Reconstructing Gene Gains and Losses with BadiRate

Pablo Librado, Julio Rozas

Reconstructing Gene Gains and Losses with BadiRate

Pablo Librado[1] and Julio Rozas[2]

1 Centre for Anthropobiology & Genomics of Toulouse, Université Paul Sabatier, Toulouse, France

2 Departament de Genètica, Microbiologia I Estadística, and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Spain

Corresponding Authors

Pablo Librado - plibradosanz@gmail.com

Julio Rozas - jrozas@ub.edu

Running Head: Estimating Gene Gain and Loss Rates with BadiRate

Abstract

Estimating gene gain and losses is paramount to understand the molecular mechanisms underlying adaptive evolution. Despite the advent of high-throughput sequencing, such analyses have been so far hampered by the poor contiguity of genome assemblies. The increasing affordability of long-read sequencing technologies will however revolutionize our capacity to identify gene gains and losses at an unprecedented resolution, even in non-model organisms. To thoroughly exploit all such multigene family variation, the software BadiRate implements a collection of birth-and-death stochastic models, aiming at estimating by maximum likelihood the gene turnover rates along the internal and external branches of a given phylogenetic species tree. Its statistical framework also provides versatility for inferring the gene family content at the internal phylogenetic nodes (and to estimate the minimum number of gene gains and losses in each branch), for statistically contrasting competing hypotheses (e.g. accelerations of the gene turnover rates at pre-defined clades), and for pinpointing gene family expansions or contractions likely driven by natural selection. In this chapter we review the theoretical models implemented in BadiRate, and illustrate their applicability by analysing a hypothetical data set of 14 microbial species.

Key Words

Gene gains; Gene losses; Gene turnover rates; Bioinformatics; Gene duplication; Gene family; Birth and Death model

1. Introduction

Advances in high-throughput sequencing (HTS) are generating massive genome sequence data, fostering comparative, population and metapopulation genomic studies even in non-model and uncultured organisms [1–3]. These studies can provide valuable insights into the evolutionary forces ultimately shaping biodiversity, in both prokaryote and eukaryote organisms, which bears multifarious applications in ecology, animal and plant breeding, conservation genetics, biomedicine, forensics, and sustainable pest control, among others [4–6].

Such biodiversity-based studies, and their potential applications, have been however limited by a number of factors, including: i) DNA sequencing errors, intrinsic to HTS technologies and their underlying chemistry, ii) the quality of genome assemblies, highly dependent on the read length, a critical issue in metagenomics; iii) and the accuracy of genome annotations, both at the structural and functional level. All these limitations deeply compromise the inference of gene gain and loss events, thereby precluding comprehensive analyses of gene family evolution, in both eukaryote and prokaryote organisms. For instance, current data complicate the study of medium-sized gene families (e.g., from 10 to ∼100 copies), where highly divergent and newly originated copies coexist, forming tandem gene arrays resulting from unequal crossing-over, during meiosis in eukaryotes or via homologous recombination in merodiploid prokaryotes [7–11]. Detailed characterization of large-sized gene families encompassing hundreds of thousands of members, which are quite common in many eukaryotic genomes, is even more problematic, and most likely unfeasible using current technologies.

Such incapacity to faithfully characterize repetitive gene copies has likely undermined the actual contribution of key evolutionary mechanisms to multigene family evolution, most notably including the exchange of DNA tracts between tandemly arranged and repetitive family members [12–15]. With costs rapidly dropping and performance gradually increasing, long-read sequencing technologies bear promising potential to overcome all these drawbacks. If more prevalent than anticipated from poor genome assemblies, gene conversion would entail deep implications; *i.e*., greater similarity between gene copies would not necessarily reflect closer phylogenetic origins, but subsequent DNA exchange only. Methods based on gene-tree vs. species-tree reconciliation, therefore, could simply be inappropriate to study gene families recently expanded, and that still undergo active gene conversion [16].

Under this scenario, methods that dispense with explicit gene trees might represent alternatives more robust to gene conversion. One possibility pertains to count-based methods, exploiting multigene family sizes at extant taxa as the major source of input information [17–19]. Such variation in gene count data across species can be inferred through orthoMCL [20], orthoFinder [21], OMA [22] or any equivalent approximation, and contains sufficient evolutionary information to characterize gene turnover processes in a tractable, explicit likelihood framework. Among other capabilities, count-based methods enable estimating the number of gains and losses in particular lineages, and the underlying gene gain and loss rates. In this chapter, we review the stochastic models implemented in BadiRate [18], and showcase their versatility through a hypothetical example closely mirroring the evolution of 14 microbial genomes.

2. Methods

2.1 Stochastic turnover models

BadiRate models two types of gene turnover processes, through density-dependent and density-independent rates (Fig. 1). The former accounts for molecular phenomena such as tandem gene duplication by unequal crossing-over, and assumes that the probability of experiencing a duplication increases proportionally with the number of existing genes (β, birth rate). Similarly, the probability of suffering a gene loss (via deletion or pseudogenization) depends on the total family size (δ, death rate). Gene loss, for example, represents a major mechanism underlying the shrinkage of gene families [23], and is especially prevalent within obligate intracellular bacteria [24]. Beyond gene family expansions and contractions, comparable density-dependent rates ($\beta = \delta$) might well reflect gene content stability, which enables to model the underlying process of gene gain and loss through a single turnover rate (λ, lambda rate).

