ANCIENT HGT TO COMMON ANCESTOR OF FIRM-5

Introduction

The presence of bacteria in animal gut is currently believed to be responsible for important functional roles such as pathogen colonization resistance, nutrition complementation and immune system responses (Ellegaard *et al.*, 2016). The adaptation to gut life occurs in a small number of bacterial phyla and is thought to be mainly shaped by horizontal gene transfer (HGT) since this process strongly drives bacterial evolution. Horizontal gene transfer (or lateral gene transfer) defines genetic material shared between organisms which are not in a "parent-offspring" relationship (Soucy *et al.*, 2015). Horizontally transferred genes can either be new, paralogs of existing genes or displaced xenologous genes (Koonin *et al.*, 2001). Clusters of genes can result from HGT and constitute genomic islands (Lu *et al.*, 2016). In the context of gut microbiota, HGT may explain how bacteria evolve adaptation to gut life since persistent acquired genetic material is assumed to provide a selective advantage to the host (Koonin *et al.*, 2001; Soucy *et al*., 2015).

Regarding social insects, the gut microbiota of honey bees (healthy adult workers) is mainly composed of 8-10 distinct species groups of bacteria: *Gilliamella apicola*, *Frischella perrara*, *Snodgrassella alvi*, *Bartonella apis*, "Alpha-2", *Bifidobacterium asteroides*, "Firm-4" and "Firm-5" (Moran *et al.*, 2012, Ellegaard *et al.*, 2016). Some of these bacterial phylotypes have also been retrieved in bumble bees. 16S rRNA community analysis shows that Firm-5 is one of the most abundant species and forms the core of social bees microbiota. Firm-5 comprises *Lactobacillus* species which are Gram-positive facultative anaerobic enabling to metabolize sugars into lactic acids. Because Firm-5 group dominates consistently the gut microbiota of social bees, we will investigate ancient HGT which may be potentially associated with social bees gut life adaptation. We will search for HGT shared by honey and bumble bees and specific to bumble and honey bees, respectively. Moreover, we will look if genes potentially acquired by HGT are located in predicted genomic islands.

Materials and Methods

All codes, commands, input and output files are available on Vital-IT. You can access the folder at:
/scratch/beegfs/monthly/mls_2016/claivaz_ricci/SAGE_Firm5_specific_HGT
Moreover, our GitHub repository is available at:
https://github.com/KamilSJaron/SAGE_Firm5_specific_HGT.git

Our project approach consisted of the following steps:

1. Selection of gene families for 3 groups of bees (Bumble bees, Honey bees, and Bumble-Honey bees)
2. Extraction of protein sequences present in gene families
3. Blastp against RefSeq database
4. Filter of best blast hits
5. Inference of hierarchical taxonomy distance
6. Selection of gene families with potential HGT

7. Inference of phylogenetic trees
8. Prediction of genomic islands

Kirsten M Ellegaard provided us different files including one containing all annotated orthologous and paralogous genes (further called orthologs) of Firm-5 strains which were assigned to gene families (obtained from OrthoMCL; Li *et al.*, 2003). From this file, we selected gene families if at least 80% of all strains for a given group of bees were present and free of other strains including outgroup ones. The Bumble bees group comprised a total of 10 strains, the Honey bees group included a total of 15 strains and the Bumble-Honey bees group resulted in 25 strains (8 outgroup strains). Since we are interested in ancient HGT from distant-related donors, we decided to compare protein sequences. A file containing protein sequences of all annotated genes was provided to us in order to extract the amino acid sequence of every orthologs present in selected gene families for each group of bees. Subsequently, we performed blastp for all protein sequences against RefSeq database (O'Leary *et al.*, 2016). This database was suitable in this project as it is a non-redundant manually curated database containing transcript, protein and genomic sequences. In order to select the best blast hits for each gene family, we set filter parameters. Only blast hits with E-value lower than $1*10^{-5}$ and an alignment length of more than 80% were selected. Moreover, the distribution of percentage of identity of all blast results allowed us to decide picking those with at least 50% of identity (**Supplement S1**). For each gene family, if less than 25 blast hits fitted these parameters, we decreased the threshold of percentage of identity by 20% until reaching 25 hits or more (percentage of identity not smaller than 20%). For every best blast hits, we then looked for their last common ancestor with orthologs on NCBI protein database (NCBI Resource Coordinators, 2016) and determine their subjective hierarchical taxonomy distance:

- Lactobacillus = 1
- Lactobacillaceae = 2
- Lactobacillales = 3
- Bacilli = 4
- Firmicutes = 5
- Bacteria = 6
- None = 7 (Archae, Eukaryota or contaminations)

For each gene family, we plotted the percentage of identity of every best blast hits against their hierarchical taxonomy distance. These plots allowed us to define which gene families are good candidates for HGT. To get insight into HGT candidates, we performed linear and polynomial regression models. If models were significantly different, we defined the polynomial model as the best one and the gene family as HGT candidate. For every manually chosen HGT candidates, we extracted protein sequences of 5 orthologs of studied strains and up to 50 blast hits. Then, we aligned protein sequences using MAFFT (v7.305) and inferred maximum-likelihood phylogenetic trees without bootstrap using RAxML (v8.2.9; Katoh *et al.*, 2013; Stamatakis, 2014). In RAxML, we set amino acid as data type and 'auto' as substitution model. Resulted best phylogenetic trees were visualized using FigTree (v1.4.3; Rambaut *et al.*, 2010). Meanwhile, we predicted genomic islands of reference genomes using IslandViewer (v4): L183 for Honey bees in addition to F237 and F245 for Bumble bees (Bertelli *et al.*, 2017). For HGT candidates (gene families), we extracted protein name of every genes for each reference genome. Thanks to GenBank files provided by K.M. Ellegaard (RAST outputs), we recovered protein coordinates and RAST-predicted function (v2.0). Finally, we mapped manually genes onto each reference genome and investigate if they cluster in a genomic island.

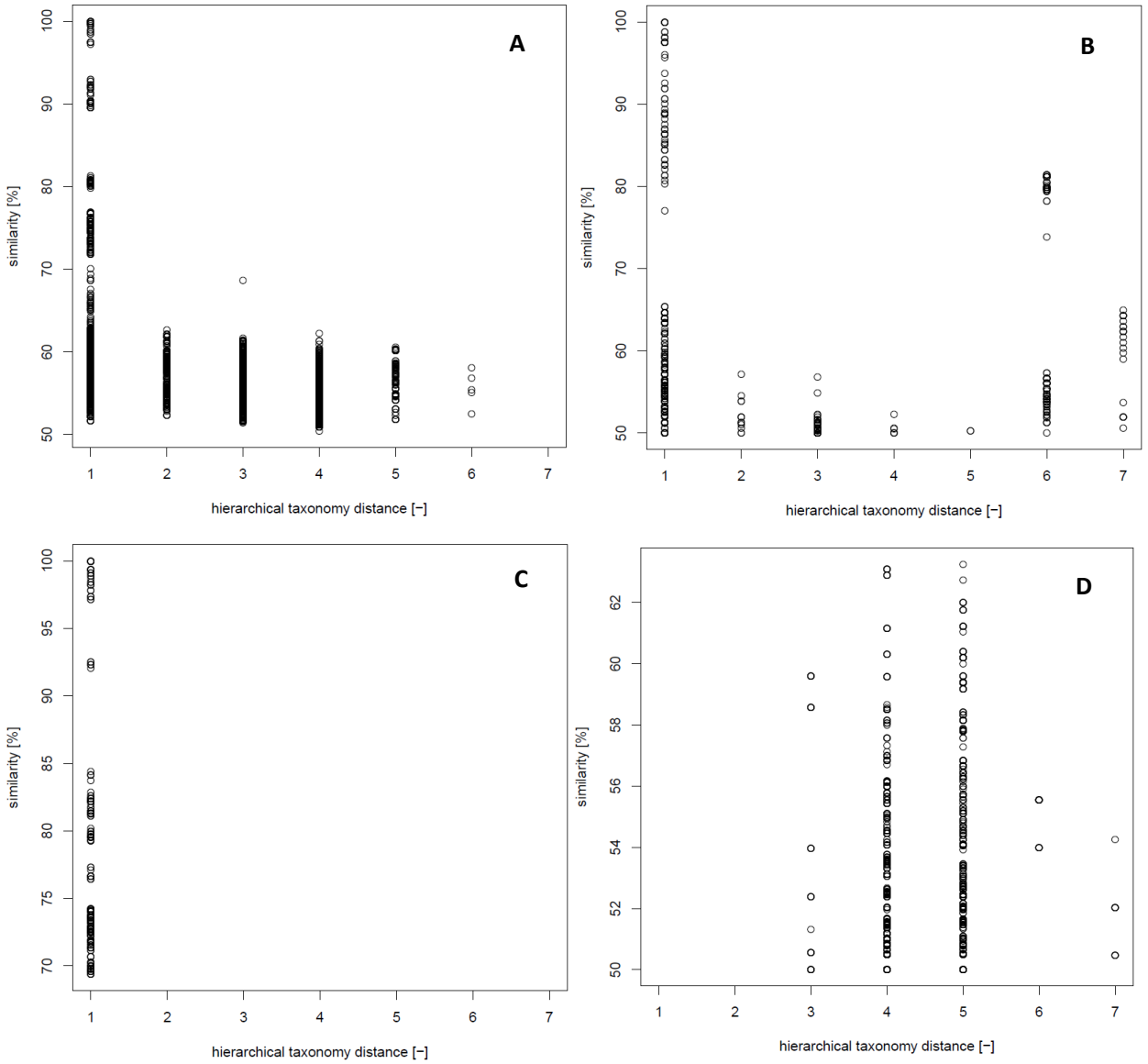

Figure 1: ***Scatter plot of sequences similarity in function of the hierarchical taxonomy distance.*** *A: negative candidate: gene family 1048 specific to bumble and honey bees; B: potential candidate: gene family 1058 specific to bumble and honey bees; C: potential candidate: gene family 1059 specific to bumble and honey bees; D: potential candidate: gene family 1674 specific to bumble bees. The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

Results

Putative ancient horizontal gene transfer

To find some putative ancient HGT specific to Firm-5, the study of orthologs present exclusively in bumble and honey bees (and not in *Lactobacillus* outgroup) can be good HGT candidates. 10, 14 and 27 gene families were specific to bumble, honey and Firm-5 (bumble and honey) clades, respectively. We

then performed blastp against RefSeq database and inferred hierarchical taxonomy distances amongst the best blast hits.

