

# Supplementary materials

Timothy Hughes<sup>a</sup> David A. Liberles<sup>b,\*</sup>

<sup>a</sup>*Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

<sup>b</sup>*Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA*

---

## Abstract

This file contains information on the supplementary materials for the article entitled "The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families". This includes: animation files, R code and figures of gene family size distribution for all models.

---

## 1 Simulation animations

Unzip the file `animations.zip`. Open the `animations.html` file in a browser. This file provides a summary of all available animations and from this file it is possible to link the details of each animation.

## 2 Simulation code

The code is written in Java and is available in the `simulationCode.zip` file. This file can be unzipped and its contents can be inspected to gain an understanding of the details of the simulations' code. A good starting point for such an inspection is the "simulations" directory which is structured in the same way as table 1. The code is extensively commented.

To actually run one of the simulations, a user will need to:

1. modify the file paths in the relevant simulation `.java` file.
2. ensure that all classes of the `.jar` file are on the classpath.

---

\* Corresponding author.

*Email address:* `liberles@uwyo.edu` (David A. Liberles).

3. ensure the maths package colt.jar file is also on the classpath (available from <http://dsd.lbl.gov/~hoschek/colt>).

Note, however, that this code was not designed as a software application and, therefore, running modified code might not be straight forward.

### **3 Figures for all models**

Figures of the distribution of gene family size for all models (see table 1 in the article for an overview over all models).