There are however evolutionary processes deemed as density-independent, by their own idiosyncrasy. The most notable instance involves horizontal gene transfer which does not result from DNA duplication, but represents an external acquisition independent of the actual number of copies [25]. Likewise, a *de novo* gene origin can represent the onset of a new multigene family without requiring a pre-existing family member. Such density-independent processes can be specifically modelled through the gene innovation rate (ι), more generally accounting for all types of orphan genes with no traceable or unknown ancestry. This also includes remote homology relationships that remain undetected by most methods, owing to excessive sequence divergence. Distinguishing gene duplications (density-dependent) and

innovations (density-independent) might be however challenging from gene count data solely. To avoid potential caveats, the gain parameter (γ) enables to group both processes, regardless of their nature, under a single turnover rate. To this end, however, the duplication rate needs to be expressed in the same units as gene innovations, collectively treating all gains as density-independent (i.e. gene gains per mya).

Combined, these four rates constitute the core of the five gene turnover models implemented in BadiRate, namely L (Lambda), LI (Lambda-and-Innovation), BD (Birth-and-Death), GD (Gain-and-Death) and BDI (Birth-Death-and-Innovation). Accommodating the whole complex range of molecular processes underlying gene gains and losses in prokaryotes, the last two turnover models are particularly appropriate for studying multigene families in microbes, as further detailed and exemplified below (Fig. 1).

The most general model, known as BDI, jointly considers two types of gene gains, those identified as gene duplications (via unequal crossing-over; β), and those arising by density-independent mechanisms such as the horizontal gene transfer (ι). The BDI model additionally implements a gene death rate (δ) that accounts for gene deletion or pseudogenization events. The other four gene turnover models simply represent particular instances of this one. Assuming that all family members arose by gene duplication ($\iota = 0$), the BDI simplifies to a BD turnover model. Likewise, equal birth-and-death rates ($\beta = \delta$) leads the BDI and BD to the LI or L models, respectively. Finally, the jointly accounting for all gene increments regardless of their molecular mechanism of acquisition (i.e., gene duplications or not) leads to the GD model. Further details are provided within the supplementary information of the original BadiRate publication [18].

2.2 Branch models

It is widely accepted that DNA (gene) gains and losses represent a major evolutionary mechanism in generating genome diversity and functional innovation [26–28]. This implies that gene family turnover rates might drastically vary across lineages, with lifestyle shifts shaping gene contents, mainly through contractions and expansions of multigene families, but also by *de novo* acquiring new genes. Comprehensively comparing such turnover heterogeneity, therefore, can reveal the molecular hallmark of adaptive variation.

Conditioning on a given phylogenetic species tree, BadiRate can fit different branch models to gene count data. The simplest one assumes that all phylogenetic lineages evolved under the same turnover rates (Global Rates model, or GR). To avoid misinterpretations, it is worth remarking that the GR model does not imply equal numbers of gene gains and losses per branch, but an overall model (with a single rate parameter for gain and loss) whose specific realization (number of gains and losses) can vary due to the stochastic process inherent to evolution. At the other end of the spectrum, the most complex model allows turnover rates to differ at each phylogenetic branch (Free Rates model, FR). Intermediate to those, a full range of flexibility, enabling any potential combination of branch classes (Branch Specific Rates models). Typically, this involves contrasting whether pre-defined groups of species (foreground branches), characterized by a distinctive lifestyle such as parasitic bacteria, evolved under differential turnover rates, in comparison to the remaining phylogenetic lineages (background branches).

2.3 Root family size model

Like many other methods in phylogenetics, BadiRate uses the Felsenstein tree-pruning algorithm to calculate the likelihood of the model given the gene count data. The Felsenstein algorithm traverses the species tree from the phylogenetic tips to the root, following a stepwise procedure. This indeed begins with an observed family size at the phylogenetic tips (external nodes), to then estimate the probability of each potential family size at their immediately ancestral phylogenetic nodes. These probabilities are conditioned by the turnover rates proposed by the current ML optimization step. This process is repeatedly iterated until reaching the phylogenetic root. To complete likelihood calculations, nevertheless, the Felsenstein algorithm also requires a statistical distribution to model the family size at the phylogenetic root. BadiRate contemplates two options for this, including either Poisson (defined by one parameter) or Negative Binomial (two parameters) distributions; in both cases, the underlying parameters are automatically estimated by BadiRate. In practice, both root family size models can accommodate the gene count data comparably well, providing similar results overall, albeit a choice needs to be supplied by the user. This choice can be possibly guided by explicit statistical contrasts of competing root models, as described below.

