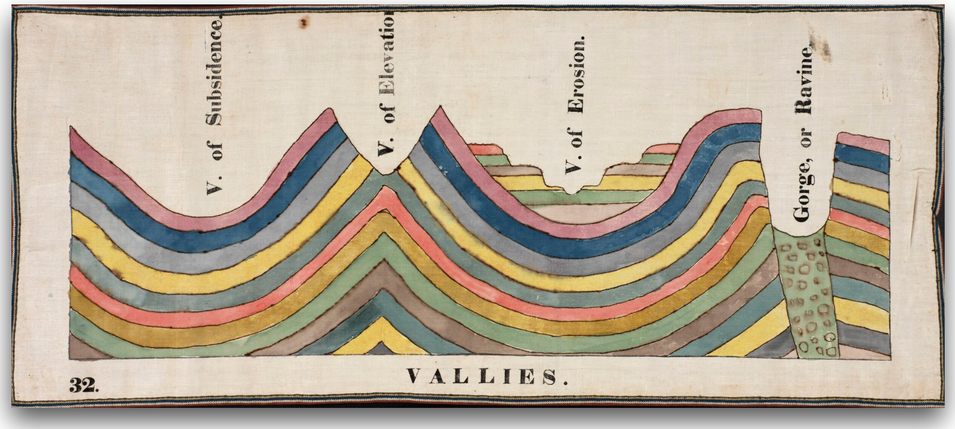
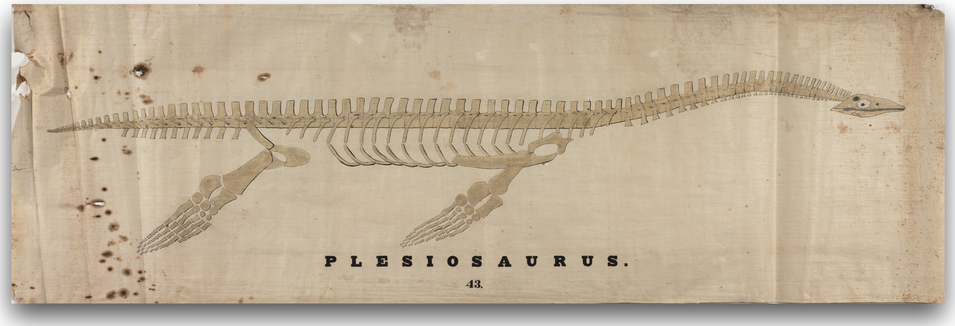
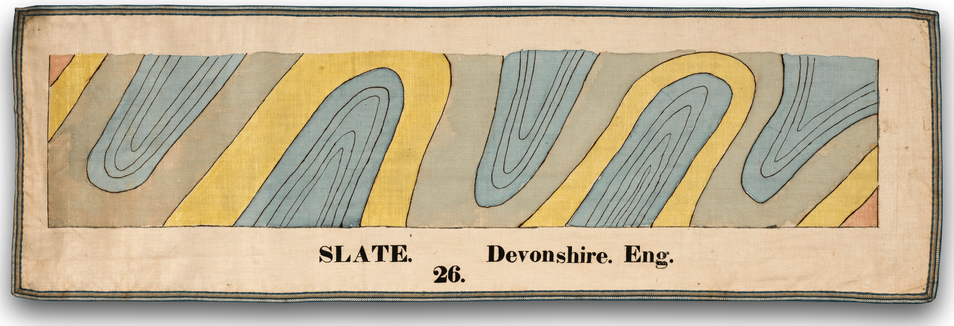






Complex models

Sequence data evolve under **heterotachy**!

That is **rate variation across sites** - and lineages of course!

