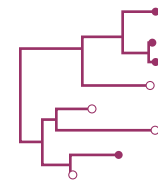
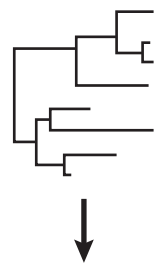
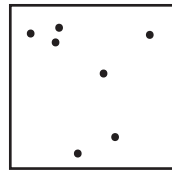


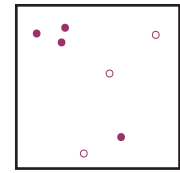
Phylogenetic reconstructions are interpreted as distance matrices, and the subset is tagged.



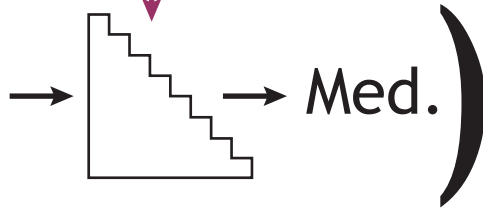
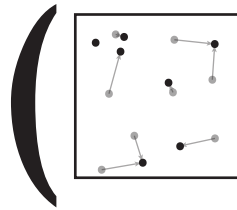
Distances are re-interpreted as points in a high-dimensional space to avoid information loss.



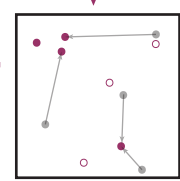
High-dimensional Nonmetric MDS



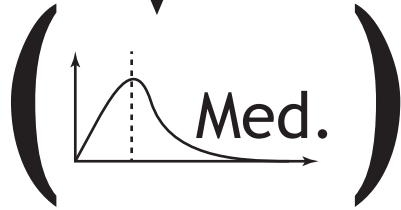
Random sampling in the new space is associated to the closest real (or labeled) points.



x1000



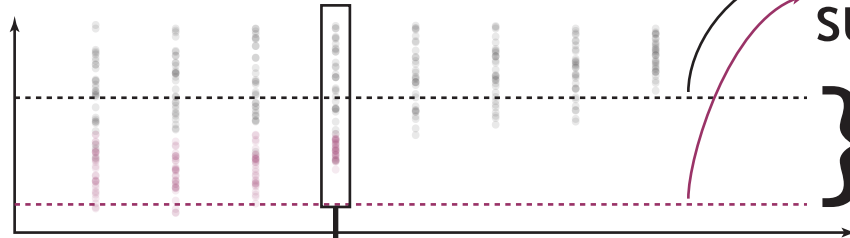
The same random sampling is repeated at different sampling efforts.



for {0 .. #Genomes}

The saturation of the samples can now be compared qualitatively against the subset.

Med



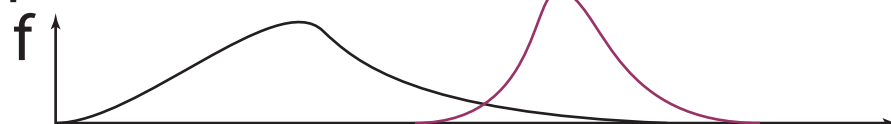
Med. w/o subsampling

Naïve difference

Sampled genomes

Empirical PDFs of distances at the same sampling effort can now be computed.

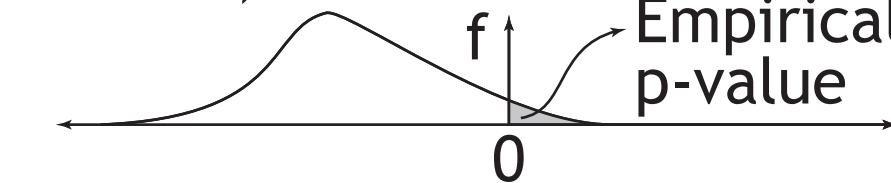
Empirical PDFs



Median distance

The distribution of the difference between total and subset can now be estimated.

(Med - Med) convolution



Normalized difference

Empirical p-value