



Shedding light on the dark landscape of tree space



quality-diversity illumination visualizes phylogenomic tree incongruence

Russell Y. Neches and Motomu Matsui

Institute for Chemical Research, Kyoto University

russell@vort.org

code: hifuku.vort.org

Abstract : Visualizing likelihood landscapes

The premise of exploratory data analysis is that a picture can surface relevant structure that no summary was designed to report. Hillis *et al.* (2005) brought this to phylogenetics: a two-dimensional layout of a tree collection reveals tree islands, posterior spread, and gene-to-gene disagreement; treespace generalized it to any tree metric. Here we develop this further, with fully realized likelihood landscapes illuminated on reusable, embeddable charts. The software is written in pure Python with JIT-compiled CUDA kernels, and renders realistic phylogenomic problems in minutes.

1 Worked example : The NCLDV core genes

Guglielmini *et al.* (2019) asked whether the conserved NCLDV genes record a single vertically-inherited core, or several discordant histories. To answer, they measured congruence by scoring each single-gene tree against the concatenated reference, tabulating by hand how many reference clades each tree recovered. Here we show how quality-diversity (QD) illumination on an anchor-plane projection offers a geometric alternative; tree congruence and likelihood distributions can be compared simultaneously for all candidate marker genes.

2 Quality-Diversity Illumination

Three anchor trees fix a two-dimensional chart. A normalized distance metric d constructed from the clade branch-score (CBS) and Robinson–Foulds metric (RF) places every tree by its distances to the anchors; both terms are squared-Euclidean, so the chart is flat by construction. A parameter w weights the distance metric between topology and branch scores. We use an equal weighting ($w = \frac{1}{2}$) here. An out-of-plane residual z measures divergence from the chart.

$$d^2 = (1 - w) \frac{CBS^2}{C_{ref}} + w \frac{RF}{R_{ref}} \quad z = \sqrt{\max(0, -c)}$$

Quality-diversity illumination then fills the chart, keeping the best tree in every niche. Its output is a map of likelihood, not one optimum.

```
loop:
  pick a filled niche; copy and vary its tree
  score the variant: its niche, its logL
  keep it if the niche is empty or logL improves
```

3 Three illuminated landscapes

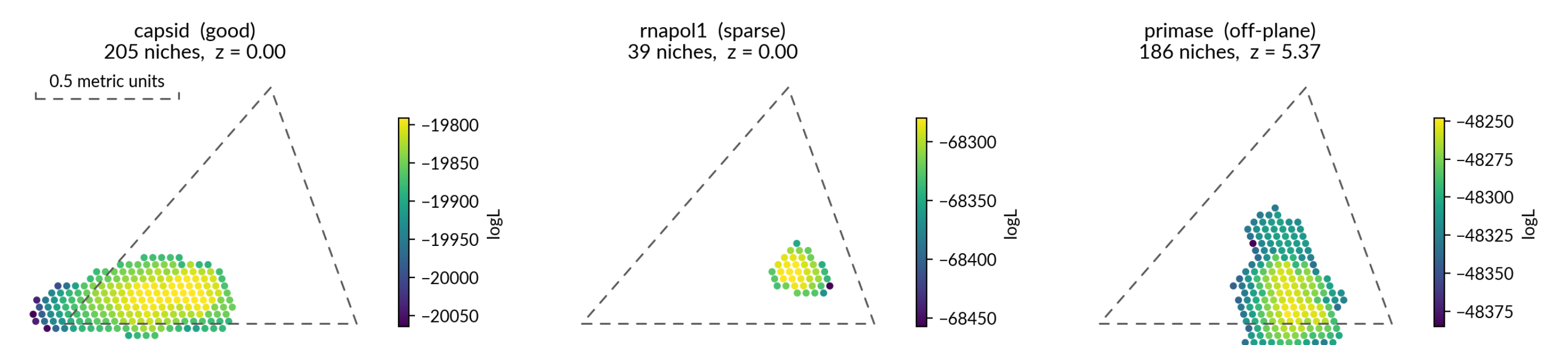


Fig. 1: *capsid* fills the chart. *rnapol1* has weak signal. *primase* sits off the chart plane.

The three corners of the anchor frame are the NJ trees for *capsid*, *tf2s* and *vltf3*. Each marker is a “niche” that sits at the barycentric combination $X = \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3$ of the anchors; a tree near a corner resembles that anchor. The filled region shows where a gene’s likely trees lie. For *capsid*, illumination reveals a broad landscape, but not for *rnapol1*. *primase* is well explored, but its trees lie far off the plane ($z = 5.4$). Its niches are projections; the flat chart is lossy. The full set of anchor distances span about five dimensions, and so it is important to select the best chart for each gene.