Heterotachy is known to **mislead** phylogenetic inference …

… and we need **complex models 🤯 to deal with heterotachy!**

Partition Models

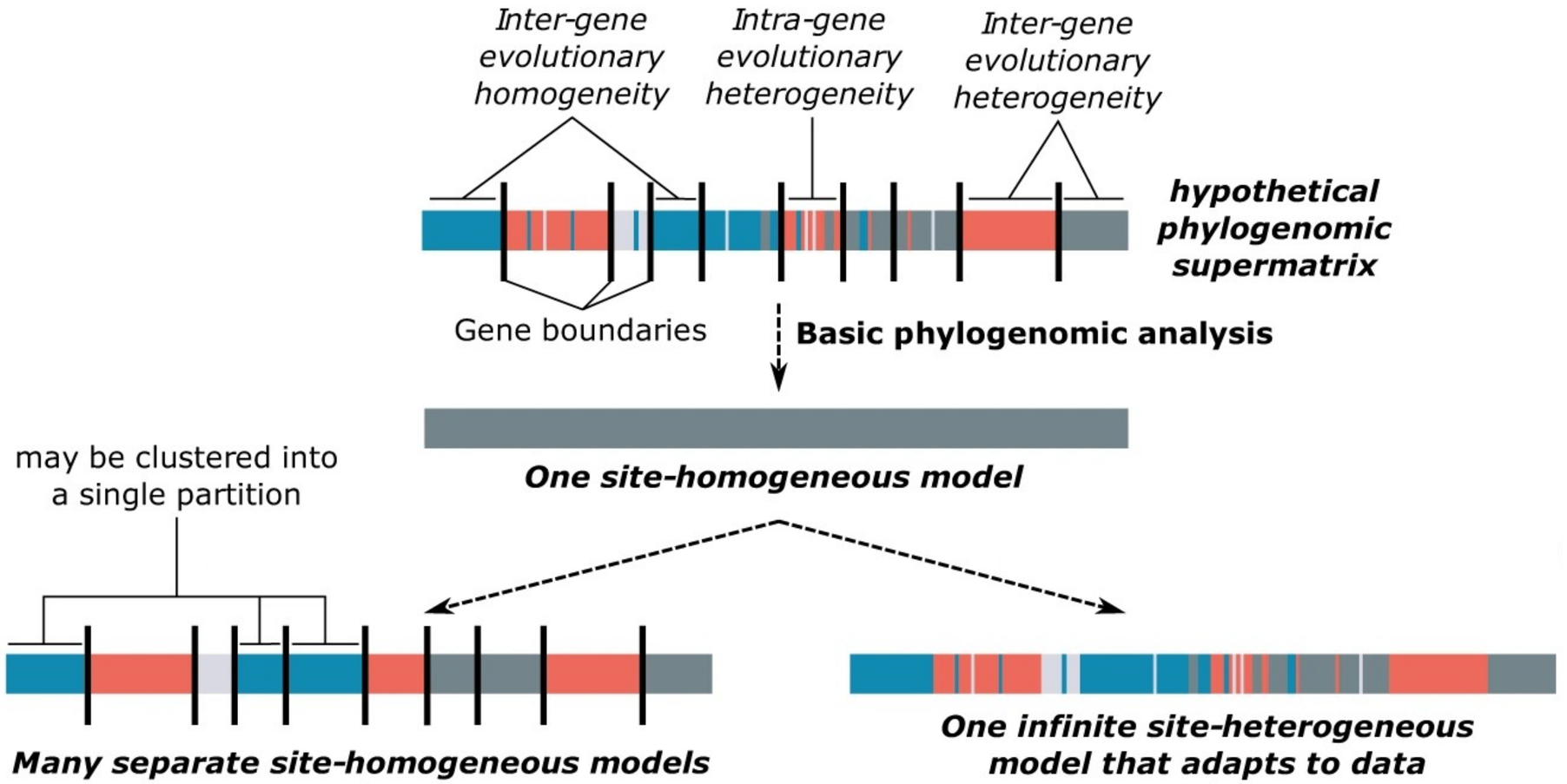
- Divide a multiple sequence alignment into distinct subsets (e.g. genes or codon positions).
- Each subset (partition) evolves under its own substitution model and parameters.
- Useful when prior knowledge about data structure is available.

Mixture Models

- Do not assign sites to specific subsets beforehand.
- Each site has a probability of belonging to multiple model components.
- Model heterogeneity within sites by combining multiple substitution processes.