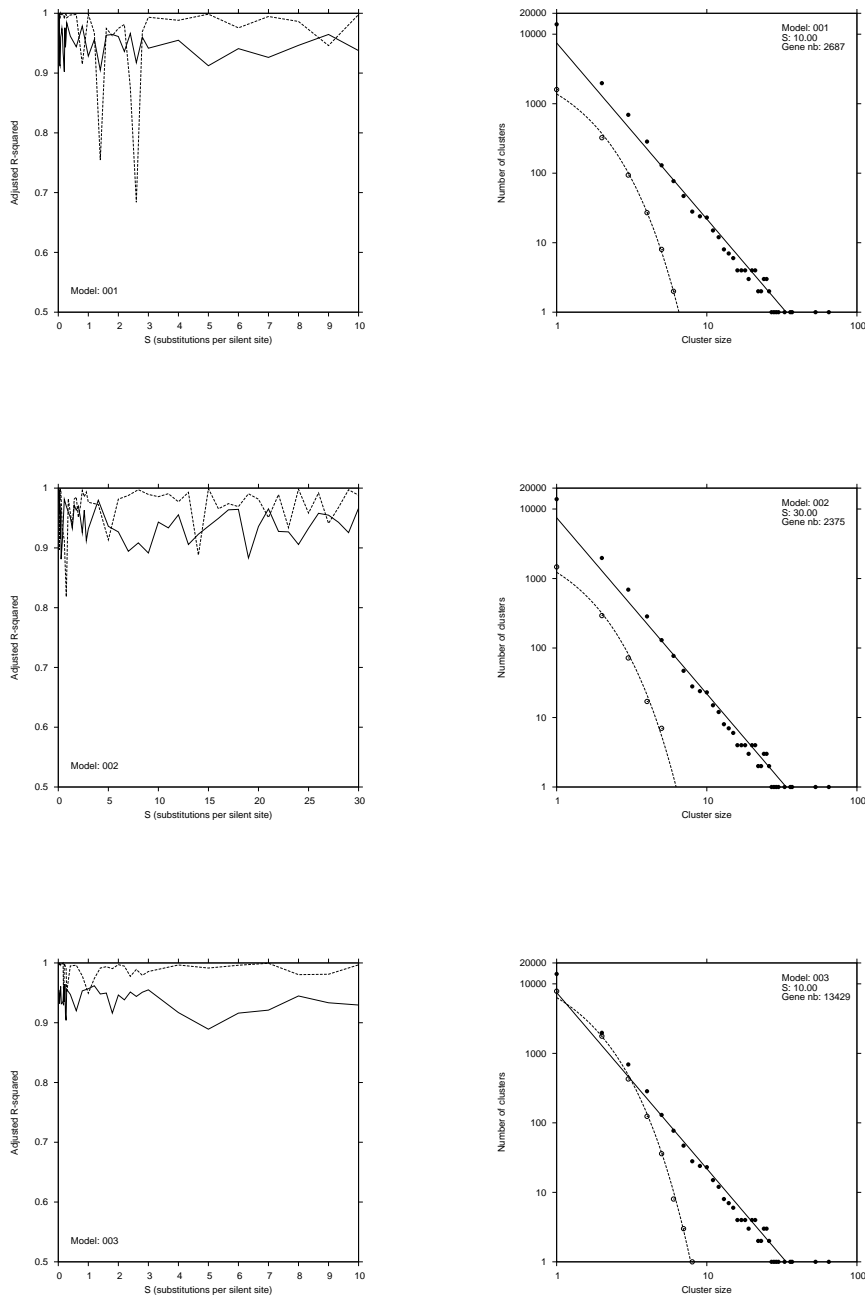


Fig. 1. Group 1 models (same hazard function for all genes)  
*Model 1*: basic; *model 2*: longer S; *model 3*: higher number of initial genes.  
*Solid line*: power-law function. *Dotted line*: exponential function.  
*Black points*: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.  
*White points*: gene family size data computed from the simulation.

