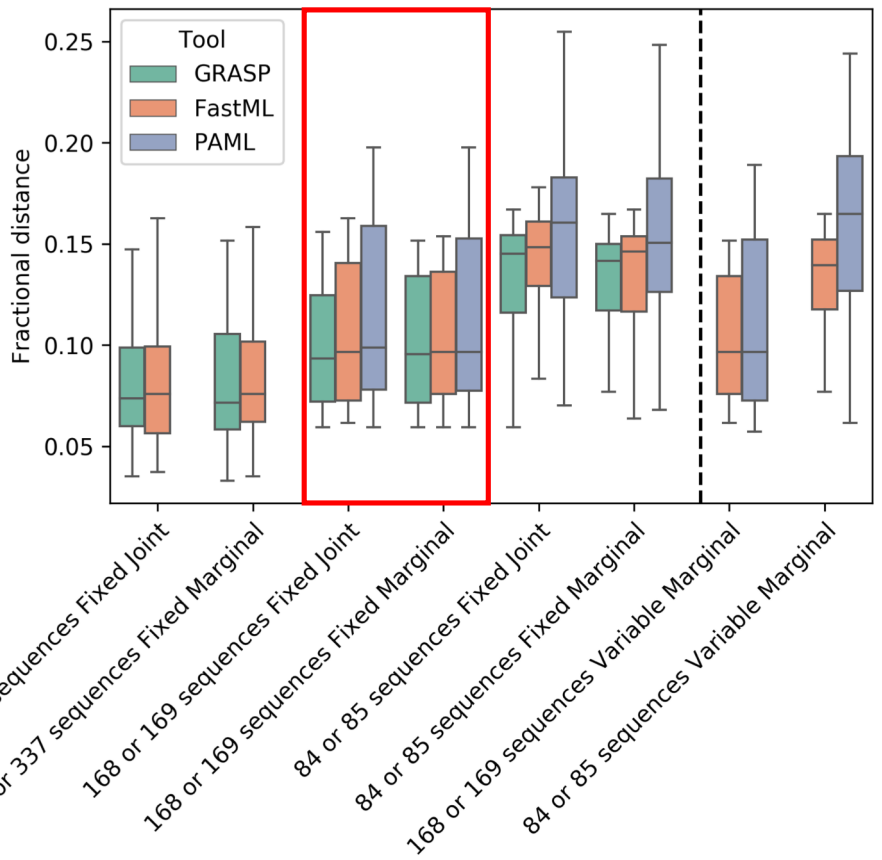
Distance between members across methods



Data consensus standard:

Generate ancestors close to that of the superset

Distance between members and superset ancestor



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

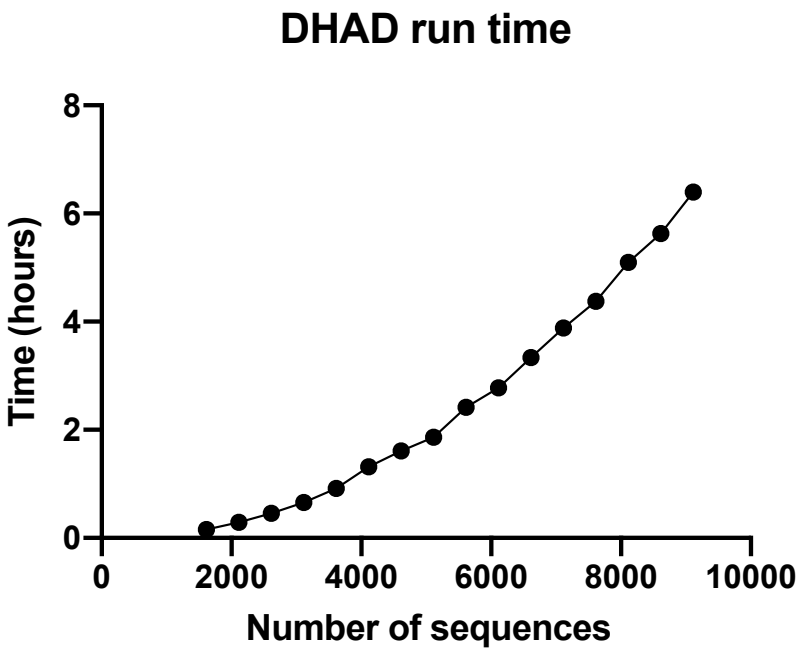
- Large volumes of data
  - mean more insertion and deletions, and
  - require attention to the computational expense of phylogenetic inference
- Inference engine in GRASP is efficient
  - uses any standard evolutionary rate matrix
  - implements an algorithm that decomposes the problem into the smallest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