Key Difference

- Partition models use predefined structure.
- Mixture models infer structure probabilistically.

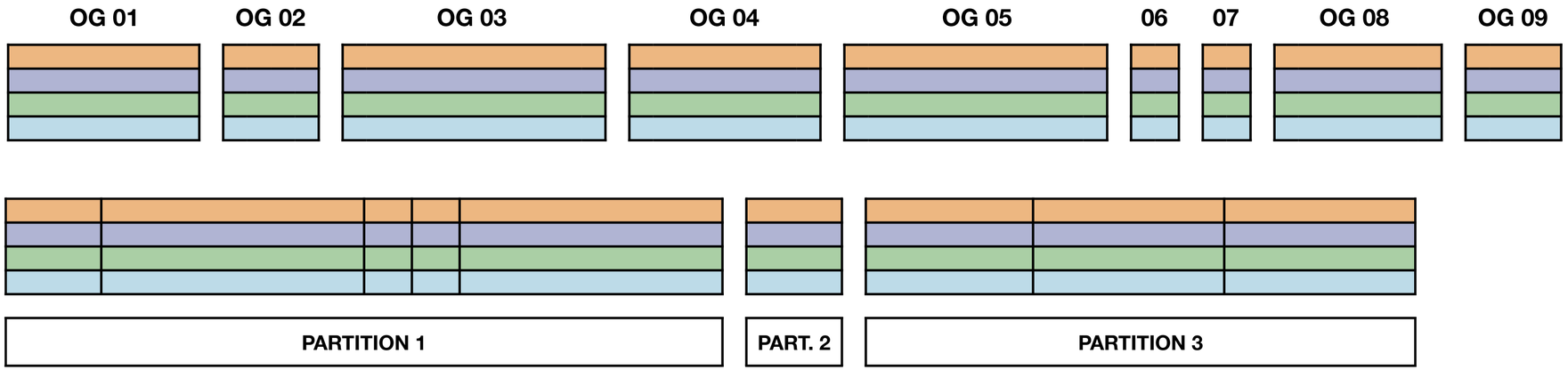

Inter-gene evolutionary homogeneity
Intra-gene evolutionary heterogeneity
Inter-gene evolutionary heterogeneity
hypothetical phylogenomic supermatrix
Gene boundaries
Basic phylogenomic analysis
One site-homogeneous model
may be clustered into a single partition
Many separate site-homogeneous models
One infinite site-heterogeneous model that adapts to data

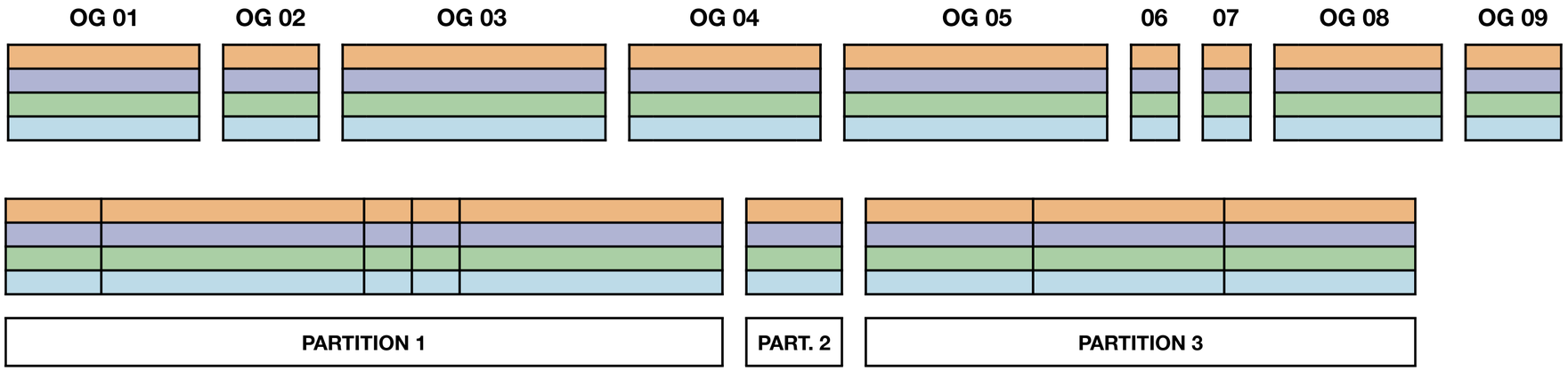
The types of **partition models** are:

- **Edge-linked partition model with equal branch lengths**
  All partitions share the same set of branch lengths.