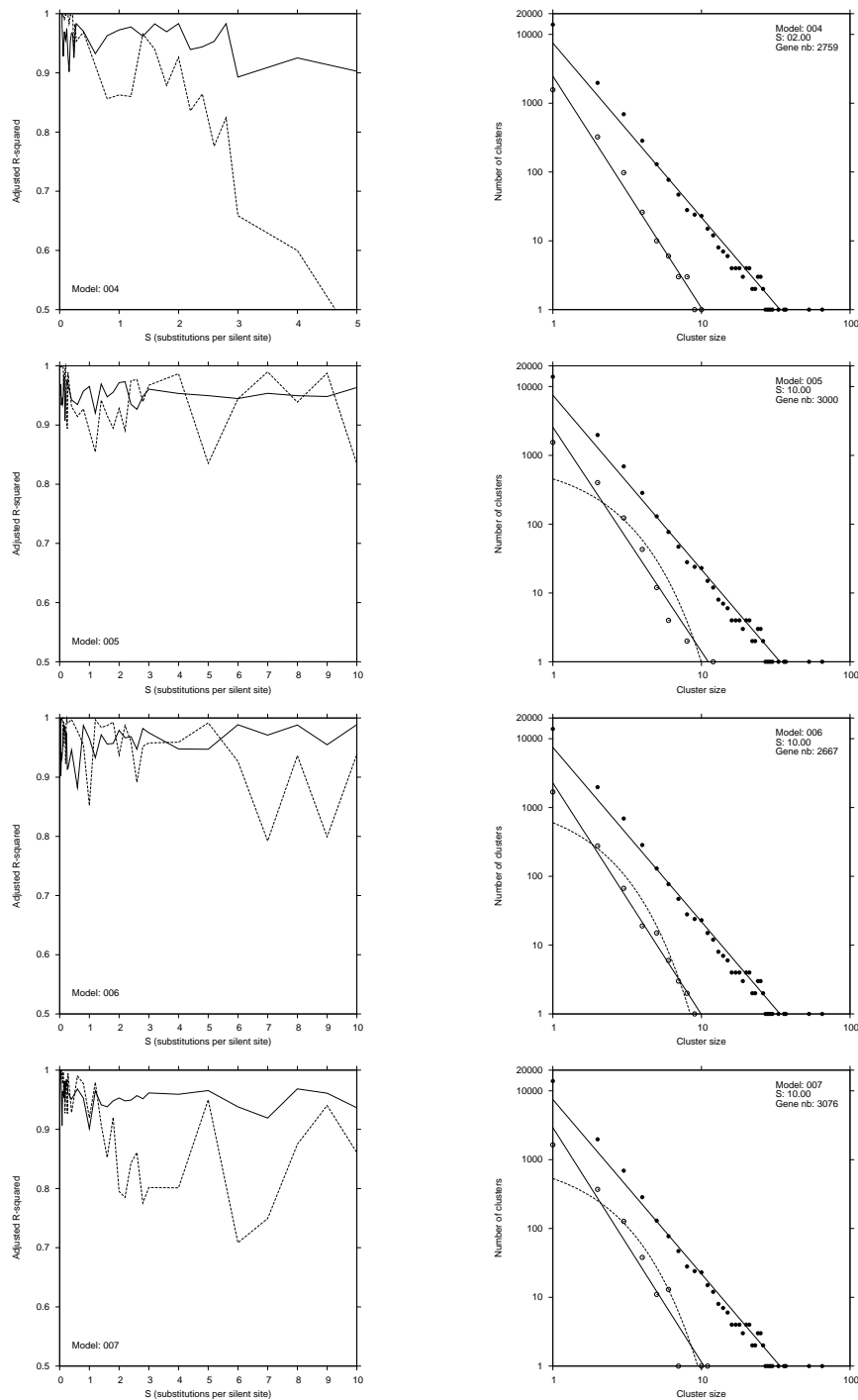


Fig. 2. Group 2 models (different hazard functions for different genes)  
*Model 4*: basic; *model 5*: error not inherited; *model 6*: error mean greater than 0; *model 7*: low error standard deviation.  
*Solid line*: power-law function. *Dotted line*: exponential function.  
*Black points*: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.  
*White points*: gene family size data computed from the simulation.

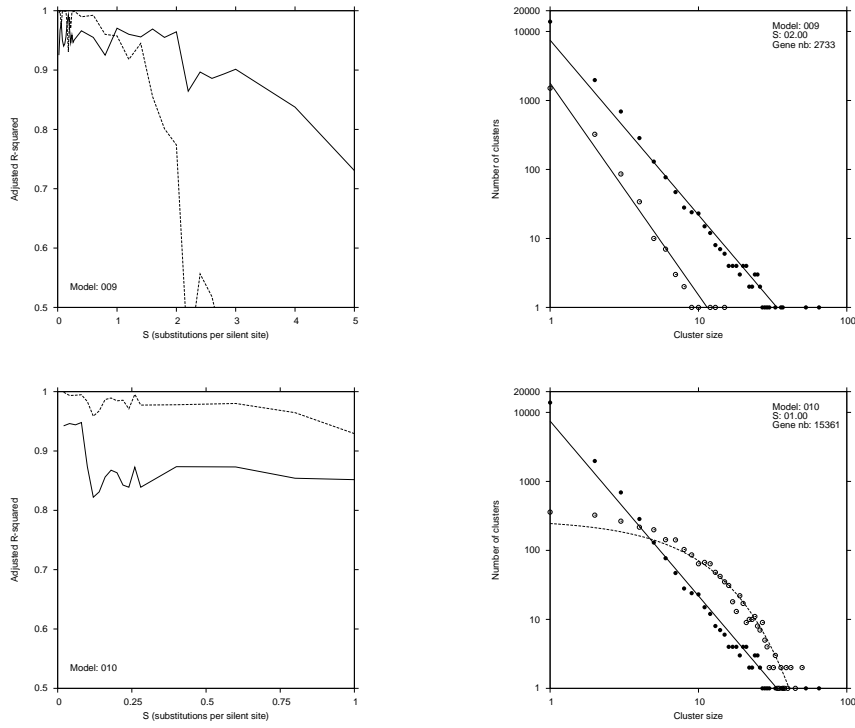


Fig. 3. Group 3 models (one process removed)

*Model 9*: no replacement substitution; *model 10*: no pseudogenisation

*Solid line*: power-law function. *Dotted line*: exponential function.

*Black points*: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.

*White points*: gene family size data computed from the simulation.