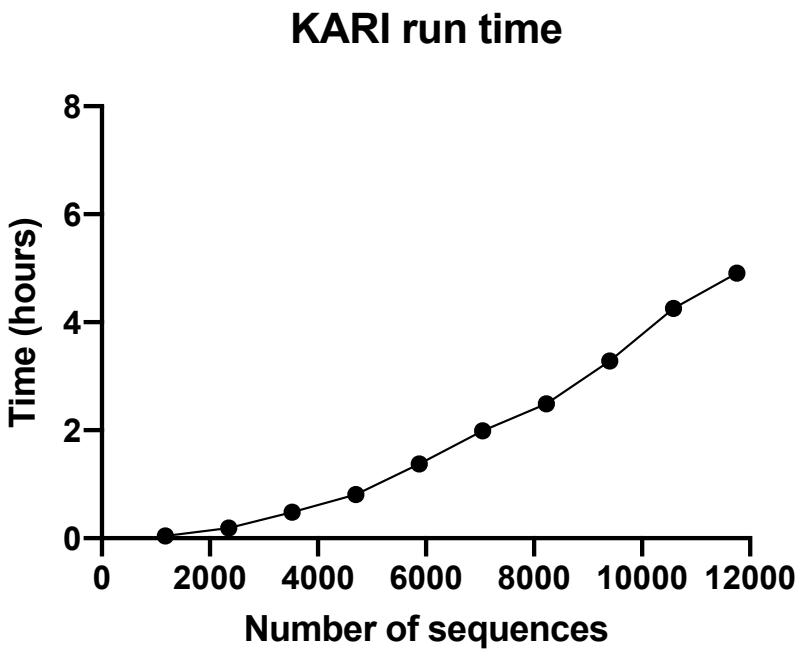
Capacity of popular tools is typically limited to less than a thousand sequences. What are the limits?

GRASP

Graphical representation of ancestral sequence predictions

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





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

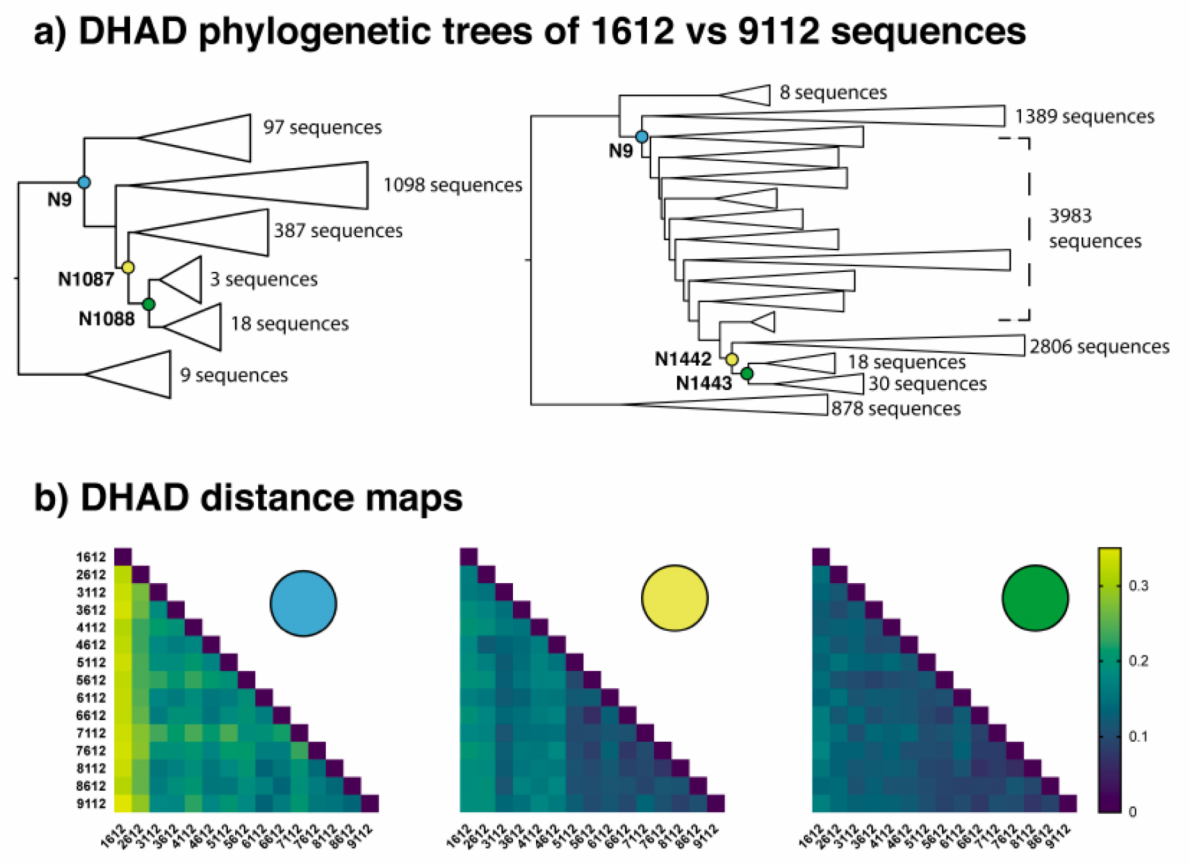
  (visual exploration offered in web service)