3 different scenarios can be considered as putative HGT (**Fig1**). **Fig1B** shows a higher percentage of similarity for distant-related species in comparison to closely-related ones, which can attest the presence of potential HGT. **Fig1C** shows orthologs specific to Firm-5, which includes best hits only related to *Lactobacillus* hierarchical distance. This means that these genes have only emerged in this bacterial genus or that they have been acquired by HGT from an unstudied donor specific to *Lactobacillus*. This case of putative HGT cannot be assessed by phylogenetic studies and will not more be considered in this project. **Fig1D** shows only hits which are recovered from distant-related taxa and not present in closer ones. In contrast, **Fig1A** shows the negative control of HGT, where the percentage of similarity decreases with the hierarchical taxonomy distance. This means that these orthologs are very ancient and present in the majority of hierarchical levels. This scenario is not considered as putative HGT. All other results are presented in **Supplement S2**. 991, 1058 and 1099 are putative gene families acquired by HGT specific to bumble and honey bees. In addition, 1674, 1675, 1678 and 1757 are putative HGT gene families specific to bumble bees. There is no HGT candidate specific to honey bees. The taxonomical distance 7 is not considered to infer putative HGT, because of possible contaminations. Indeed, most of those results are from whole organism study of *Apis sp.,* which may be contaminated by the microbiota genomes.

Putative HGT phylogeny inference

To confirm the ancient HGT in Firm-5 gene families, phylogenetic analyses are based on amino acid sequences. The tree topology allows the appreciation of the relationship amongst the different hierarchical distance groups. An unordered hierarchical taxonomy tree can attest HGT acquisition for a specific gene family. An expected positive HGT can be *Lactobacillus sp.* clustered with distant-related strains (such as *Firmicutes*, taxonomical distance = 5) in comparison with closer related ones (such as *Lactobacilliales*, taxonomical distance = 2).

Fig2B presents the best candidate of HGT: gene family 1058. Indeed, it seems that *Lactobacillus sp.* extracted from bumble gut (F230 and F233) are more related to *Bifidobacterium sp.* (taxonomical distance = 6) than to other closer hierarchical distance groups and even to reference genomes from honey gut (F259, F260, L186). This result means that a potential HGT occurred between *Lactobacillus sp.* (of bumble) and *Bifidobacterium sp.*, which are both present in the natural microbiota of bees. In contrast, **Fig2A** shows the negative HGT result of the gene family 1048. The tree topology is more structured in function of the hierarchical taxonomy distance, meaning that these orthologs are potentially present in the most recent common ancestor and subsequently diverged in the different taxa. The other phylogenetic trees of HGT gene family candidates are present in **Supplement S3.** As the HGT process is less explicit than for gene family 1058, the different topologies present in **Supplement S3** cannot attest to HGT process. 1099 contains only hierarchical taxonomy distance equals to 1, 2 and 3 and misses distant-related groups. All other gene families miss closely-related groups to safely infer HGT candidates.

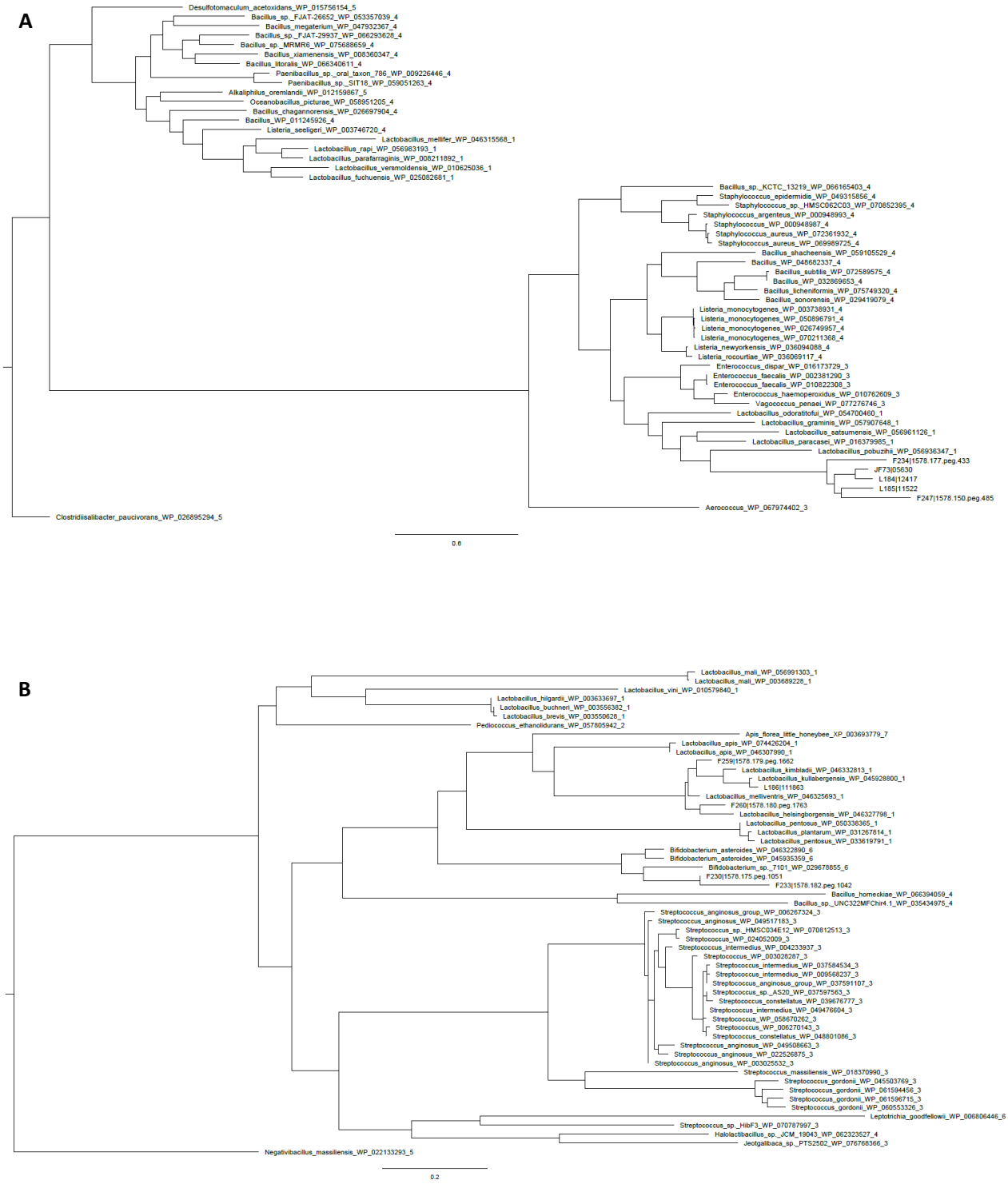

Figure 2: ***Cladogram specific to bumble and honey bees, inferred by LG model.*** *A: 1048 gene family (negative result); B: 1058 gene family. A and B: 50 random BLAST hits and 5 reference genomes were used for those analyses; each random BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance; reference genomes contain only the strain and the protein identifier.*

Map gene family into genomic island

To confirm HGT process for the different gene family candidates, genomic islands are inferred by Sighunt and IslandViewer prediction tools. We hypothesized that different candidates acquired by HGT are clustered in the same genomic island. For this purpose, the gene families are mapped into their reference genomes.

L183 is used as honey reference genome, while two reference genomes (F245 and F237) are used for the bumble group. We chose those reference genomes because of the limitation of both tools which unexpectedly infer the most part of reference genomes as a genomic island. The results, present in **Supplement S4**, show dispersed HGT candidates into reference genomes. They do not cluster together in a genomic island as expected. Even if a genomic island is predicted along the majority of the genome, it does not change the fact that genes are not clustering together in a potential genomic island. To refine this step, the use of different tools inferring more effectively genomic islands need to be used.

RAST function of the putative HGT

Finally, RAST-predicted function specific to the putative HGT gene family are recovered. Concerning the gene family 1058, RAST predicts a ribosomal-protein-alanine N-acetyltransferase involved in the utilization of the acyl-coenzyme A, which is an intermediate compoud between the glycolysis and Krebs' cycle. Other gene families are also predicted as intermediary for energy production: gene families 55 and 1086 code for fumarate reductase flavoprotein subunit (intermediate enzyme of Krebs' cycle) and acetyltransferase from the GNAT family (utilization of coenzyme A for acylation of different substrate), respectively. Gene families 1059 and 1099 are involved in drug resistance by beta-lactamase and activator tipA or regulator MerR transcription, respectively. The predicted functions of gene families 1097 and 1674 are more ambiguous as they are less studied. They code for transcriptional regulators from LysR family and PadR family, respectively. Both are potentially involved in many pathways, such as quorum sensing, virulence, motility and abiotic stress. Concerning the other considered gene families present in **Supplement S4**, RAST cannot predict a precise function; either the predicted functions among reference genomes are different or the function referred to "hypothetical protein".

Discussion

In this project, we investigated ancient HGT potentially associated with social bees gut life adaptation by searching for incongruences between the species tree and our phylogenetic trees based on protein sequences of orthologs. Moreover, we determined if horizontally transferred genes are clustering in a genomic island.

