Evolutionary selective constraints acting on the stop codon across land plants

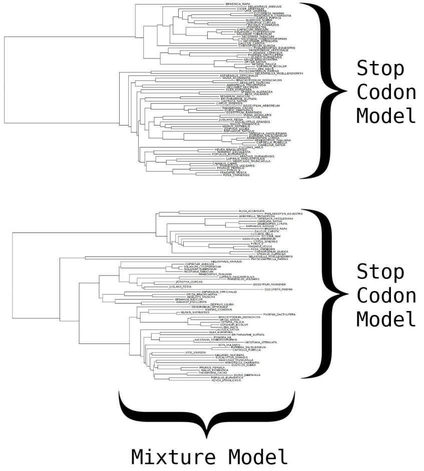
Presented by: Vishvesh Karthik

Motivation

- Common substitution models exclude stop codons.
- Stop codons suppression is more common*
- Read-through mechanisms and ribosomal stalling act as protein regulatory mechanisms$^+$
- Egs: AMD1, selenoproteins etc[1]
- Stop codons involvement in protein synthesis is understated

COA: Course of action

- Plant orthologs were obtained[1]
- An extended model of substitution was applied[2]
- A mixture model using the estimated model parameters was implemented on the stop codons
- The final estimates are bootstrapped for certainty



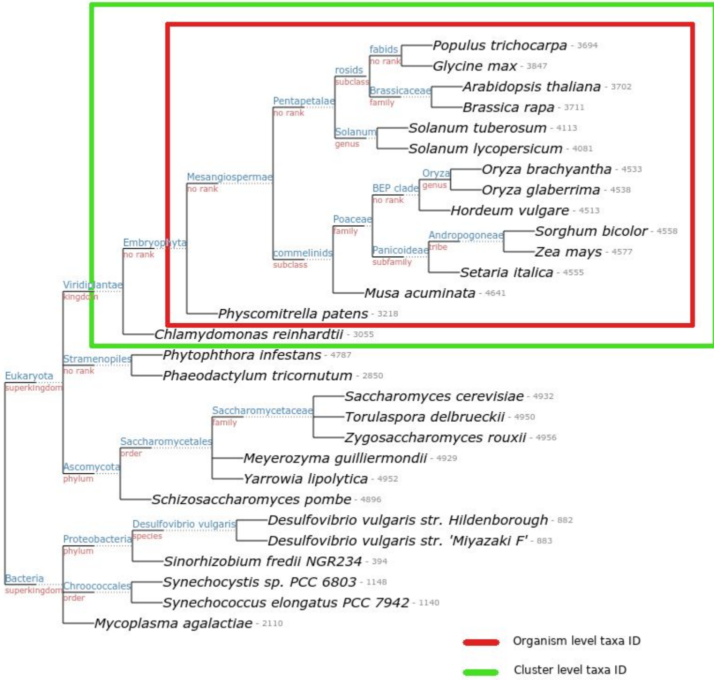
Stop Codon Model

Stop Codon Model

Mixture Model

Methods

- Flat files are downloaded from OrthoDB[1]
- Convert “gene-based” clusters to “organism-based” clusters
    - “CLUSTER_ID:ORG_ID_ID:GENE_ID” -> “CLUSTER_ID:ORG_ID”
- Get counts[2]
    - “CLUSTER:NO_ORGS”
    - “ORG:NO_CLUSTERS”

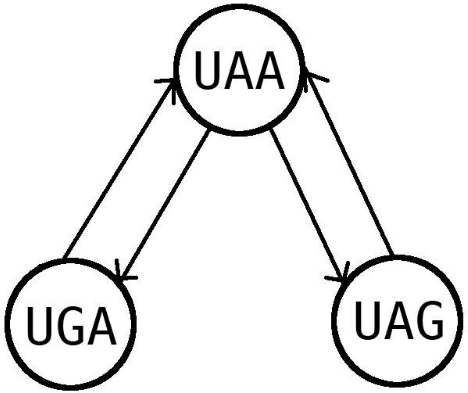


Methods

- Select clusters which fall under “Viridiplantae” taxonomic division
- Rank organisms and choose top 40
- Rechoose the clusters based on selected organisms[1]