4 Assembling a multi-gene map

No plane holds both the distances and the neighborhoods of a high-dimensional cloud. We choose to preserve neighborhoods; nearby points represent trees with similar topologies. Highly dissimilar trees will have distorted distances, but will never be neighbors.

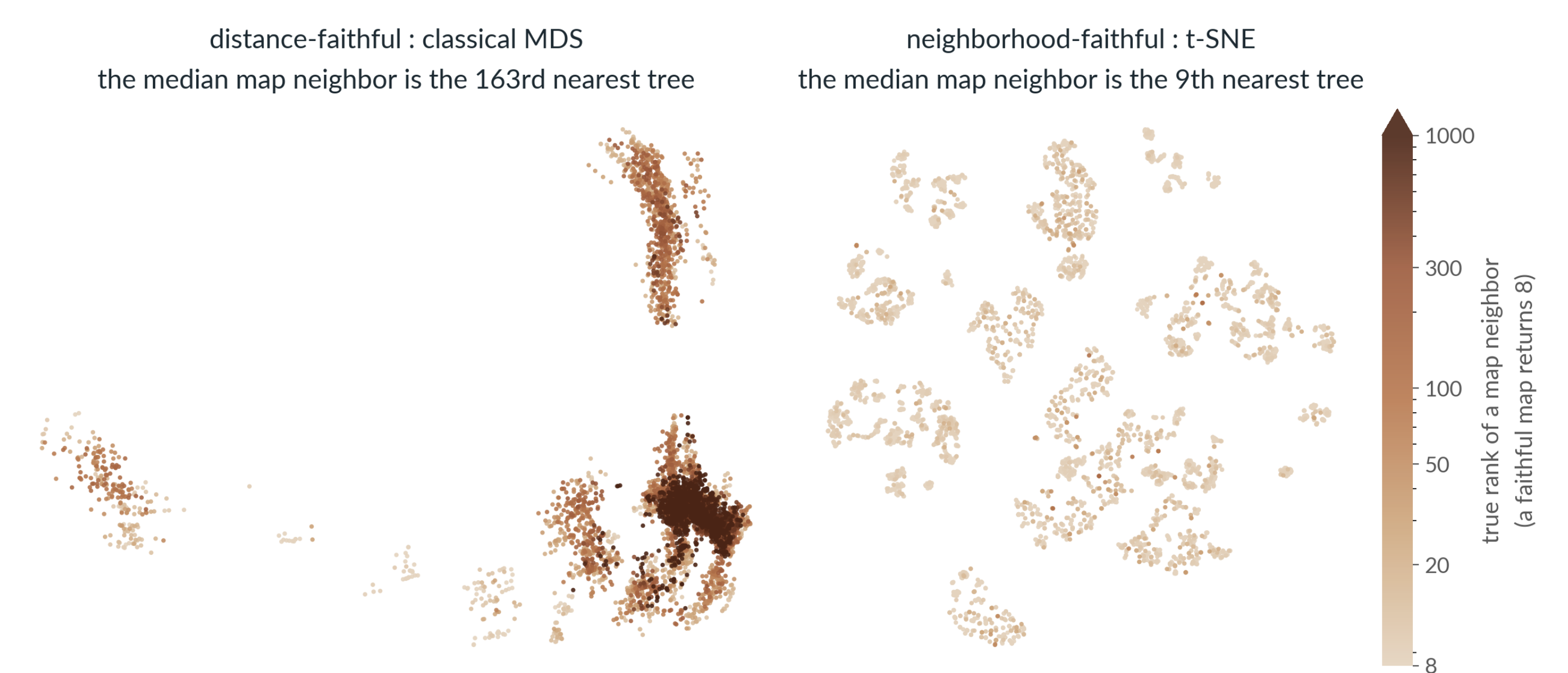


Fig. 2: 5772 trees, colored by the true tree-space rank of the neighbors each map draws (faithful = 8). MDS keeps the largest distances (0.82) but a quarter of its adjacencies join trees ranked beyond 1000. t-SNE sacrifices global distance fidelity, but preserves neighborhoods.