Table 1: ***Summary of the number of gene families taken into account at each step of the project.***

Total number of gene families		4307
Number of gene families per group	Bumble bees	10
	Honey bees	14
	Bumble-Honey bees	27
Number of gene families per group after filtering blast hits	Bumble bees	8
	Honey bees	14
	Bumble-Honey bees	27
Number of gene families potentially acquired by HGT	Bumble bees	4
	Honey bees	0
	Bumble-Honey bees	3
	Firm-5 specific	8
Number of confirmed gene families acquired by HGT	Bumble bees	0
	Honey bees	0
	Bumble-Honey bees	1

At each step of the project, we filtered gene families to find those acquired by HGT (**Table 1**). From a total of 4307 gene families, we ended with only one HGT candidate specific to bumble and honey bees: gene family 1058.

Gene family 1058 is the best putative HGT. Regarding the phylogenetic tree, we observe that there is a clade containing the majority of *Lactobacillus sp*. in addition to *Bifidobacterium sp*. Interestingly, Firm-5 specific to bumble is closely-related to *Bifidobacterium sp*. However, Firm-5 specific to honey is more divergent from Firm-5 specific to bumble. It suggests that the HGT event may happen between *Bifidobacterium sp.* and Firm-5 specific to bumble. It is more likely that HGT occur from Firm-5 specific to bumble to *Bifidobacterium sp*. Indeed, there is no other *Bifidobacterium sp*. in the lineage containing most *Lactobacillus sp*. To confirm, we could infer a phylogenetic tree based on protein (or nucleotidic) sequences of *Bifidobacterium sp*. and all Firm-5 strains. It would give insight into potential co-evolution between those strains, unique event of HGT or HGT occurring several times between them. All other HGT gene family candidates cannot allow us to attest to HGT event.

In this project, there are several limitations to assess reliable phylogenetic tree results. First of all, the use of outgroup to root the tree would be necessary since tree topology depends on the rooting process. We could use the worst hit of blast results as a defined outgroup root. Moreover, increasing the number of considered sequences would result in a better resolution of phylogenetic trees (using bootstrapping process). By randomly selecting sequences, we do not take into account all clades reflecting every taxonomical distances. When there is no blast hits in closer taxonomical distance (scenario **Fig1D**),

phylogenetic trees are not really informative for HGT inference. Concerning blast hits only from *Lactobacillus sp*., it potentially shows recent HGT which need to be further investigated with another strategy (cf. Group 6).

In perspective, we could adapt this pipeline for several aims. For example, we could change the sorting process to infer Firm-5 specific gene content, bumble and honey bees Firm-5 specific gene content strains. Moreover, we could reproduce our analysis by using nucleotidic sequences to infer recent HGT within Firm-5 group. We could then investigate how selective pressure drives horizontally transferred genes evolution.

Currently, genomic island inference tools do not allows us to investigate properly if HGT candidates cluster in a genomic island. In this project, we identified one best HGT candidate. In the case of several HGT candidates, it would be very important to find effective tools predicting genomic islands.

The predicted function of the gene family 1058 is a ribosomal-protein-alanine N-acetyltransferase. In Eukarya, protein acetylation is well known as a post-translational modification involved in cell signaling and as response of external or internal perturbations (Drazic *et al.*, 2016). In procarya, post-translational modification functions remain unclear (Nesterchuk *et al.*, 2011). It would be interesting to study deeply the protein function, it might play a role in the communication between the bee host and Firm-5 microbiota.

References

Moran NA, Hansen AK, Powell JE & Sabree ZL. 2012 Distinctive Gut Microbiota of Honey Bees Assessed Using Deep Sampling from Individual Worker Bees. *PLoS One*. **7**:e36393 (doi:10.1371/journal.pone.0036393)

Ellegaard KM & Engel P. 2016 Beyond 16S rRNA Community Profiling: Intra-Species Diversity in the Gut Microbiota. Frontiers in Microbiology. **7**:1475 (doi: 10.3389/fmicb.2016.01475)

Soucy SM, Huang J & Gogarten JP. 2015 Horizontal gene transfer: building the web of life. *Nature Review Genetics*. **16**:472-482. (doi:10.1038/nrg3962)

Lu B & Leong HW. 2016 Computational methods for predicting genomic islands in microbial genomes. *Computational and structural biotechnology journal*. **14**:200-206 (doi:10.1016/j.csbj.2016.05.001)

Koonin EV, Makarova KS & Aravind L. 2001 Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*. **55**:709-42

Li L, Stoeckert CJ & Roos DS. 2003 OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*. **3**:2178-2189 (doi:10.1101/gr.1224503)

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W,

Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD & Pruitt KD. 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. **4;44**:D733-45 (doi:10.1093/nar/gkv1189)

NCBI Resource Coordinators. 2016 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. **44**(Database issue):D7-D19. doi:10.1093/nar/gkv1290.

Katoh K & Standley DM. 2013 MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. **30**: 772-780 (doi:10.1093/molbev/mst010)

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

Rambaut A & Drummond A. 2010 FigTree v1. 3.1. Institute of Evolutionary Biology, University of Edinburgh

Bertelli C, Laird MR, Williams KP, Fraser S University Research Computing Group, Lau BY, Hoadl G, Winsor GL & Brinkman FSL. 2017 IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*. (doi:10.1093/nar/gkx343)

Drazic A, Myklebust LM, Ree R & Arnesen T. 2016 The world of protein acetylation. *Biochimica et Biophysica Acta - Proteins and Proteomics.* **1864:** 1372-1401 (doi:10.1016/j.bbapap.2016.06.007)

Nesterchuk MV, Sergiev PV & Donstova OA. 2011 Posttranslational Modifications of Ribosomal Proteins in Escherichia coli. *Acta Naturae.* **3:** 22-33

Supplement Materials

S1: distribution of percentage of identity of every blast hits

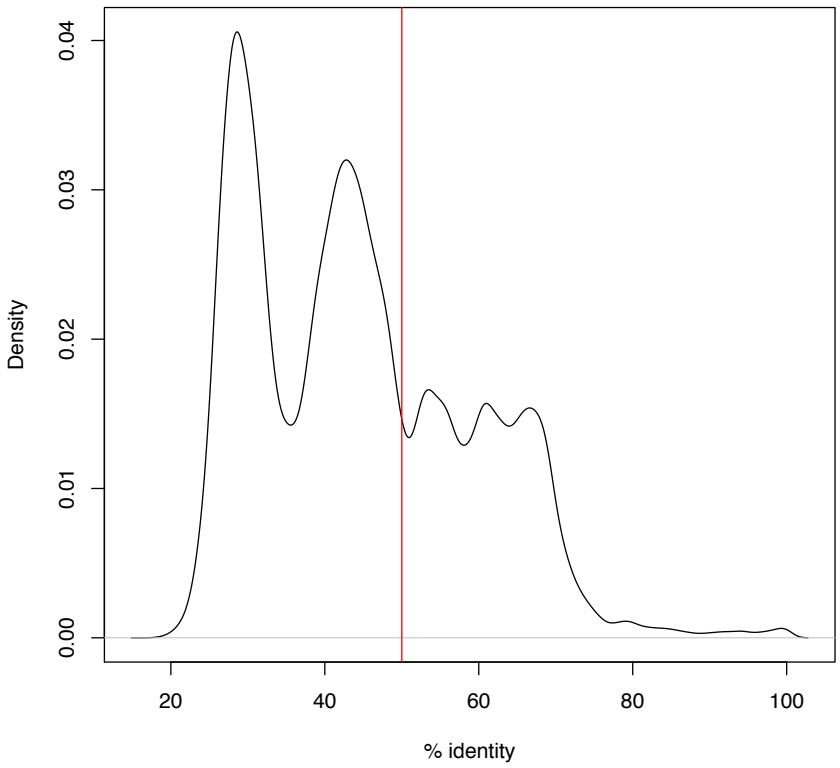

Figure 3: ***Density plot of the percentage of identity of every blast hits.*** *Vertical red line represents the chosen threshold (50%).*