Methods

$$\mathrm{q}_{ij} = \begin{cases} \pi_{jk} \text{ , synonymous transversion } (i, j \in S) \\ \kappa\pi_{jk} \text{ , synonymous transition } (i, j \in S) \\ \omega\pi_{jk} \text{ , non-synonymous transversion } (i, j \in S) \\ \kappa\omega\pi_{jk} \text{ , non-synonymous transition } (i, j \in S) \\ 0 \text{ , } > \text{ 1 nucleotide difference} \\ \phi\kappa\pi_{jk} \text{ , synonymous transition } (i, j \in N) \\ 0 \text{ , i } \in \mathrm{N} \oplus \text{ j} \in \mathrm{N} \end{cases}$$



Methods

- For each cluster:
    - Lookup each gene ID on OrthoDB using API and fetch NCBI gene ID (if available)
        - Download the NT CDS from NCBI
        - Perform a check to see if the sequence is in-frame and has a stop codon

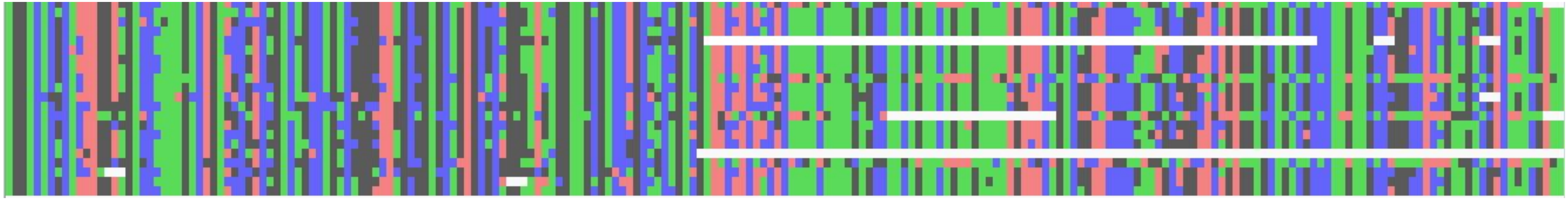
Methods

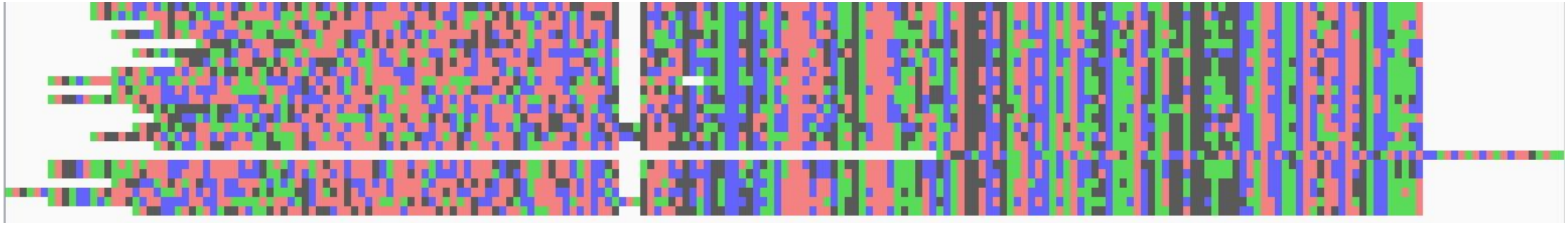
- For each cluster:
    - Perform an initial alignment to estimate the general relativity of the sequences
    - Divide(sub-cluster) the sequences based on the aligned stop codon positions
    - Discard the sub-clusters which do not have more than 3 sequences

Methods - Caveats - Why was the above done?

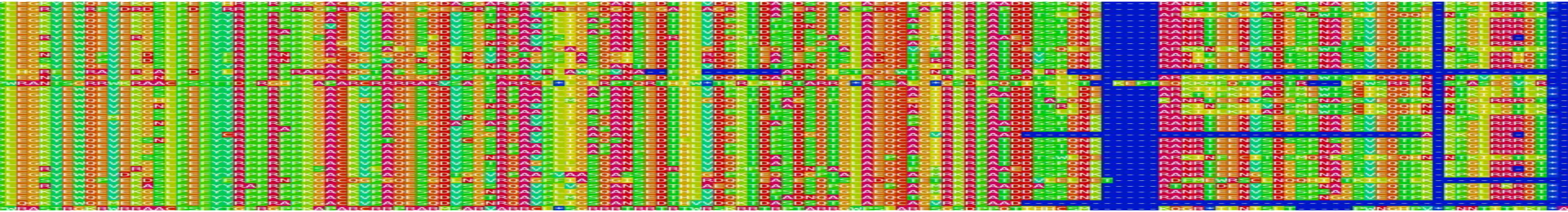
- The extMG, extended from Muse & Gaut model, assumes
  - The last codon in an alignment is either a stop codon or gaps.
  - Sequences have no in-frame nonsense mutations
  - Di-nucleotide substitutions are not possible
- Plant sequences utilize the degeneracy of codons to the full extent (how? contd.)