5 Ten gene likelihood landscape map

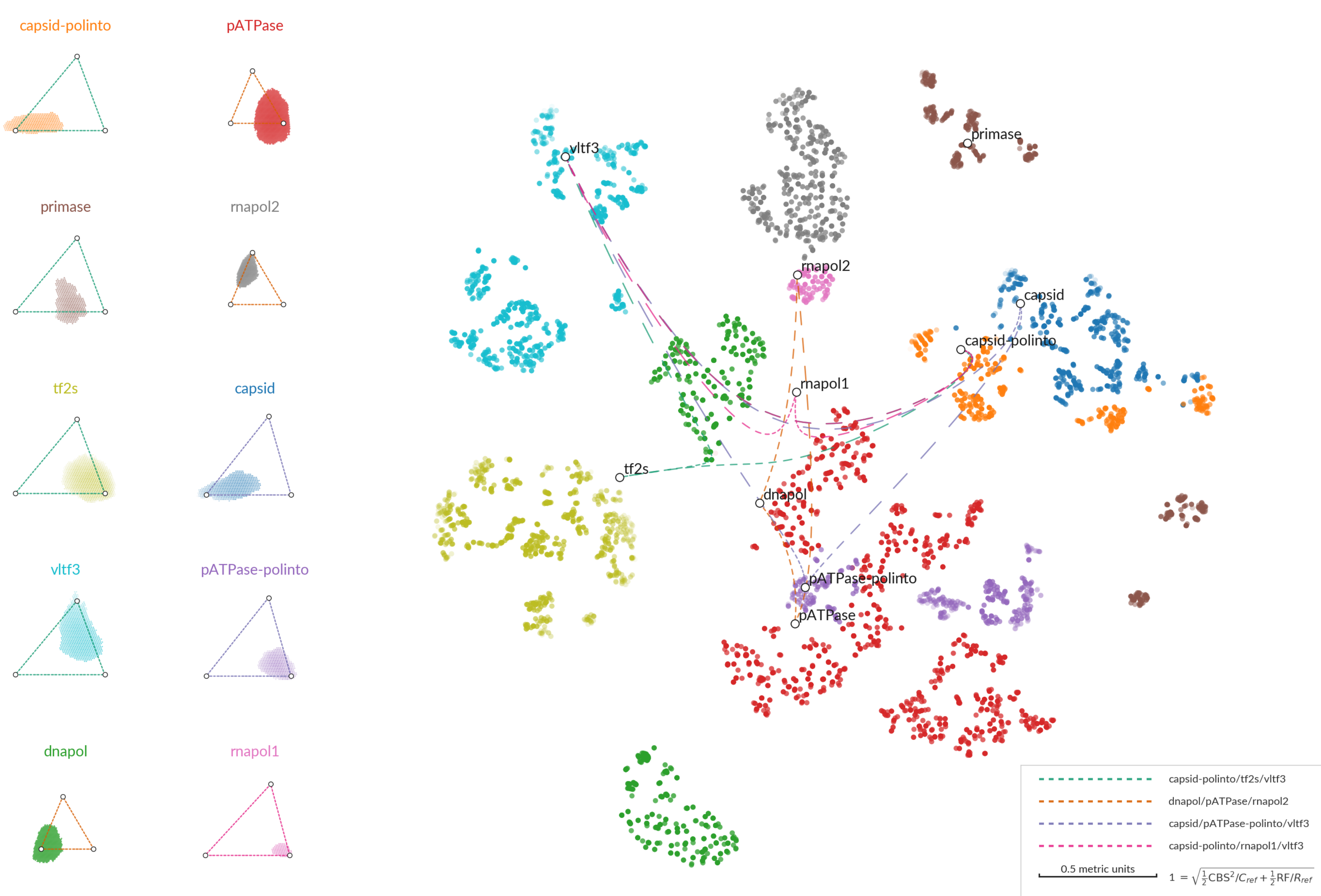


Fig. 3: Ten NCLDV core genes on one map. Nearby genes share phylogenetic signal; distant genes conflict. *primase* sits apart, the divergent gene. Left: the illuminated surveys in their anchor frames. Right: all surveys projected into common coordinates by t-SNE, the anchor frames drawn as a graticule that shows the projection distortion. Opacity encodes log-likelihood.

6 New ideas

- Quality-Diversity Illumination applied to phylogenetic inference
- A shared 2D chart, flat by construction, with an out-of-plane fidelity measure
- Many charts stitched into one map with projection distortion drawn as a legible graticule

References

QD illumination	Mouret & Clune 2015
tree likelihood	Felsenstein 1981
branch-score distance	Kuhner & Felsenstein 1994
topological distance	Robinson & Foulds 1981
NJ anchors	Saitou & Nei 1987
tree-space geometry	Billera <i>et al.</i> 2001
visualizing tree space	Hillis <i>et al.</i> 2005
tree landscapes	Jombart <i>et al.</i> 2017
t-SNE	van der Maaten & Hinton 2008
NCLDV core genes	Guglielmini <i>et al.</i> 2019

Acknowledgements

This work was supported by JST DICP Grant Number JPMJND2206, GteX Grant Number JPMJGX23B2, and JSPS KAKENHI Grant Numbers 23KK0122, 24K01889, and 25K09119 to M.M.