S1: putative ancient HGT

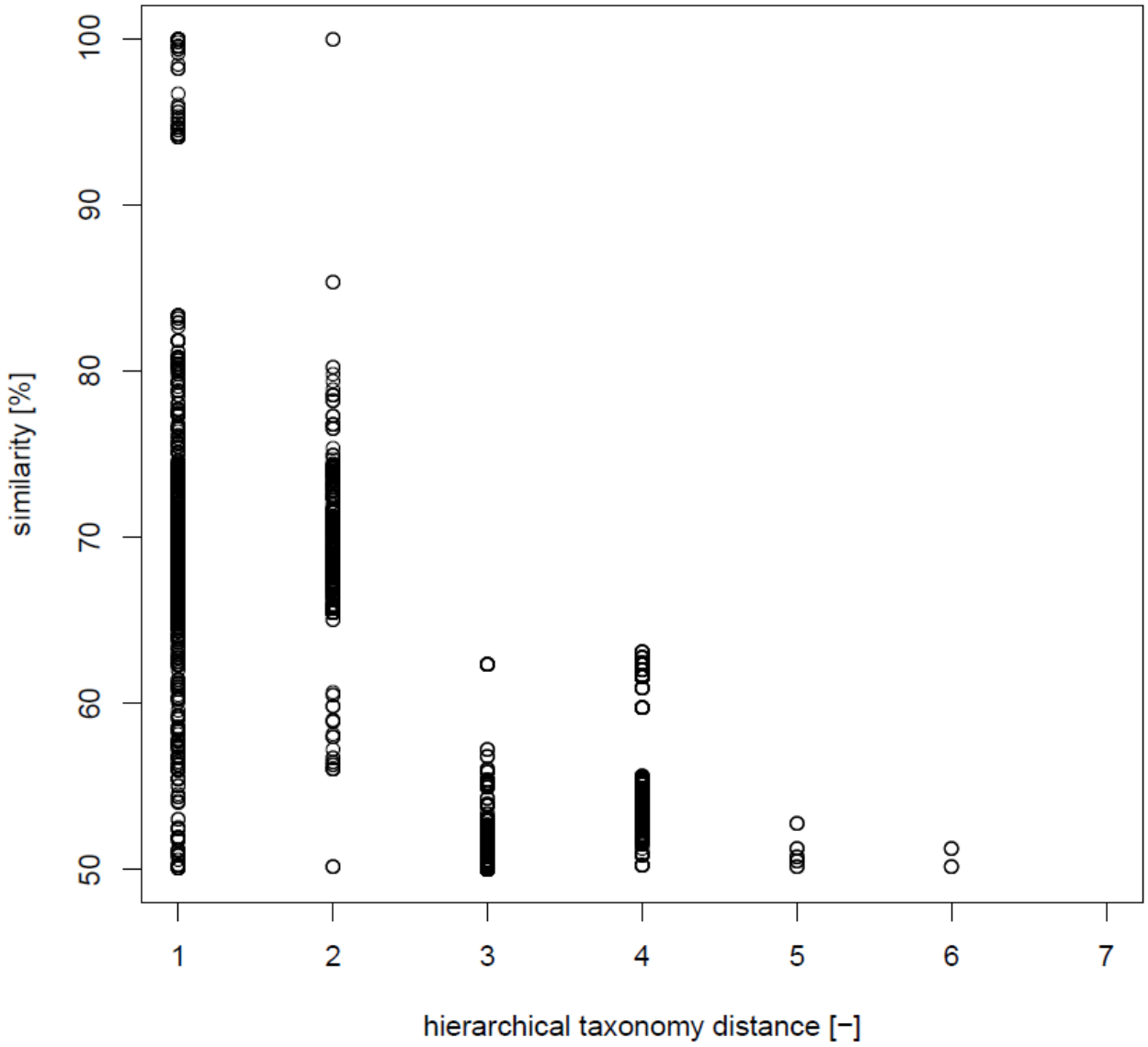

Figure 4: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1060, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

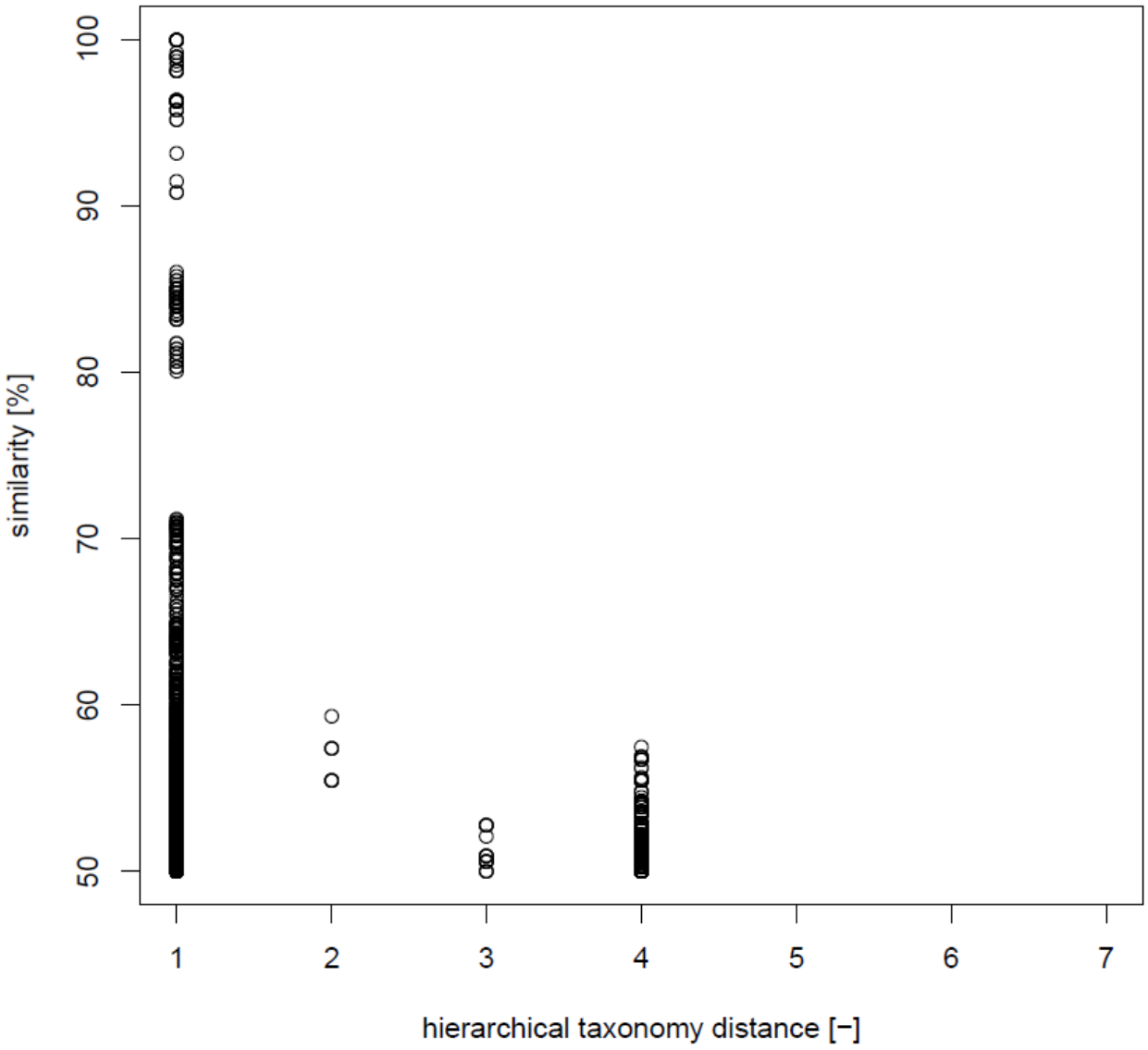

Figure 5: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1061, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

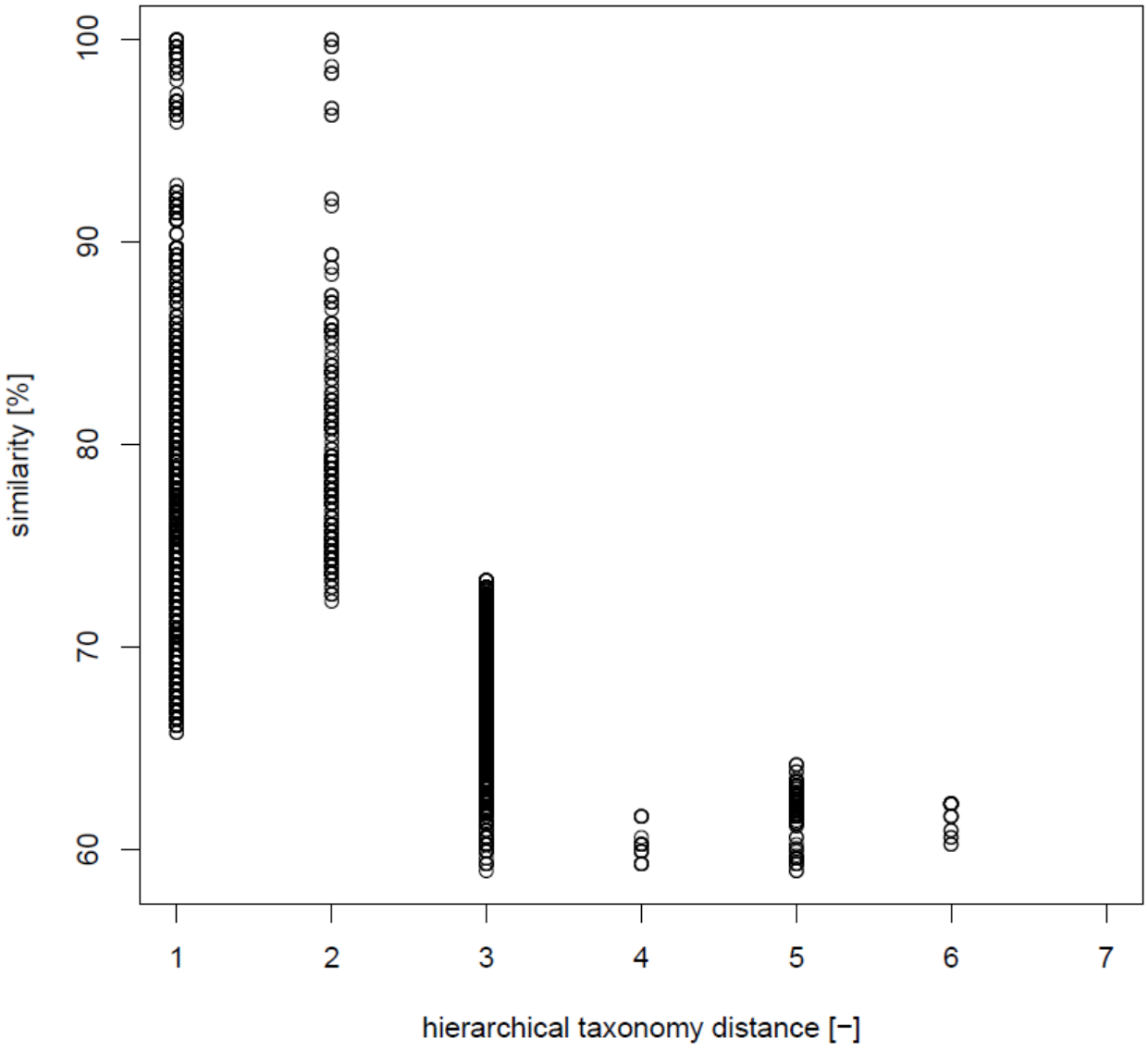

Figure 6: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1062, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