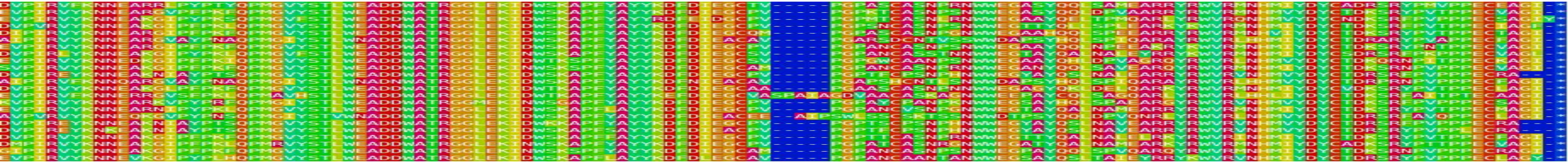
NT Alignments





Protein Alignments





Methods

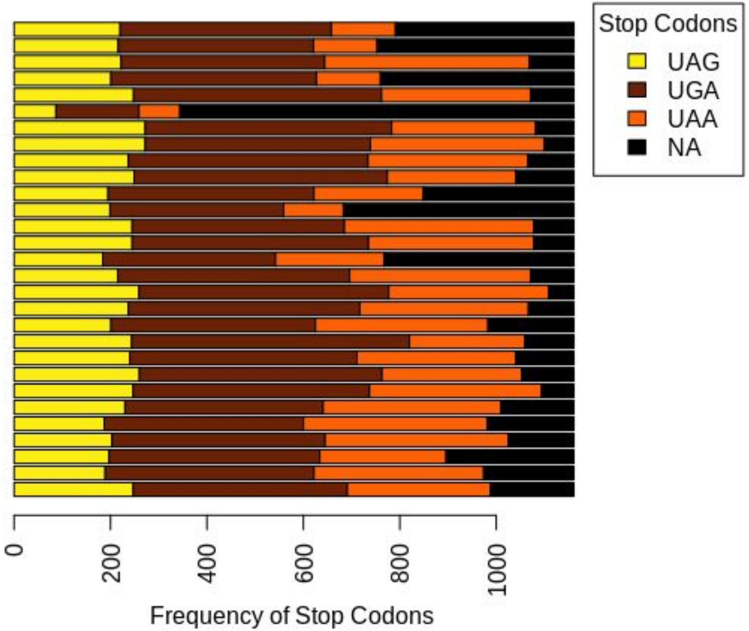
- Protein alignments express inter-cluster relativity well
- Except for the conserved regions, nucleotide sequences of plants, are hard to align
- Thus to gain common ground, sequences are aligned based on AAs and the alignments are mapped back to nucleotide sequences.

Methods

- For each cluster:
  - Apply the extMG to the aligned NT sequences and save the parameters
  - Infer phylogenetic trees using codonPhyML
- For all clusters:
  - Apply the mixture model
  - Perform bootstrapping