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

Hypothesis 2:

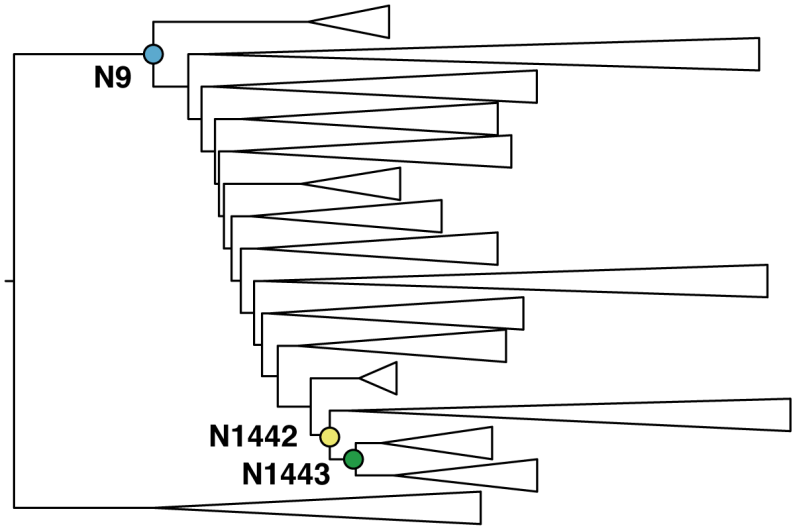
More sequences, accounting for indels improves robustness of ancestor inference; also enables a new approach to survey activity of members "by-ancestor proxy"

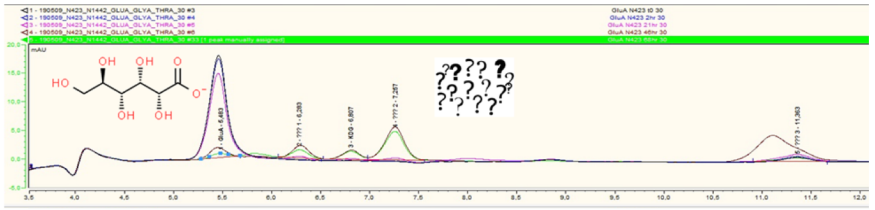
Increasing data volume constrains ancestral predictions



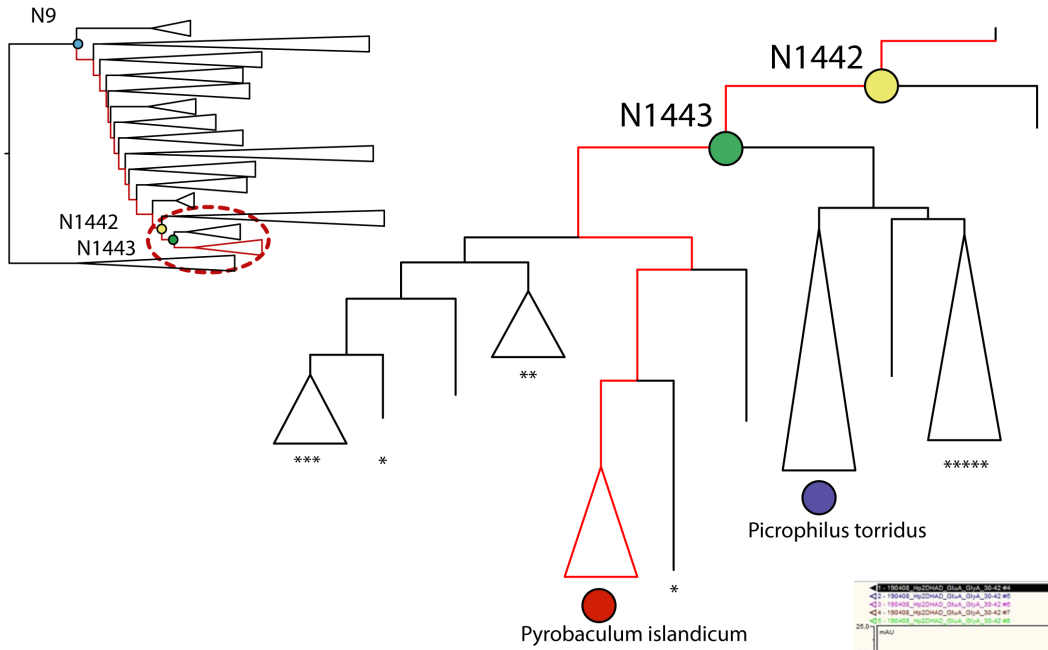
Surveying function across clades?