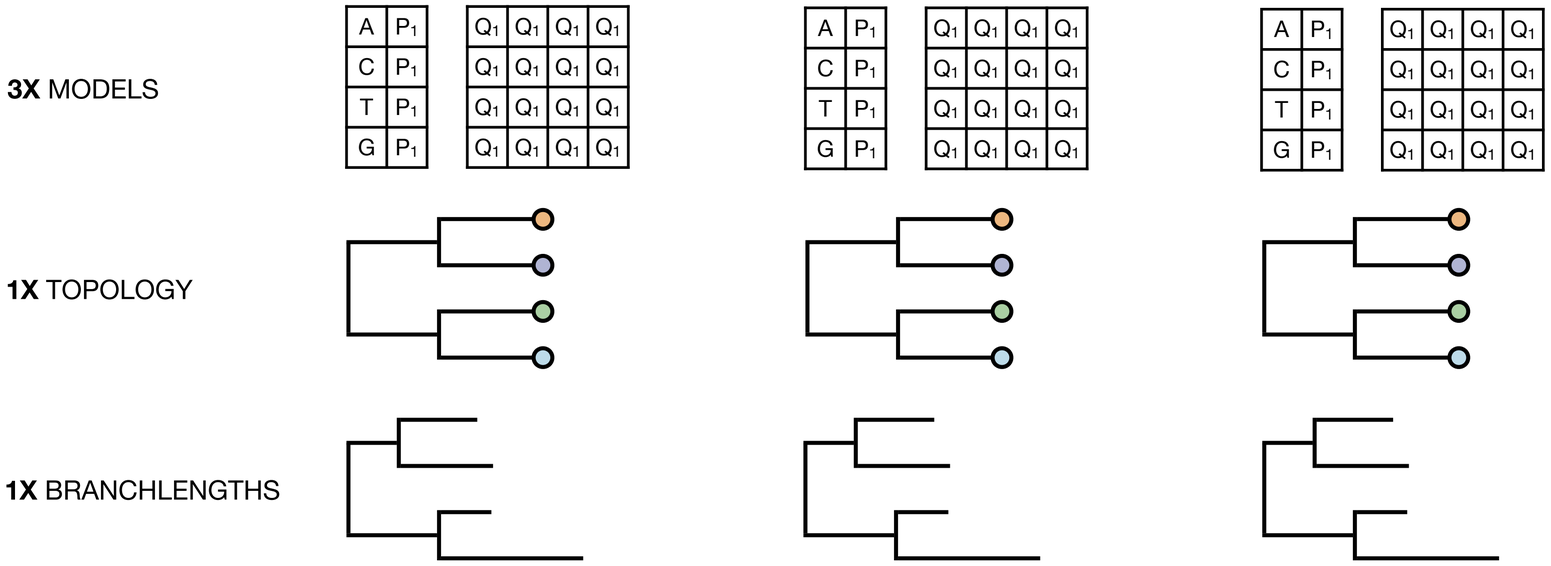
- **Edge-linked partition model with proportional branch lengths**
  Each partition has branch lengths that are proportional to a shared set.