2.4 Contrasting hypotheses under different models

We refer to any specific combination of gene turnover, branch and root models as BadiRate supermodels. Contrasting the fit of such supermodels to gene count data is paramount to assess the statistical significance of different evolutionary scenarios and/or molecular processes, such as branch-specific accelerations of the BDI rates, whether birth and death rates are statistically different, the most appropriate root model, or the relevance of gene innovation itself.

Several options to identify the best-fit supermodel exist, including Likelihood Ratio Tests (LRT) and the Akaike Information Criterion (AIC). These two methods control for the number of parameters, as more complex models will always fit the data better, yet the aim is to explain evolutionary processes with the simplest possible model. The LRT approach can only be implemented to contrast two models that are nested, in which the simplest one represents a particular instance of that most complex. This can refer to competing branch models such as GR (simplest) vs FR (most complex), and to competing turnover models such as BD (simplest) vs. BDI (most complex). More versatile, the AIC method can simultaneously contrast multiple BadiRate supermodels at once, whether nested or not. The number of parameters can be obtained from the supermodel specified. As an example, a supermodel assuming the same GD turnover model across all branches (GR branch model) and a Poisson distribution for the root (one additional parameter) will be characterized by three parameters (γ, δ and the additional parameter to model the root family size through a Poisson distribution). Replacing the GR by a FR branch model, in a rooted tree summarizing the phylogenetic relationship of 6 species (therefore 2 x 6 - 2 = 10 branches; 6 external and 4 internal), would increase the number of parameters to 21, while additionally considering a Negative Binomial instead of a Poisson distribution for the root would involve 22 parameters. Together with the likelihood calculated by BadiRate, this provides flexibility to formally test as many biological hypotheses as conceived by the user.

3. Running the program

In this chapter we will leverage a hypothetical toy example designed to provide useful guidelines and tips on how-to use BadiRate, particularly noting potential (common)

misinterpretations. This comparative panel comprises a total of 14 species at different evolutionary distances, and closely mirrors the gene content of prokaryotic species (Fig. 2A), which can be appropriately analyzed using the GD model.

3.1 Prepare the software and the data

BadiRate v1.35 is open-source and standalone software, available at http://www.ub.edu/softevol/badirate/. It does not require installation, and runs in any command-line interface under the Perl interpreter. BadiRate requires two sources of input data, namely the species tree, and the gene counts for each family and species. The branch lengths of the species tree should not be expressed in units of substitution rate, since these are shaped by lineage-specific evolutionary forces, such as demographic histories, and are thus relative. The species tree needs to be instead ultrametric, warranting that contemporaneous species evolved during the same amount of time, in absolute scale (e.g., mya), since their most recent common ancestor. This can be attained by using first RAxML v8 (Stamatakis 2014), followed by r8s (Sanderson 2003) (Fig. 2A; Note 1). The latter implements a semi-parametric approach to search for the absolute time scale that best fits the relative substitution rates, and given calibration points. As such calibration, we placed the divergence time of the root at 358 mya. The ultrametric tree (newick format) that we finally used was `DataSet_ultrametric.tree` (supplementary data file):

```
((Sp_Out_L:338.947317,(Sp_Out_K:321.606396,(Sp_Out_J:289.
680865,(Sp_Out_I:270.157513,(Sp_Out_H:254.615370,(Sp_In_A
:56.744225,(Sp_In_B:32.900113,((Sp_In_C:9.143547,Sp_In_D:
9.143547):4.542730,(Sp_In_E:9.521834,(Sp_In_F:7.991694,Sp
_In_G:7.991694):1.530140):4.164443):19.213836):23.844112)
```

```
:197.871145):15.542143):19.523352):31.925531):17.340920):
19.052683,(Sp_Out_M:335.000000,Sp_Out_N:335.000000):23.00
0000);
```

The second input required for BadiRate is the family size file, which is tabulated to summarize the gene count per gene family and species. This can be obtained using OMA [22] or similar algorithms (see note 2), and in our example consists of 3,402 finely-grained groups of homologous genes, also known as gene families, phyletic profiles or orthogroups (supplementary data file 2; Notes 2 and 3). This file was additionally supplemented with the counts of orphan genes for each species, in order to account for all protein-coding genes. In total, this resulted in a file with 14,071 rows excluding the header (`inputHOG.tsv;` supplementary data file), where the first four orthogroups are:

```
Group  Sp_In_F      Sp_In_D      Sp_Out_I     Sp_In_B
       Sp_In_A      Sp_Out_H     Sp_In_E      Sp_Out_M
       Sp_Out_K     Sp_In_G      Sp_Out_N     Sp_Out_L
       Sp_In_C      Sp_Out_J
HOG00004        1       1       0       1       2       0
1       0       0       1       0       0       1       0
HOG00007        1       1       1       1       1       0
1       1       0       1       1       1       1       1
HOG00009        1       1       1       1       0       1
1       1       1       1       0       0       1       1
```

```
HOG00015        2       2       3       2       1       1
1       2       0       2       0       0       2       1
```

It is important to note that the species names must be exactly the same in both the newick and the family size files. To further describe an example, the second gene family, with id HOG00007, contains a single gene copy in all species, except in Sp_Out_H and Sp_Out_K where the gene family HOG00007 is absent.