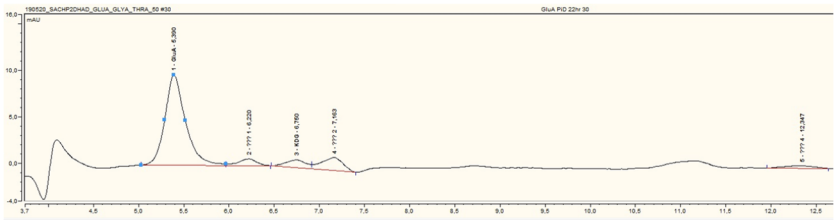
- Two reconstructions at the extremes (585 v 9112 sequences)
  - Characterised three ancestors in each, “middle” ancestor had 72.6% sequence identity
- Of 9112 sequences only
  - ~5% have high quality annotations (SwissProt)
  - 22 organisms are experimentally characterised (BRENDA database)



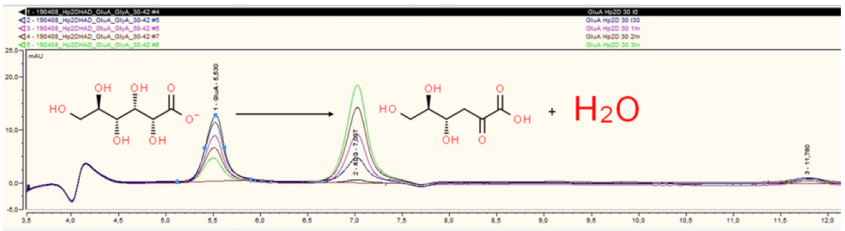


N1442 DHAD incubation





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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Boden lab

Gabriel Foley
Ariane Mora
Marnie Lamprecht
Julian Zaugg
Alexandra Essebier
Brad Balderson
Rhys Newell

UQ Gillam lab

Elizabeth Gillam
Connie Ross
Raine Thomson
Luke Guddat
Gary Schenk
Bostjan Kobe
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)

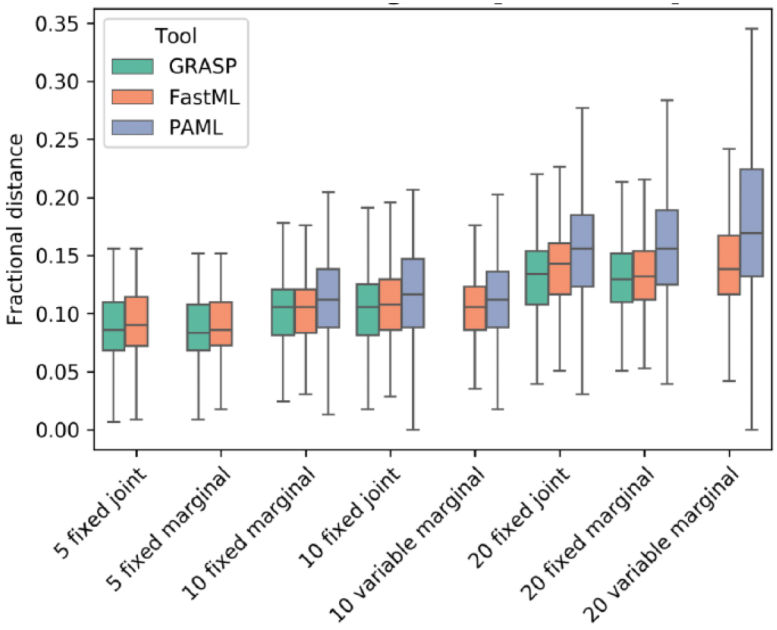
GRASP
Graphical representation of ancestral sequence predictions

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