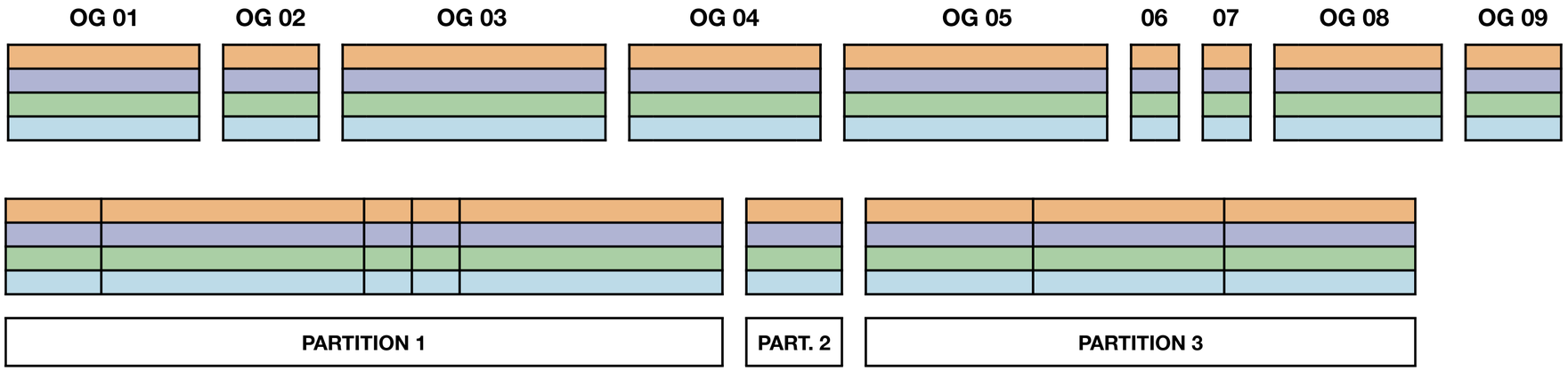
- **Edge-unlinked partition model**
  Each partition has a different set of branch lengths.