3.2 Running BadiRate

3.2.1 Inference under the GR model

Once prepared, the family size file (`inputHOG.tsv`) and the ultrametric species tree (`DataSet_ultrametric.tree`) are ready for use as input to BadiRate. As first supermodel, we considered:

- a gain-and-death turnover model (`-rmodel GD`),
- a Global Rates branch model (GR, by default), and
- a Poisson distribution for the root model (`-root_dist 1`)

In total, the four parameters of this supermodel were optimized by maximum likelihood (ML). Abbreviated as GD-GR-ML, the corresponding running command is:

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -out
GR.out -unobs 1 -anc -outlier
```

Using this command (the order of the command line options is irrelevant), the output will be saved into the file supplied after the `-out` flag (`GR.out`), but would be printed into the standard output channel otherwise. This command also has three other flags, `-unobs 1`, `-outlier` and `-anc`. The option `-unobs 1` corrects for the (small) possibility that a gene family existed in the past, but its genes became lost via pseudogenization or deletion in all 14 species. Overlooking these unobserved families, therefore, would lead to an incomplete and partial view of the process, hence, to biased parameter estimations. We thus always recommend activating `-unobs 1`, especially in genome-wide analyses, unless users aim at estimating the turnover rates specifically in a well-defined subset of gene families, and nowhere else.

The last two flags served to conduct further downstream analyses, both conditioned on the three parameter values estimated assuming the GD-GR-ML supermodel. More specifically, the `-anc` option performs a joint reconstruction of the number of genes in the ancestral nodes of the species tree, *given* the estimated GD rates, and separately for each gene family. Once the ancestral gene count is reconstructed, at the gene family level, it is then straightforward to compute the number of gene gains and losses that occurred at each gene family and phylogenetic branch, as the mere subtraction between the corresponding number of genes in the parental and descendant nodes. The total number of gene gain and losses per branch, then, is just the accumulated sum over all gene families. Such a minimum number of gene gains and losses is automatically calculated by BadiRate, provided that the `-anc` option is activated.

The last option, `-outlier`, pinpoints gene families that *unlikely* evolved under the estimated turnover rates; this is, that the probability of generating the minimum number of gene gains and losses in a particular phylogenetic branch and gene family, and under the inferred GD rates, is very low. This option is then extremely useful to identify specific gene families that underwent significant contraction or expansion bursts, potentially uncovering the impact of relaxed selection or adaptive evolution. If the outlier families were however systematically associated with the same phylogenetic branch, this option could solely reflect that the GR branch model is too simple, and that fitting the data requires accounting for turnover rate heterogeneity across branches.

3.2.2 Specifying branch-specifics models

The extant number of genes across the 14 species reveals that Sp_In_A, Sp_In_C and Sp_In_F have unusually high numbers of protein-coding genes, suggesting that these species evolved through massive gene gain processes. To accommodate these potential lineage-specific expansions, we applied two additional branch models, allowing distinctive GD rates for these three species. The first one assumes that there are two different types of GD rates (two branch classes), one for Sp_In_A, Sp_In_C and Sp_In_F, and the other for the rest (supermodel GD-BS1-ML, 5 parameters), as specified by the following command:

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -bmodel
"18->6:10->8:14->12" -out BS1.out -unobs 1 -anc -outlier
```

Note the addition of an extra flag, `-bmodel "18->6:10->8:14->12"`. That flag value indicates to BadiRate that the phylogenetic branches 18->6, 10->8 and 14->12 evolved under their own turnover rates, potentially different from the rest of lineages, and that these three branches have identical GD rates (as indicated by colon separators). These three branch identifiers correspond to the terminal branches leading to Sp_In_A (18->6), Sp_In_C (10->8) and Sp_In_F (14->12). BadiRate offers a simple way to know the identifier of each phylogenetic branch by performing a dry run, a test execution, with the `-print_ids` option:

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -
print_ids
```

The outcome of that execution is:

```
((Sp_Out_L_1:338.947317,(Sp_Out_K_2:321.606396,(Sp_Out_J_
3:289.680865,(Sp_Out_I_4:270.157513,(Sp_Out_H_5:254.61537
0,(Sp_In_A_6:56.744225,(Sp_In_B_7:32.900113,((Sp_In_C_8:9
.143547,Sp_In_D_9:9.143547)10:4.542730,(Sp_In_E_11:9.5218
34,(Sp_In_F_12:7.991694,Sp_In_G_13:7.991694)14:1.530140)1
5:4.164443)16:19.213836)17:23.844112)18:197.871145)19:15.
542143)20:19.523352)21:31.925531)22:17.340920)23:19.05268
3,(Sp_Out_M_24:335.000000,Sp_Out_N_25:335.000000)26:23.00
0000)27;
```

where the external node identifier is given after the species name, both separated by the underbar symbol. The identifiers of internal phylogenetic nodes, in turn, are embedded within the field usually reserved to bootstrap support values, as defined by the newick nomenclature. To obtain the identifier of each branch, then, the user simply needs to concatenate the id of the parental node, followed by a directed arrow and the id of the descendant node (Fig. 2A).

Since Sp_In_A, Sp_In_C and Sp_In_F do not form a monophyletic clade, it is likely that their multigene family expansions resulted from lineage-specific evolutionary pressures. The following model allows these three lineages to have evolved under independent turnover rates (GD-BS2-ML):

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -bmodel
"18->6_10->8_14->12" -out BS2.out -unobs 1 -anc -outlier
```