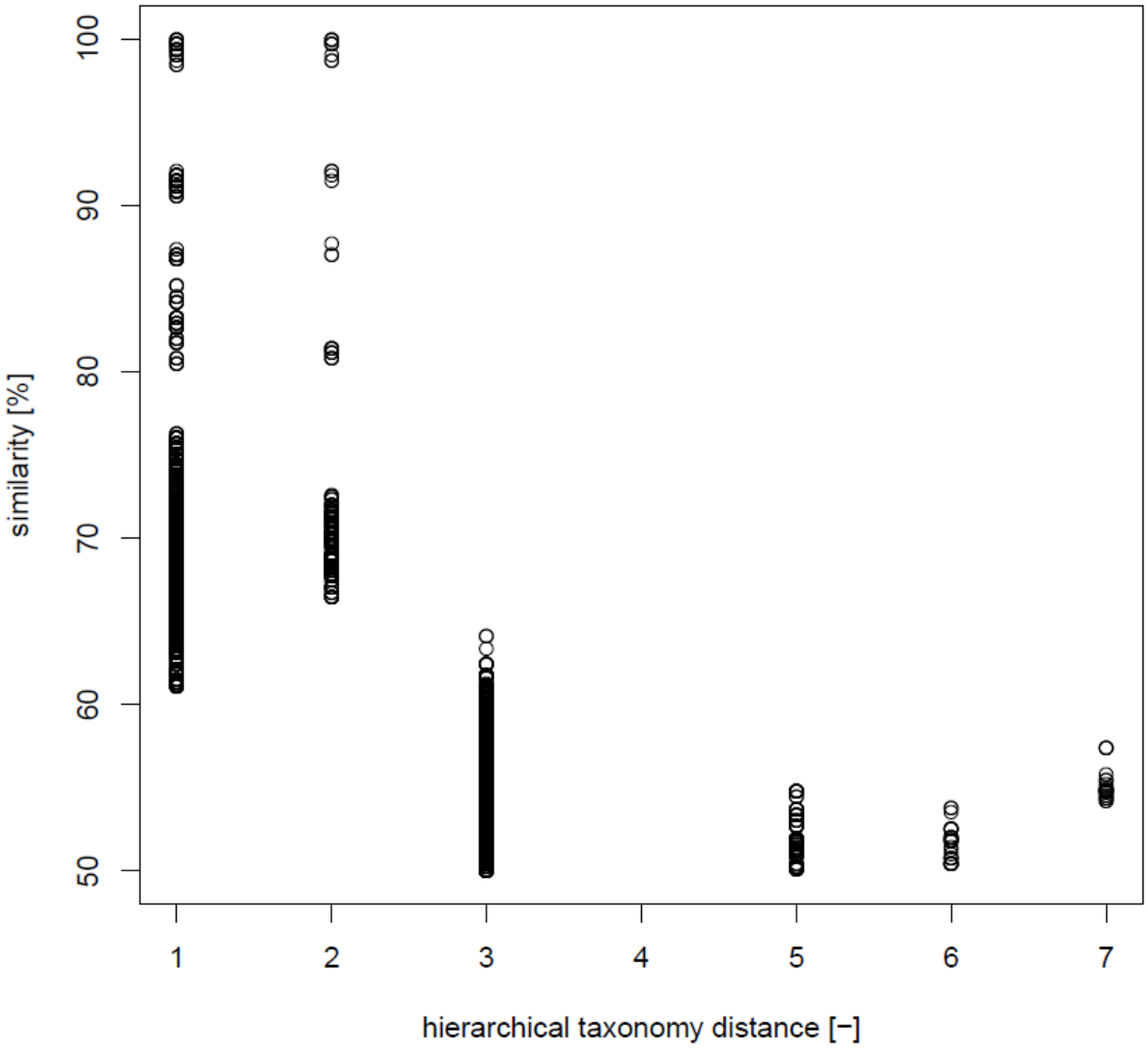

Figure 7: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1063, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

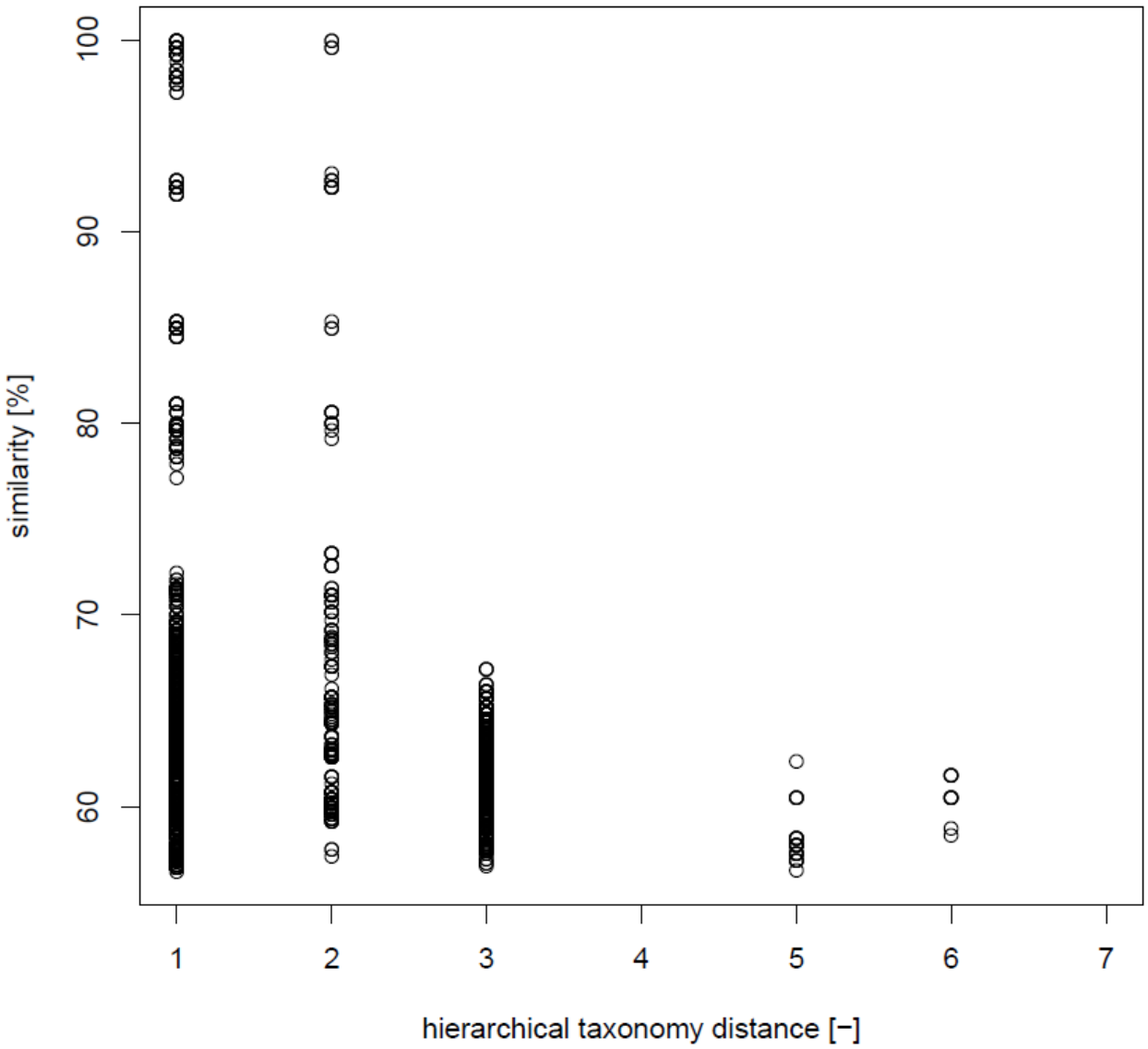

Figure 8: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1064, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