OG 01
OG 02
OG 03
OG 04
OG 05
06
07
OG 08
OG 09
PARTITION 1
PART. 2
PARTITION 3

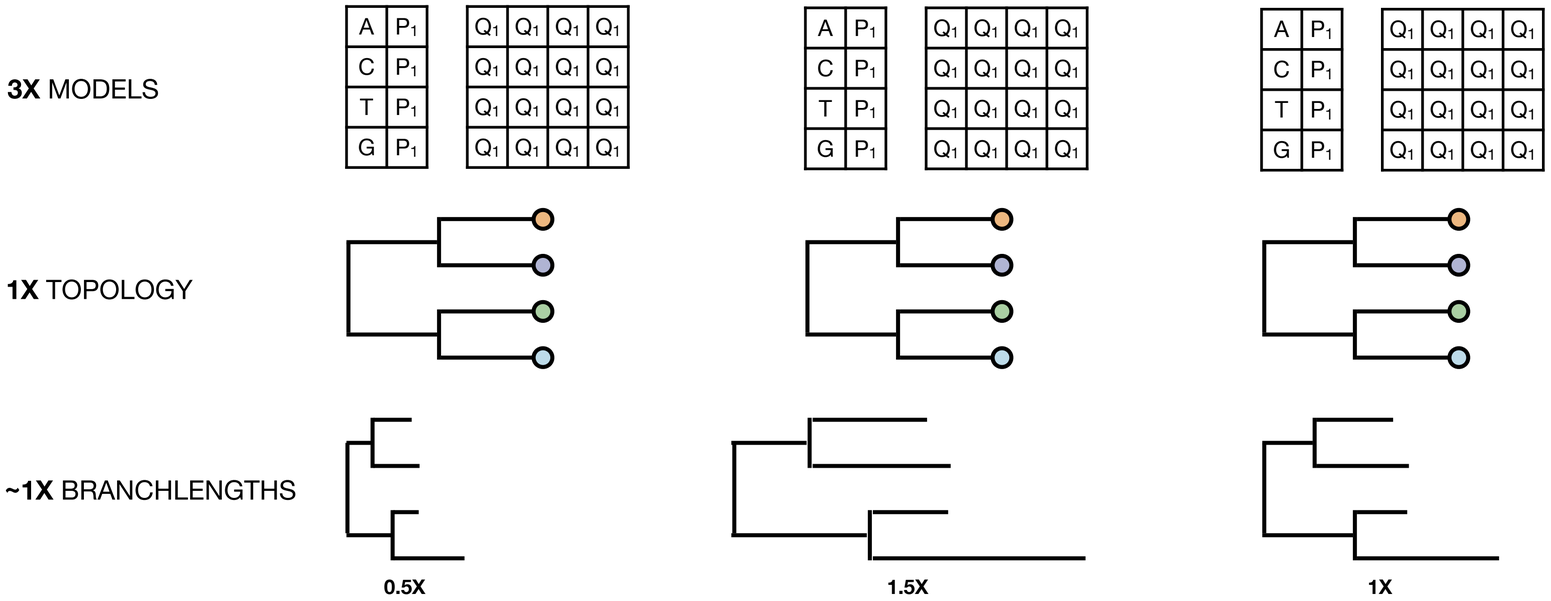


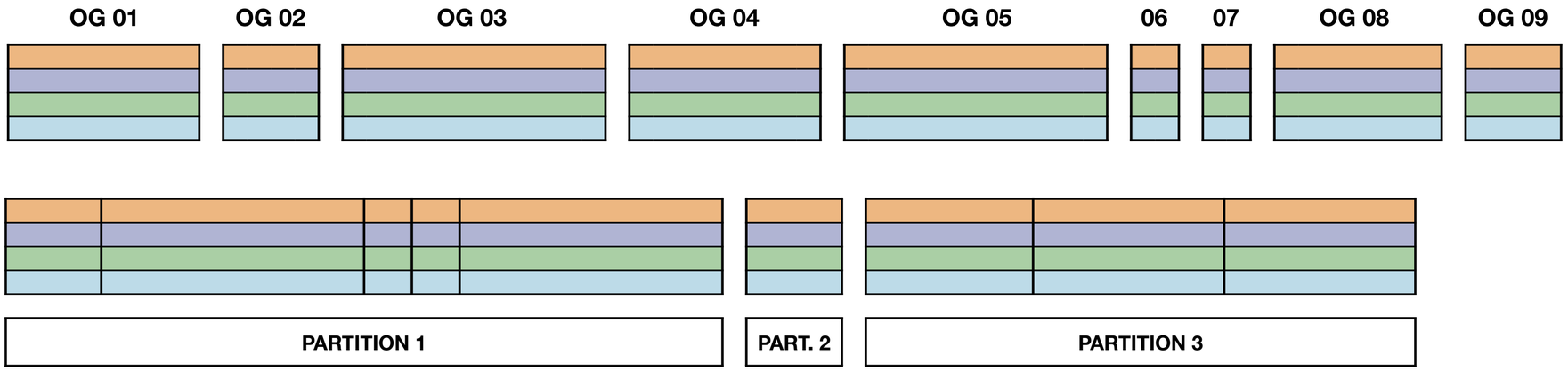
EDGE - LINKED PARTITION MODEL





EDGE - PROPORTIONAL PARTITION MODEL





EDGE - UNLINKED PARTITION MODEL

3X MODELS

A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

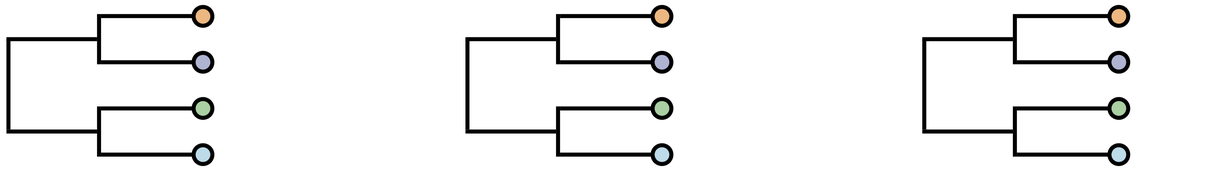
A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