Branch ids are now separated by underbars, `"18->6_10->8_14->12"`, instead of colon symbols**.** This informs BadiRate that these three lineages represent free, independent branch classes, whereas all the remaining branches belong to another independent, background, class (9 parameters).

BadiRate can accommodate as many branch classes as needed, and each class can include multiple branches. As an example, the following command:

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -bmodel
"18->6:10->8_14->12" -out BS3.out -unobs 1 -anc -outlier
```

would indicate that the lineages leading to Sp_In_A (18->6) and Sp_In_C (10->8) evolved under the same turnover rates, while Sp_In_F (14->12) represented another foreground lineage, both differentiated from the background (the rest) branch class.

3.2.3 Inference under the most complex model, the FR model

Finally, we also evaluated the most complex branch model, which assumes that each phylogenetic branch evolved under its own turnover rates. With 14 species in the analysis, thus 2 x 14 - 2 = 26 phylogenetic branch classes, the total number of parameters of this model amounted to 53. Since separating each branch identifier by an underbar might be tedious, and prone to errors, BadiRate implements a shortcut for specifying the Free Rates (FR) branch model, whereby the specific running command simplifies to:

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -bmodel
FR -out FR.out -unobs 1 -anc -outlier
```

ML optimization is still perfectly feasible despite the much higher dimensionality in the parametric space. Yet, this analysis comes with additional computational costs. In certain situations, users might however aim to perform a more rapid, preliminary, exploration of the turnover process. To this end, BadiRate also implements two faster

estimation procedures, based on parsimony. To switch to these estimation procedures, the flag `-ep` should be supplied, followed either by CSP (Sankoff parsimony; GD-FR-CSP) or CWP (Wagner parsimony; GD-FR-CWP).

```
perl BadiRate.pl -sizefile inputHOG.tsv -treefile
DataSet_ultrametric.tree -rmodel GD -root_dist 1 -bmodel
FR -out FR.CSP.out -unobs 1 -anc -outlier -ep CSP
```

Parsimony-based inference can certainly help explore the turnover rates in only a few minutes at most, even under the most complex scenarios, such as the FR model. This can provide, in a first glimpse, the guidelines to group branches into a number of classes, each with its own distinctive turnover rate. This immediately disregards a large fraction of branch models deemed as incompatible. For example, it might be unnecessary to test for different turnover rates in branches A and B, if both branches show very similar turnover rates under FR parsimonious inference. Contrasting the remaining subset of branch models might be instead prioritized, through the more rigorous and formal framework for hypothesis testing provided by ML. Analysing by ML the four supermodels described in this chapter yielded the following fit to the data:

Model	Ln(lkl)	#parameters	AIC
GD-GR-ML	-89,559.4853	3	179,124.9706
GD-BS1-ML	-80,687.1326	5	161,384.2652
GD-BS2-ML	-78,423.4202	9	156,860.8404

GD-FR-ML	-72,239.7226	53	144,585.4452

As indicated in the table, the GD-FR-ML supermodel substantially improved the likelihood (lkl) (i.e., greater Ln(lkl)), in comparison to simpler branch models. Importantly, this better fit not only owes to its greater number of parameters, as the Akaike Information Criterion (AIC), which would penalize for a potential excess of parameters (AIC = 2 x #parameters - 2 x ln(lkl)), is considerably lower for the FR model than for the other three models. At this point, it is important to remember that the four branch models reported here serve to exemplify the BadiRate usage, and to provide intuitive and detailed insights into its versatility, but do not intend to represent an exhaustive analysis of the multigene family evolution across the 14 hypothetical species.

3.3 Output

This section will describe the output file corresponding to the FR model, the model that best fits the data. As all BadiRate output files, the `FR.out` file is organized in two different sections. The first is delimited by the `INPUT` and `END INPUT` tokens, and simply summarizes the input parameters, once parsed by BadiRate, helping users to validate that all options are correctly specified, and interpreted by the software. It also includes a subsection associating each phylogenetic branch with its corresponding class, according to the stipulated branch model.

