More genes or more taxa?

• Choosing the amount of information
• Choosing your ingroup/outgroup taxa
• (morphology/behavior/molecules...)
• Choice of data

Designing your phylogenetic analyses


Figure 2. Absolute and adjusted error as a function of taxon sample size for a dataset generated with the HKY model of evolution on the Rosendal tree topology. Weighted parsimony data points represent the average error over all subsample diameters.
Methods of Phylogenomic Inference

- Alignment
- Orthologous genes
- Homology assessment
- DNA-string comparisons
- Methods based on whole-genome features
- Methods based on gene content
- Methods based on gene order
- LINE and SINE integration
- Gene expression and tissue
- Rare genomic changes
- Repeat position indices and signatures
- Gene 4
- Gene 3
- Gene 2
- Gene 1

Supermatrix
- Orthologs
- Large-scale gene tracing
- Organisms
Figure 3 | Phylogenomics and the resolution of...


Deluc F, Brinkmann H, Chourot D, and Philippe H. 2006. Tunicates and not cephalochordates are the closest


Delsuc F, Brinkmann H, Chourout D, and Philippe H. 2006. Tunicates and not cephalochordates are the closest


Xenoturbella EST analysis
Gene accumulation curves

(Repeat with 50 random addition replicates)
ESTs processing

• Raw trace files need to be processed:
  - trace2dbEST

• EST reads are then assembled using PartiGene
  - (with original quality files allowed)

• PartiGene calls CLOBB to assign sequences to clusters of like EST sequences and PHRAP to assemble the EST sequences assigned to each cluster into a contiguous sequence for that cluster.

• The single contiguous sequence for each cluster is then translated with prot4EST.

ESTs processing
Orthology assignment

- A local database of all Homo sapiens, Canis familiaris, Gallus gallus, Drosophila melanogaster, and Anopheles gambiae sequences that have orthology assignments in Homologene is constructed.
- Masked sequences were queried against these sequences with blastp.
- The blast results were passed to TribeMCL for Markov Chain Clustering (MCL). The MCL inflation parameter was varied in intervals of 0.1 from 1.4 to 2.5 to identify the value that generated the maximum number of clusters.
- TribeMCL groups that had homologene sequences that were assigned to multiple TribeMCL groups were discarded, as were groups that had sequences from the homologene dataset.
- Only groups with sequences from at least 33% of the taxa are preserved.
- The number of sequences for each species represented within each TribeMCL group is calculated, and groups with a median greater than one are discarded.

A local database of all Homo sapiens, Canis familiaris, Gallus gallus, Drosophila melanogaster, and Anopheles gambiae sequences that have orthology assignments in Homologene is constructed.
After passing the taxon sampling and orthology criteria, we obtained 196 genes for 56 species. On average, each species was represented by 48% of the genes included in the 146-gene matrix of Philippe et al. (2005). Our matrix of 196 genes contains only 50 of the genes included in the combined matrix consists of 32,975 amino acids (19,362 positions are parsimony-informative). The combined matrix contains 32,975 amino acids (19,362 genes). On average, each species was represented by 48% of the 196 genes obtained for 56 species. After passing the taxon sampling and orthology criteria, we...