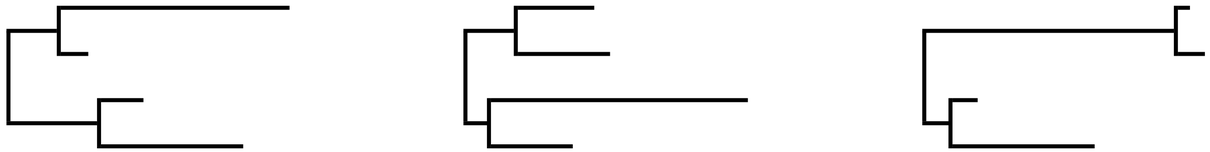
A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

1X TOPOLOGY



3X BRANCHLENGTHS



Which partition model?

- edge-equal model is typically unrealistic
  Every part of the alignment is treated as if it evolved at the same speed along the same branches. However we know that different genes, codon positions, or functional regions often evolve at different rates.

- edge-unlinked model can overfit if there are many short partitions
  Partitions contain few informative sites and not enough data in each partition to reliably estimate its own branch lengths … the model begins to fit noise in the data rather than the true evolutionary signal.

- edge-proportional model is recommended as it represents a good tradeoff between oversimplification and overparametrization 🧐

RATE MIXTURE MODEL

Site evolves under the same substitution model, but with a different rate multiplier - e.g. +G or +K.

PROFILE MIXTURE MODEL

Model variation in frequencies (aa or nt) across sites, but they share the same substitution matrix.

Empirical profile models use fixed frequency profiles derived from large datasets - e.g. LG+C60.
Bayesian profile models infer the number and composition of profiles from the data - e.g. CAT.

FULL MIXTURE MODEL

Each site may evolve under a different substitution model, not just a different profile or rate.

RATE MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C

PROFILE MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C
0.1	0.1	0.2	0.8	0.3	0.2	0.1	0.2	0.7	0.1	0.3	0.1	0.8
0.7	0.7	0.5	0.1	0.4	0.7	0.1	0.5	0.2	0.1	0.4	0.8	0.1
0.2	0.2	0.3	0.1	0.3	0.1	0.8	0.3	0.1	0.8	0.3	0.1	0.1

FULL MIXTURE MODEL

A	C	G	C	T	G	C	A	A	T	A	C	T
A	C	C	C	T	C	C	A	T	G	C	C	T
A	C	C	C	A	G	C	A	G	C	T	C	C
A	C	C	C	A	A	C	A	A	C	A	G	C
0.1	0.1	0.2	0.8	0.3	0.2	0.1	0.2	0.7	0.1	0.3	0.1	0.8
0.7	0.7	0.5	0.1	0.4	0.7	0.1	0.5	0.2	0.1	0.4	0.8	0.1
0.2	0.2	0.3	0.1	0.3	0.1	0.8	0.3	0.1	0.8	0.3	0.1	0.1

K_1

K_2

K_3

3X GAMMA CATEGORIES

A	P_1
C	P_1
T	P_1
G	P_1

A	P_2
C	P_2
T	P_2
G	P_2

A	P_3
C	P_3
T	P_3
G	P_3

3X FREQUENCIES

A	P_1
C	P_1
T	P_1
G	P_1

Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1
Q_1	Q_1	Q_1	Q_1

A	P_2
C	P_2
T	P_2
G	P_2

Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2
Q_2	Q_2	Q_2	Q_2

A	P_3
C	P_3
T	P_3
G	P_3

Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3
Q_3	Q_3	Q_3	Q_3

3X SUBSTITUTION MODELS

Both partition and mixture models allow more than one substitution model along the sequences and accommodate heterogeneity in evolutionary processes. **However**:

Partition Models:

- **Sites assignment**: each site is assigned to a specific partition based on predefined *criteria*.
- **Model application**: distinct substitution models are applied to each partition.
- **Assumption**: the assignment of sites to partitions is known and fixed prior to analysis.

Mixture Models:

- **Sites assignment:** no assignment of sites to a class, probability of belonging to multiple classes.
- **Model application**: a weighted average of substitution models is used.
- **Assumption**: the site-to-class assignment is unknown.

FINISH