The second section begins with the token `OUTPUT`, and is further subdivided into subsections. Starting by the token `##Family Turnover Rates`, the first such subsection reports the turnover rate estimates per branch class (Fig. 2B), as well as the

likelihood of the GD-FR-ML supermodel. The remaining subsections disclose the results from the optional run modes, if any. For example, by activating the `-anc` option, BadiRate appends a long subsection to the output file, with the reconstruction of the estimated number of copies in the ancestral nodes. This subsection begins with the token `##Ancestral Family Size`, and includes as many rows as multigene families are present in the input family size file. Each row represents a tree in newick format, where the number of genes in the ancestral nodes is given, again, in the field reserved for the bootstrap values. Still within this subsection, the final newick tree outlines the total number of gene copies per node, as summing up the ancestral reconstructions over all families. This is followed by a shorter tabulated subsection, delimited by the `##Minimum number of gains and losses per branch` token, which summarizes the minimum number of gene gains and losses per phylogenetic branch. The `-outlier` option, in turn, activates another subsection that starts with `##Outlier Families per Branch`, which lists all gene families that unlikely evolved under the corresponding turnover rates, identified as previously described (False Discovery Rate with a significance cut-off value of 5%).

3.3.1 Interpretation

In line with the FR representing the best-fit branch model, the estimated GD rates were found to vary highly across phylogenetic branches (Notes 4-6). More specifically, gain rates ranged from $\gamma = 0.0000$ to a maximum of 0.0086 gene gains per mya. Likewise, death rates went from $\delta = 0.0000$ to 0.0157 gene losses (gene deletions or pseudogenizations) per ancestral gene and mya. Note that only the death rate is density-dependent, and thus normalized by the number of ancestral (pre-existing) genes, in contrast to the gain rate.

Considering the evolutionary time spanned by the species' phylogeny (the phylogenetic root is 358 mya), the underlying GD turnover rates predict thus a highly dynamic scenario, involving numerous gene gain and loss events. This high turnover is well illustrated by the joint ancestral reconstruction (Fig. 2C), which portrays an overall increase in the number of genes toward more recent times (Note 7). This is particularly marked along the branches leading to the Sp_In_A, Sp_In_C and Sp_In_F, which nearly doubled their gene repertoire in relation to their immediately ancestral nodes. Investigating whether the corresponding outlier families are enriched for specific functional categories (e.g., through Gene Ontology analyses) would additionally provide biological insights into their adaptive potential, in relation to each species' biology. Multiple procedures to assess for functional enrichment exist, and are already well-described elsewhere, beyond the scope of this chapter.

4. Notes

1) The accuracy of the phylogenetic species tree will depend on both the underlying multiple sequence alignment (MSA) and the phylogenetic reconstruction. As an example, users can leverage M-COFFEE v10.00.r1607 [29] to build the MSA, and then filter for poorly (uncertain) aligned regions, with trimAL [30]. The final MSA was then used to infer the phylogenetic relationships with RAxML v8 [31], and to generate a dated ultrametric tree with the r8s software [32], leveraging the divergence at the root as calibration point. Equivalent approaches for generating ultrametric trees are also feasible.

2) Our hypothetical example was created using OMA [22] to define orthogroups. Other software/algorithms are also commonly used to this end, such as OrthoMCL [20] or OrthoFinder [21]. Based on slightly different approaches, with different levels of sensibility and specificity, these methods can however lead to different orthogroup definitions [33], that could eventually yield somewhat incongruent downstream results, including in BadiRate.

3) The resulting family size file might include count data from protein-coding genes and transposable elements. By their own idiosyncrasy, both have different turnover dynamics. Consequently, BadiRate might fail in finding turnover rates that fit both types of elements at once, reporting the following error message: "*`WARN: Try using a more complex model, or changing the starting values. See the -rmodel, -bmodel and -start_val options`*". This indicates that the count data is excessively heterogeneous to be fit within the

stipulated supermodel, and that more complex alternatives need to be considered. This is however not always enough to fix the problem, and filtering out transposable elements and/or outstandly outlier families is then advisable.

4) The methods implemented in BadiRate are tailored to investigate genome-wide data. Some researchers, however, aim to estimate gene turnover rates for a single or small subset of families only. We caution that reduced gene family data might provide limited information on the gene turnover process, and that BadiRate does not yet provide the variance around the point ML estimates. If dealing with a moderate number of orthogroups, turnover uncertainty could be potentially assessed by bootstrapping or jackknifing them, as commonly done in non-parametric statistics. Alternatively, it could be convenient to perform some complementary analyses based on the parsimony-based methods (e.g., the `-ep CSP` or `-ep CWP` options) [34].

5) Although this will be mitigated in the forthcoming years, the continuity of current genome sequences largely varies across species. In species with highly fragmented assemblies, determining the exact number of gene copies might be challenging, and most likely underestimated as repetitive gene copies remain undetected. These missing gene copies do not represent deletion or pseudogenization events, but mere methodological artefacts that can inflate the death turnover rates, creating spurious heterogeneities in the turnover rates across branches. In such exceptional cases, it might be worth considering alternative and publicly-available approaches that explicitly model uncertainty in the gene count data [19], even though it might be considerably safer to disregard species with extremely poor genome assemblies from these types of analyses. Evaluating gene completeness, or the assembled genome size

relative to the expected one, might constitute recommendable practices prior to further analyses and interpretations.