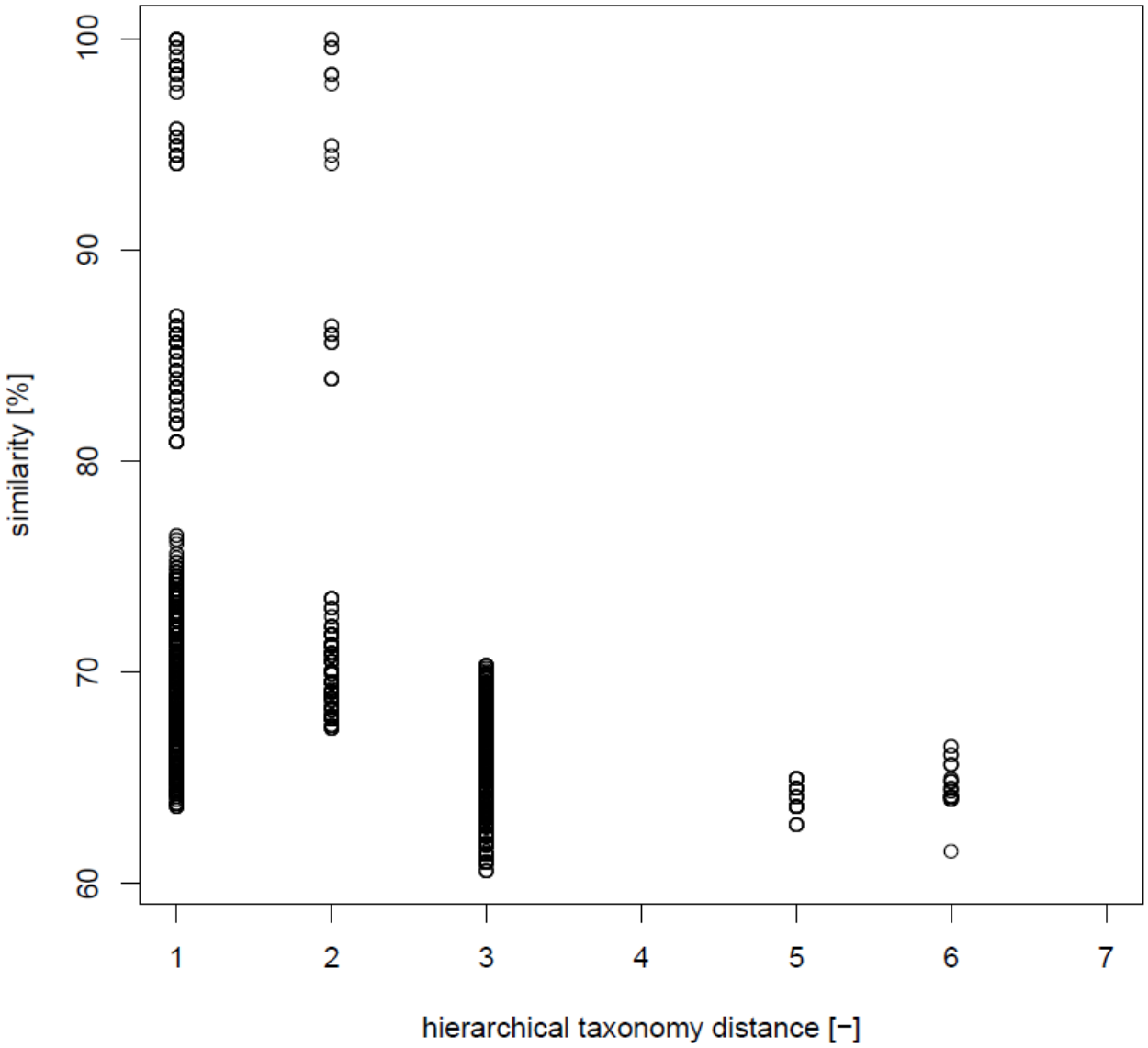

Figure 9: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1066, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

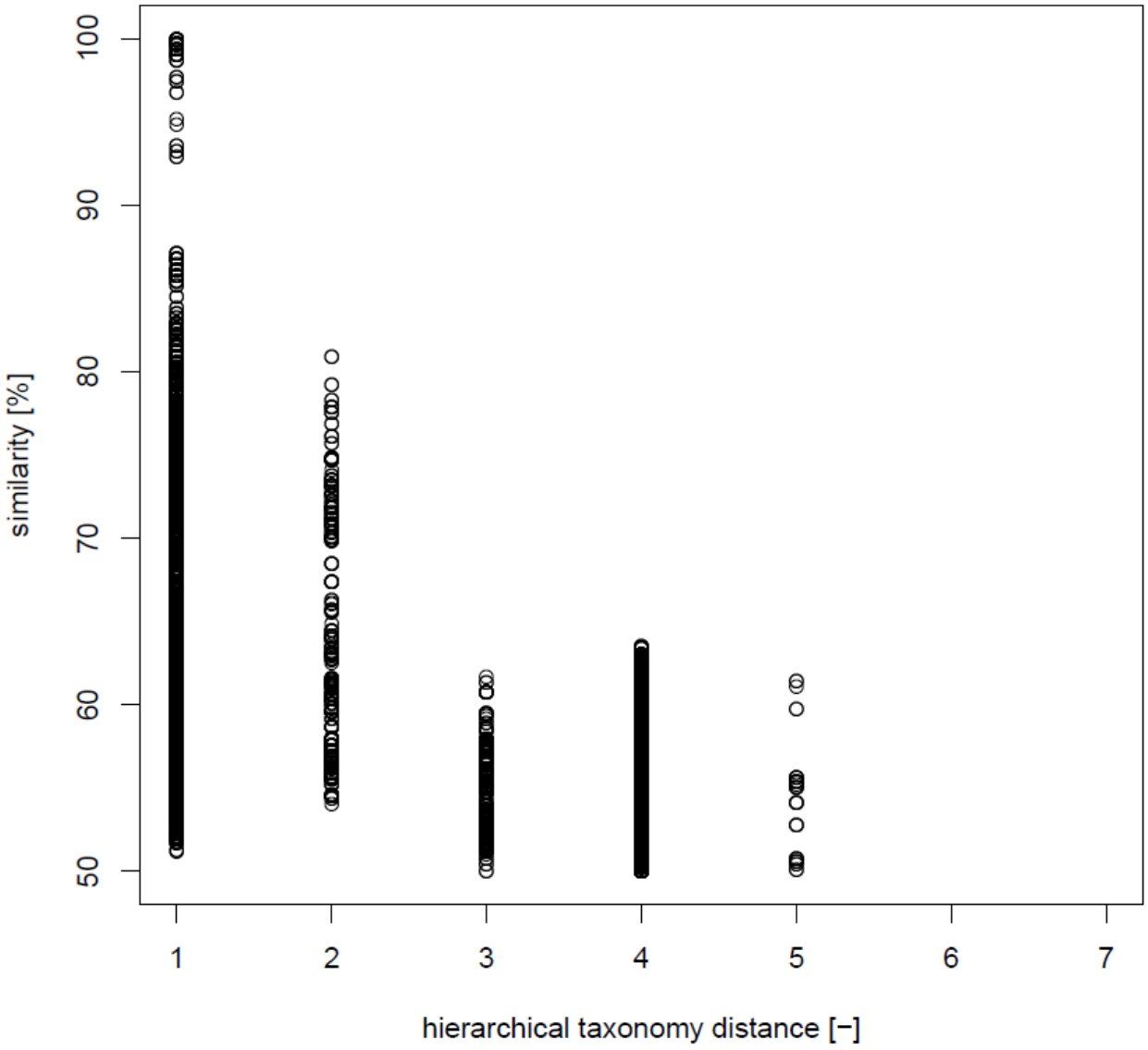

Figure 10: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1067, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

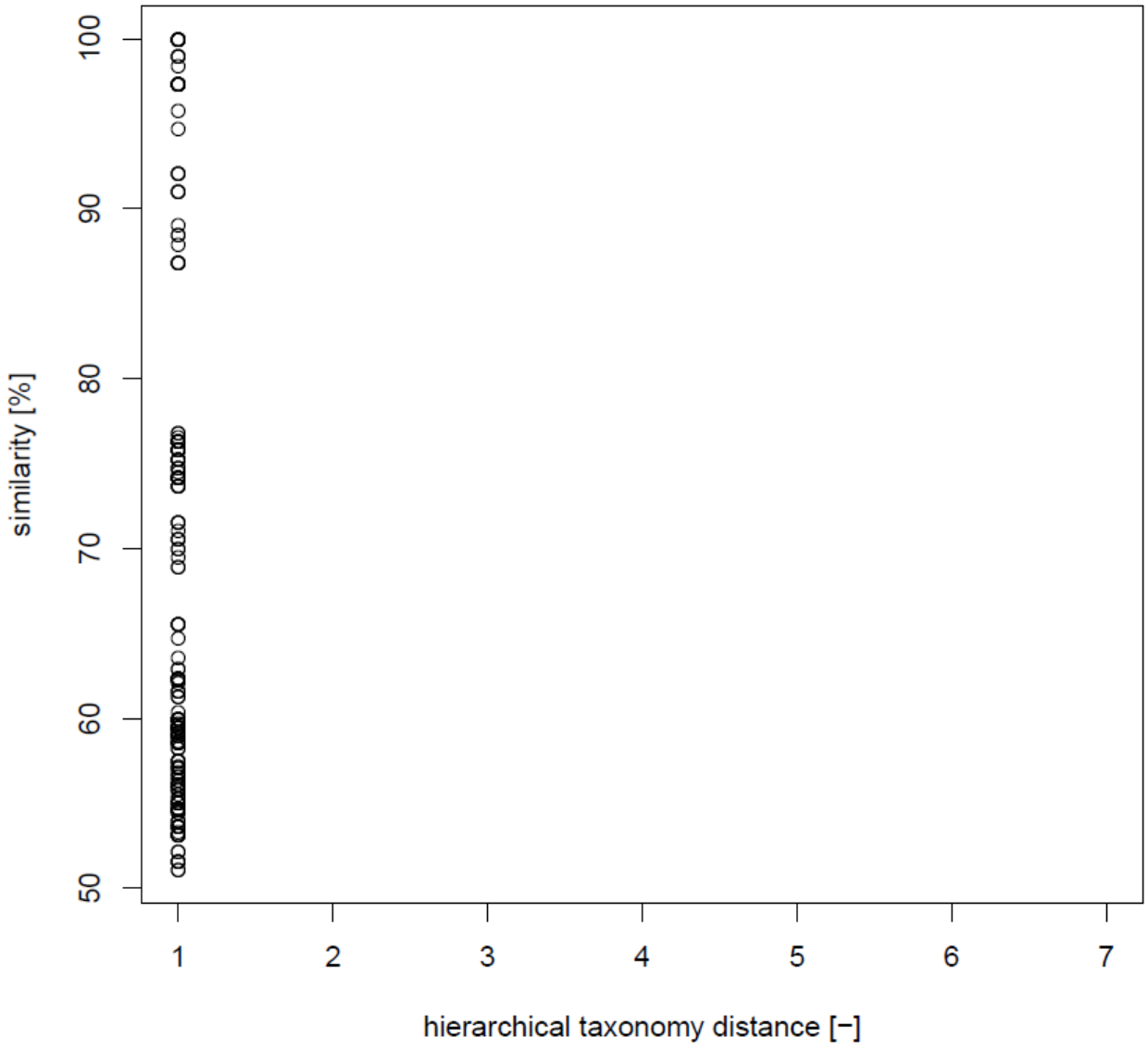

Figure 11: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1072, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