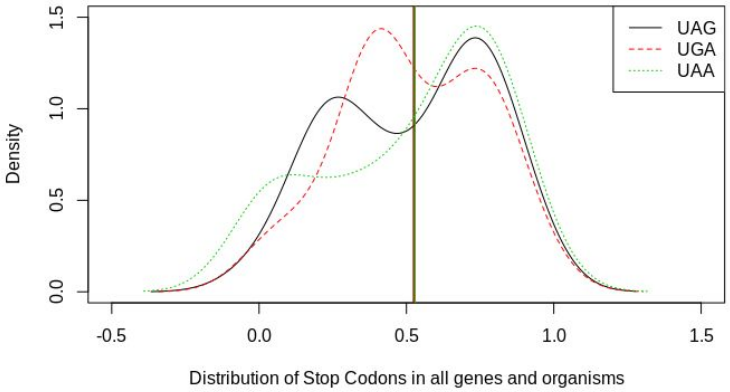
Results

- UGA is the most abundant stop codon



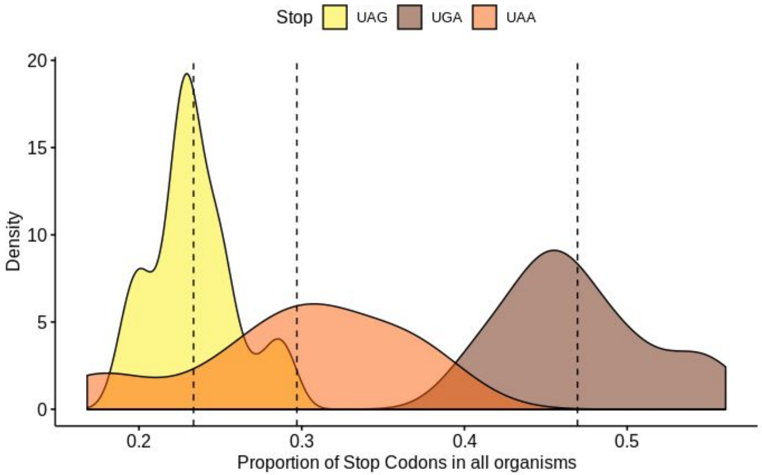
Results

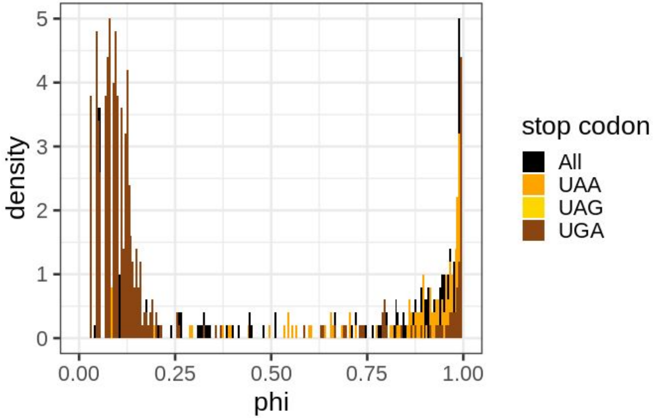
- UGA is the most abundant stop codon
- The selected organisms can be divided into two groups based on codon preference[1]



Results

- The mixture model estimates a ϕ of 0.3
- ~60% of the genes are under purifying selection[1]
    - 50% of UGA is preserved
    - 70% UAG -> UAA
    - 60% UAA -> UGA/UAG

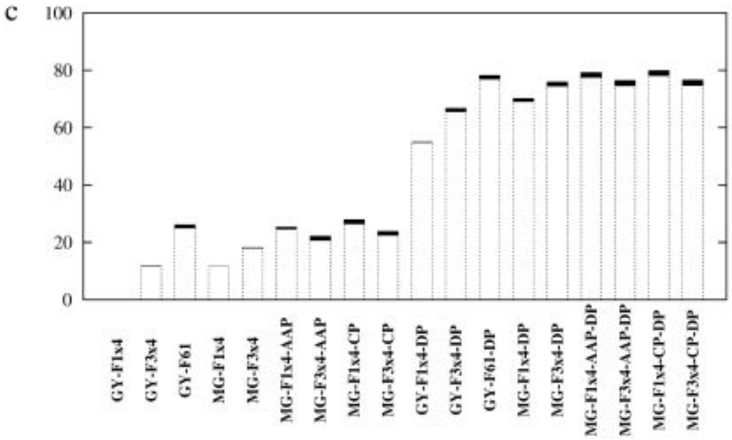




Rodrigue N, Lartillot N, Philippe H. Bayesian comparisons of codon substitution models. *Genetics*. 2008;180(3):1579–1591.

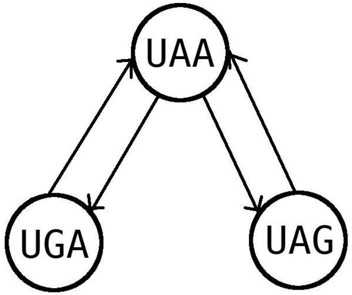
Discussion

- MG by itself is not the best performing model, thus it can be improved by adding codon-preference statistics to improve estimation results



Discussion

- The markov chain for stop codons can be made irreducible and the TPM can be shrinked to 4 dimensions by including a 4th state “NNN” which warrants transition from any state to any other state



Conclusion

- The extMG model provides a chance to predict additive effects of stop codons.
- This information can be used to isolate genes which might be under readthrough contexts.
- Would unlock more information as to the existence of those genes and such mechanisms.
- Can also be used to predict whether a gene is being lost or gained based on stop codon preference.

Thank You!

Questions?