6) Ancestral polymorphisms represent a well-known problem in phylogenetics [35], also for BadiRate. Briefly, genomic variation in an ancestral species might be unevenly inherited by two or more descending lineages, during their speciation. As differentiating the newly formed species, ancestral polymorphisms - already present in their common ancestor - might be then misinterpreted as evolutionary divergence after speciation, hence overestimating the true evolutionary rate. Expectedly, the impact of ancestral polymorphisms is more perceptible immediately after speciation, when new divergence is still relatively small compared to ancestral variation. With increasing evolutionary divergence, however, the relative contribution of ancestral polymorphisms to variation tends to be neglectable. The same occurs at the multigene family level. Individuals from a hypothetical ancestral species might have carried Copy Number Variation (CNV), and have then evolved as separate species with different numbers of genes. The FR branch model can accommodate (but not explicitly model) such branch variation, but simpler branch models as GR might fail. A prudent sanity-check, prior to any branch model comparison, is thus to quantify the correlation between the inferred turnover rates and the corresponding branch lengths. This can be rapidly done assuming an FR model and under a parsimonious estimation procedure (`-ep CWP` or `-ep CSP`). Theoretically, there should be no such correlation, as turnover rates are already normalized by absolute units of evolutionary time (e.g., γ stands for the number of gene gains per mya). Nevertheless, greater turnover rates in shorter branches could reveal issues with ancestral polymorphisms, either at the sequence level while inferring the ultrametric species tree, or at the CNV

level hinting for large ancestral CNV variation erroneously misattributed as gene turnover after speciation. Specifying separate evolutionary classes for short branches with inflated turnover rates can mitigate the problem, but excluding closely-related species represents a more conservative alternative.

7) The joint ancestral reconstruction performed here revealed an increasing number of genes toward the tips of the species tree, particularly marked in the lineages leading to Sp_In_A, Sp_In_C, and Sp_In_F. Similar trends have been consistently reported by other scholars, regardless of the species studied. The universality of this trend is legitimately suspicious, and might suggest a potential bias in ancestral gene reconstructions. This however should not be attributed to BadiRate, which clearly outperforms competing methods in simulation-based benchmarks. Such bias, in fact, is introduced at earlier stages, during multigene family definition. Understandably, the task of defining gene families suffers from decreasing power with more remote homologies, as protein sequences become highly divergent, unrecognizable. This implies that ancestral but remote genes remain undetected as shared across species, and thus that the number of genes at ancestral nodes appears to be smaller than really is, creating this spurious trend of universally expanding gene families. The extent of this bias remains understudied, but expected to be pronounced if dealing with deeply divergent species, while largely ameliorated by gathering a set of species that provide more even and denser phylogenetic coverage over the time-course of evolution.

Acknowledgments

We thank Vadim A. Pisarenco for help with drawing some figures. This work was supported by the Ministerio de Economía y Competitividad of Spain (PID2019-103947GB).

References

1. Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution 29:51–63. https://doi.org/10.1016/j.tree.2013.09.008

2. Bleidorn C (2016) Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity 14:1–8. https://doi.org/10.1080/14772000.2015.1099575

3. Douglas GM, Langille MGI (2019) Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes. Genome Biology and Evolution 11:2750–2766. https://doi.org/10.1093/gbe/evz184

4. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS (2018) Long reads: their purpose and place. Hum Mol Genet 27:R234–R241. https://doi.org/10.1093/hmg/ddy177

5. Pérez-Losada M, Arenas M, Galán JC, Bracho MA, Hillung J, García-González N, González-Candelas F (2020) High-throughput sequencing (HTS) for the analysis of viral populations. Infection, Genetics and Evolution 80:104208. https://doi.org/10.1016/j.meegid.2020.104208

6. Álvarez-Lugo A, Becerra A (2021) The Role of Gene Duplication in the Divergence of Enzyme Function: A Comparative Approach. Frontiers in Genetics 12:1253. https://doi.org/10.3389/fgene.2021.641817

7. Vieira FG, Sánchez-Gracia A, Rozas J (2007) Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. Genome Biology 8:R235. https://doi.org/10.1186/gb-2007-8-11-r235

8. Librado P, Rozas J (2013) Uncovering the functional constraints underlying the genomic organization of the odorant-binding protein genes. Genome Biol Evol 5:2096–2108. https://doi.org/10.1093/gbe/evt158

9. Han K, Li Z, Peng R, Zhu L, Zhou T, Wang L, Li S, Zhang X, Hu W, Wu Z, Qin N, Li Y (2013) Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Sci Rep 3:2101. https://doi.org/10.1038/srep02101

10. Johnston C, Caymaris S, Zomer A, Bootsma HJ, Prudhomme M, Granadel C, Hermans PWM, Polard P, Martin B, Claverys J-P (2013) Natural Genetic Transformation Generates a Population of Merodiploids in Streptococcus pneumoniae. PLOS Genetics 9:e1003819. https://doi.org/10.1371/journal.pgen.1003819