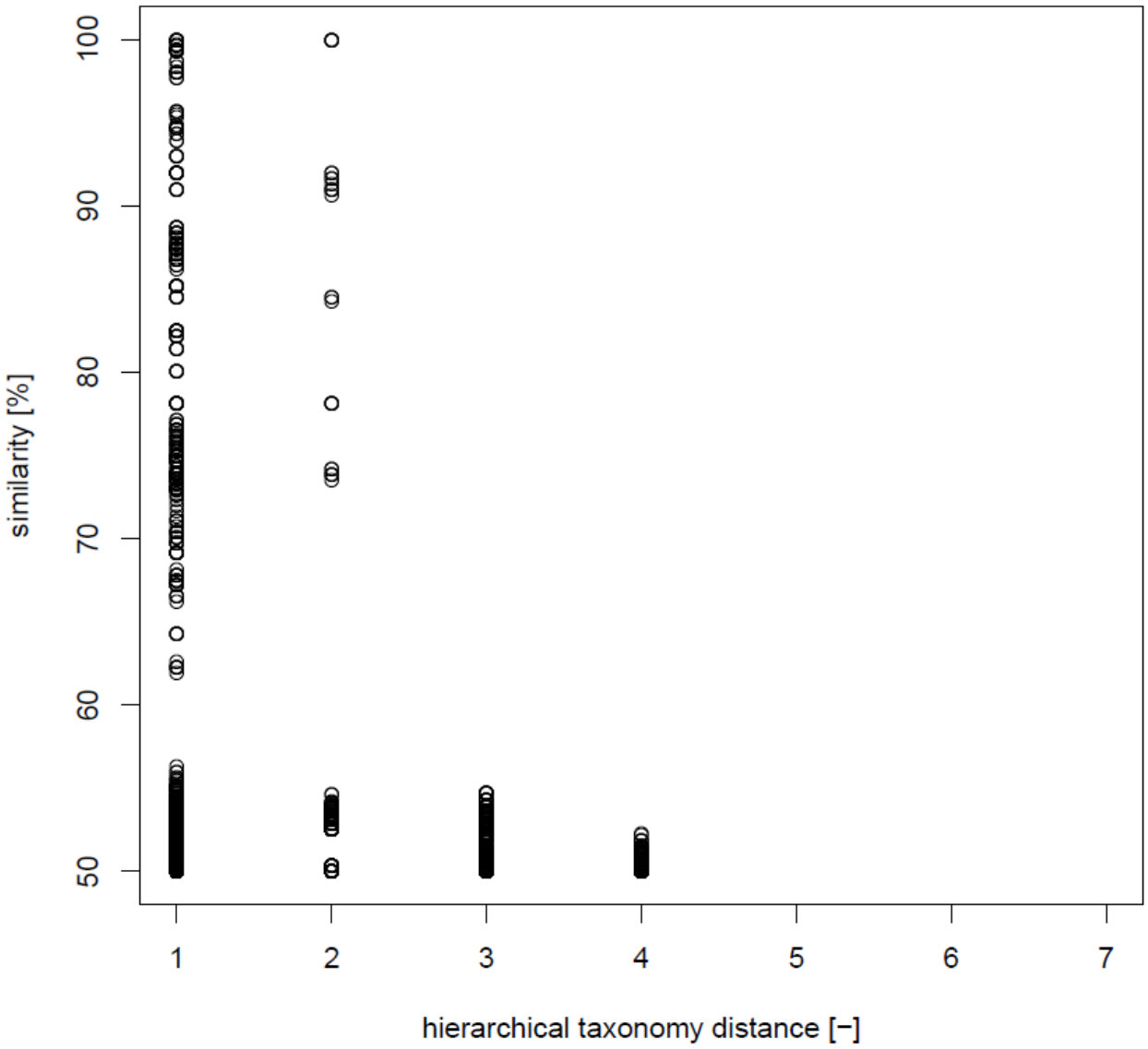

Figure 12: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1076, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

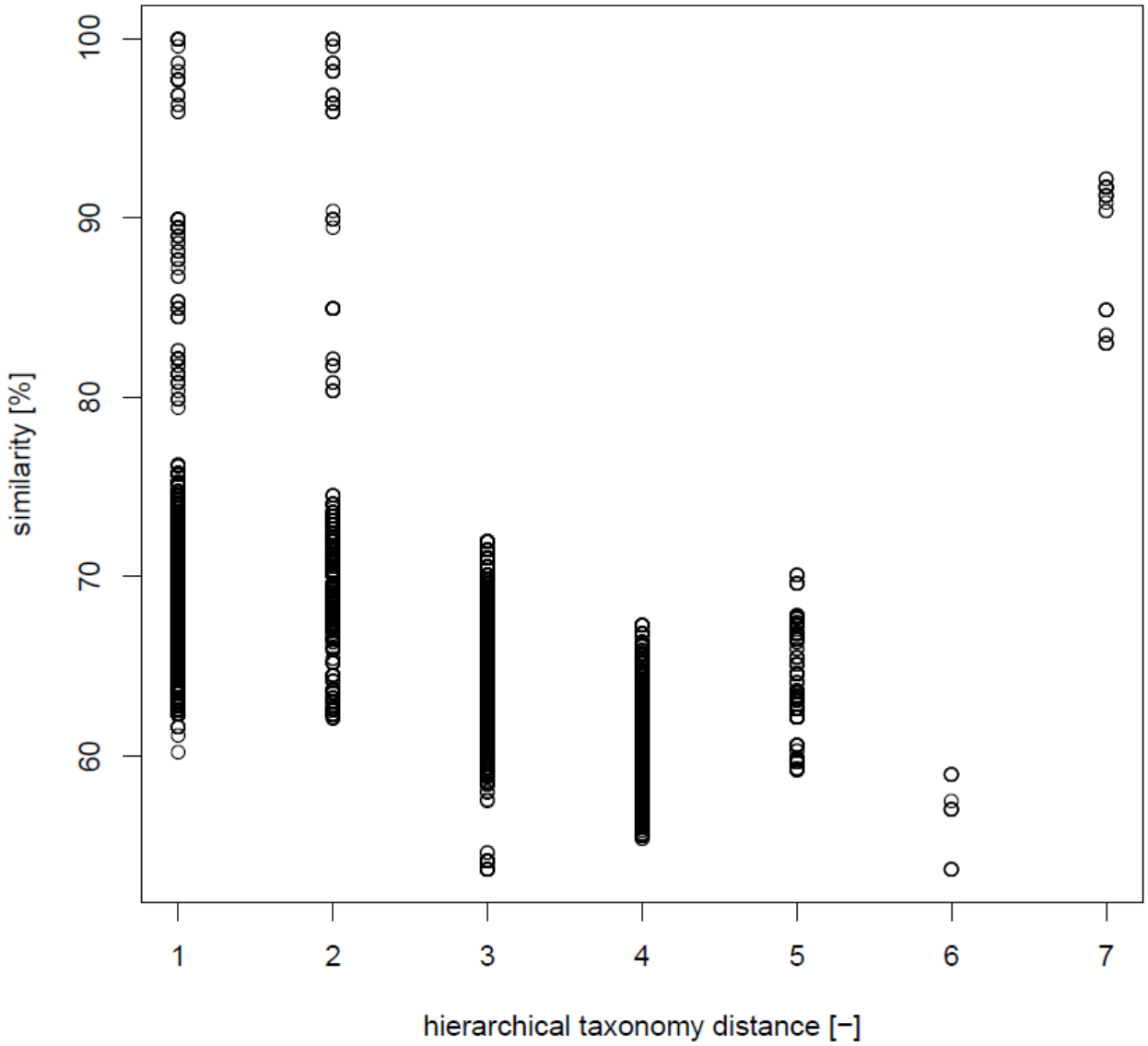

Figure 13: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1077, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