11. Clifton BD, Jimenez J, Kimura A, Chahine Z, Librado P, Sánchez-Gracia A, Abbassi M, Carranza F, Chan C, Marchetti M, Zhang W, Shi M, Vu C, Yeh S, Fanti L, Xia X-Q, Rozas J, Ranz JM (2020) Understanding the Early Evolutionary Stages of a Tandem Drosophilamelanogaster-Specific Gene

Family: A Structural and Functional Population Study. Molecular Biology and Evolution 37:2584–2600. https://doi.org/10.1093/molbev/msaa109

12. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240

13. Eirín-López JM, Rebordinos L, Rooney AP, Rozas J (2012) The birth-and-death evolution of multigene families revisited. Genome Dyn 7:170–196. https://doi.org/10.1159/000337119

14. Reams AB, Roth JR (2015) Mechanisms of Gene Duplication and Amplification. Cold Spring Harb Perspect Biol 7:a016592. https://doi.org/10.1101/cshperspect.a016592

15. Wang S, Chen Y (2018) Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes. Commun Biol 1:1–11. https://doi.org/10.1038/s42003-018-0014-x

16. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biology 8:R141. https://doi.org/10.1186/gb-2007-8-7-r141

17. Csűös M (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912. https://doi.org/10.1093/bioinformatics/btq315

18. Librado P, Vieira FG, Rozas J (2012) BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281. https://doi.org/10.1093/bioinformatics/btr623

19. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 30:1987–1997. https://doi.org/10.1093/molbev/mst100

20. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503

21. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20:238. https://doi.org/10.1186/s13059-019-1832-y

22. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Vesztrocy AW, Dalquen DA, Müller S, Telford MJ, Glover NM, Dylus D, Dessimoz C (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res 29:1152–1163. https://doi.org/10.1101/gr.243212.118

23. Bolotin E, Hershberg R (2016) Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. Sci Rep 6:35168. https://doi.org/10.1038/srep35168

24. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biology Direct 4:13. https://doi.org/10.1186/1745-6150-4-13

25. Tria FDK, Martin WF (2021) Gene duplications are at least 50 times less frequent than gene transfers in prokaryotic genomes. Genome Biology and Evolution. https://doi.org/10.1093/gbe/evab224

26. Dittmar K, Liberles D (2011) Evolution after Gene Duplication

27. Ohno S (2013) Evolution by Gene Duplication. Springer Science & Business Media

28. Ehrenreich IM (2020) Evolution after genome duplication. Science 368:1424–1425. https://doi.org/10.1126/science.abc1796

29. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34:1692–1699. https://doi.org/10.1093/nar/gkl091

30. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348

31. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

32. Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302. https://doi.org/10.1093/bioinformatics/19.2.301

33. Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train C-M, Bork P, Lecompte O, von Mering C, Xenarios I, Sjölander K, Jensen LJ, Martin MJ, Muffato M, Gabaldón T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430. https://doi.org/10.1038/nmeth.3830

34. Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J (2014) Family size evolution in Drosophila chemosensory gene families: a comparative analysis with a critical appraisal of methods. Genome Biol Evol 6:1669–1682. https://doi.org/10.1093/gbe/evu130

35. Charlesworth D (2010) Don't forget the ancestral polymorphisms. Heredity 105:509–510. https://doi.org/10.1038/hdy.2010.14

Figure Legends

Fig. 1 Schematic representation of three of the birth-and-death models implemented in BadiRate.

Panel A. BDI (birth-death-and innovation) model, including the density-independent innovation (i) rate. In contrast to the β (birth) and δ (death) rates, ι is not normalized by the number of pre-existing genes, but solely expressed as the number of gene gains per unit of time (commonly expressed per million of years). This model is appropriate to deal with biological data that eventually incorporates new copies by HGT, by *de novo* origin or to model genes with untraceable gene homologies. This model, therefore, permits the transition from 0 to 1 gene.

Panel B. The standard BD (birth-and-death) model. This model is described by two density-dependent parameters, β (birth) and δ (death), representing the number of gene gains (for birth) or gene losses (for death) per ancestral number of copies (number of genes in the ancestral node of the phylogenetic tree) and unit of time. Under this model the state transition from 0 to 1 is impossible, as new copies can only be acquired through duplications of pre-existing genes

Panel C. GD (gain-and-death) model. This model has a density-independent gain parameter (γ) and the density-dependent death parameter. This model is thus particularly appropriate to deal with biological data that includes substantial HGT or *de novo* origin events, such as transcription factor binding sites (TFBSs), small non-coding RNAs (miRNAs, piRNAs, etc) and prokaryotic organisms, as exemplified in this chapter.

Fig. 2: A) Ultrametric species tree used for this analysis. The total number of protein-coding genes of the corresponding genome are indicated in parenthesis. The numbers

in the internal and external nodes indicate their IDs. B) Cladogram of the phylogenetic tree showing the gain rates (green) and death rates (red). Rate values are multiplied by 1,000 to improve readability. Estimates in panels B and C were obtained from the best-fit GD-FR-ML model. C) Cladogram of the phylogenetic tree showing the number of gene gains (values in green) and gene losses (red) per phylogenetic branch. The number of genes in ancestral nodes are indicated in yellow boxes.