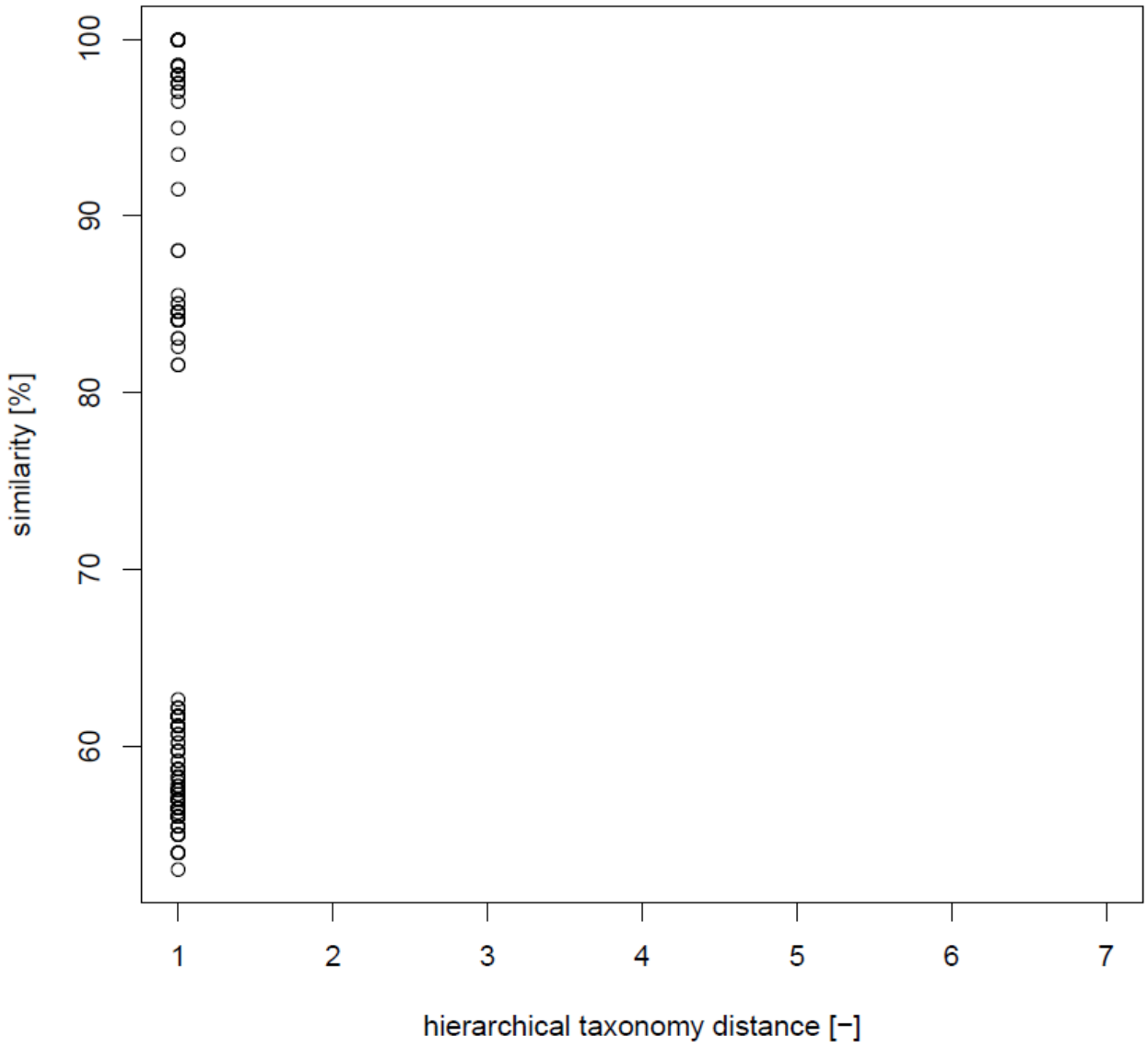

Figure 14: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1078, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

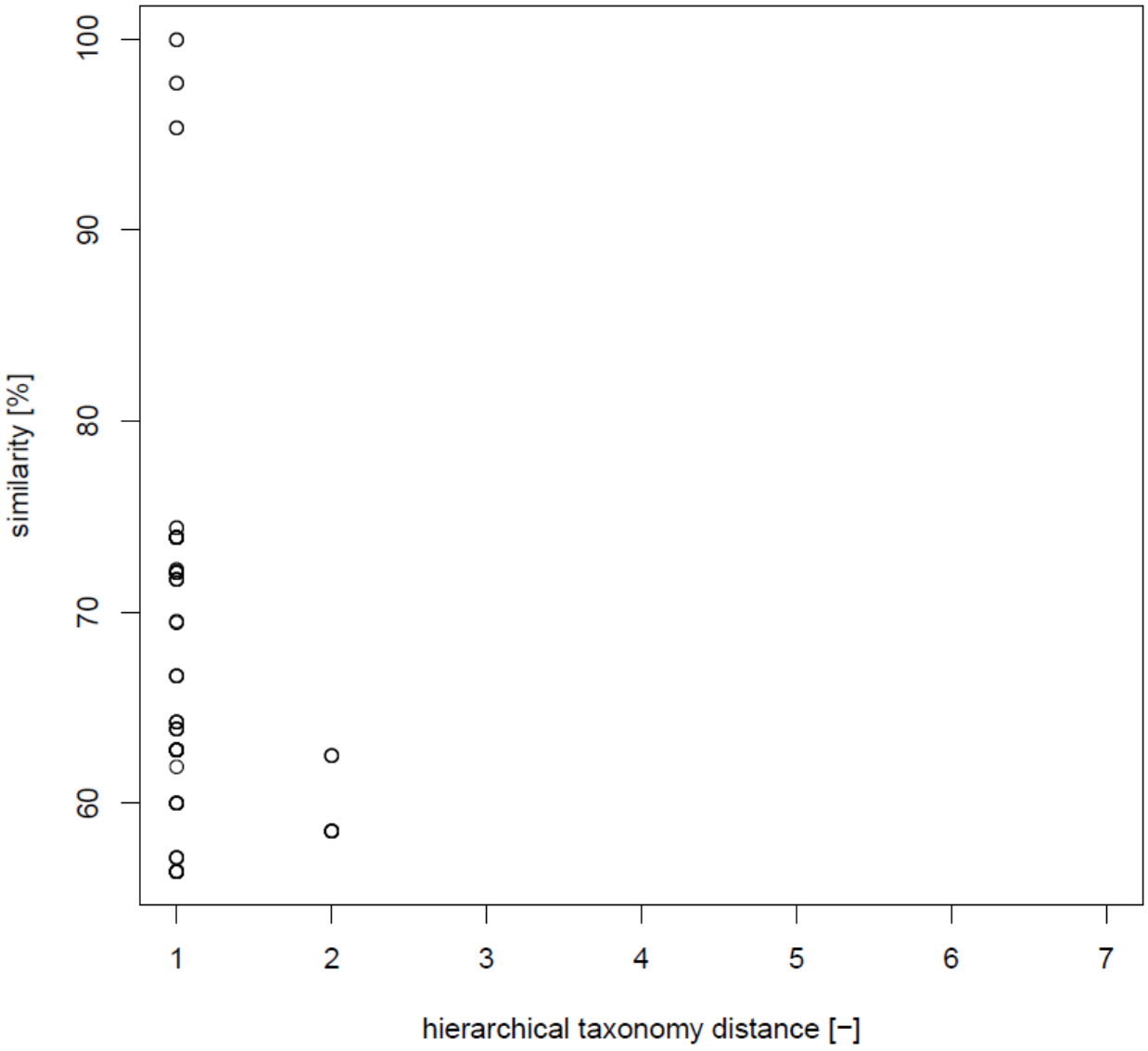

Figure 15: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1080, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

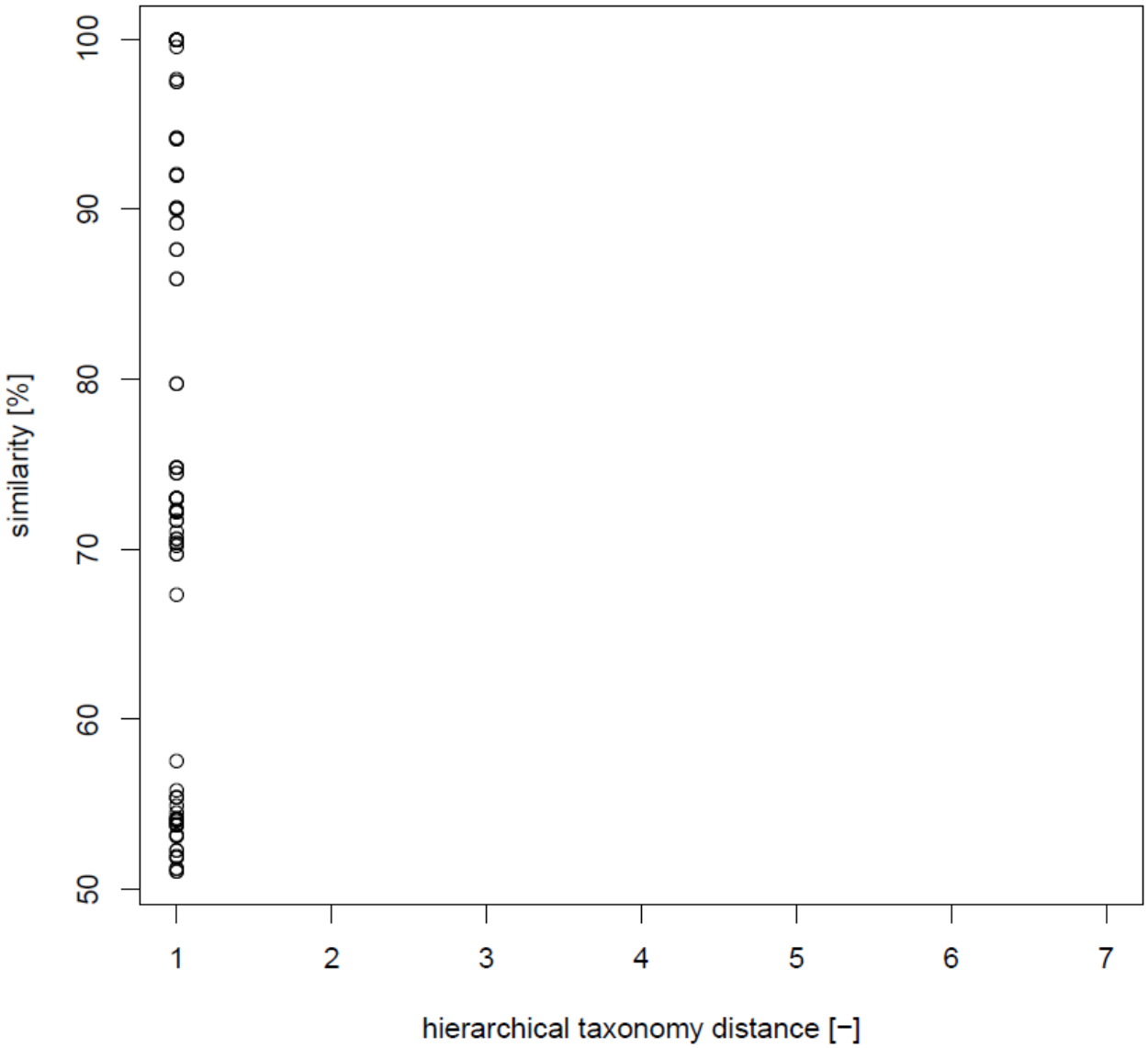

Figure 16: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1083, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

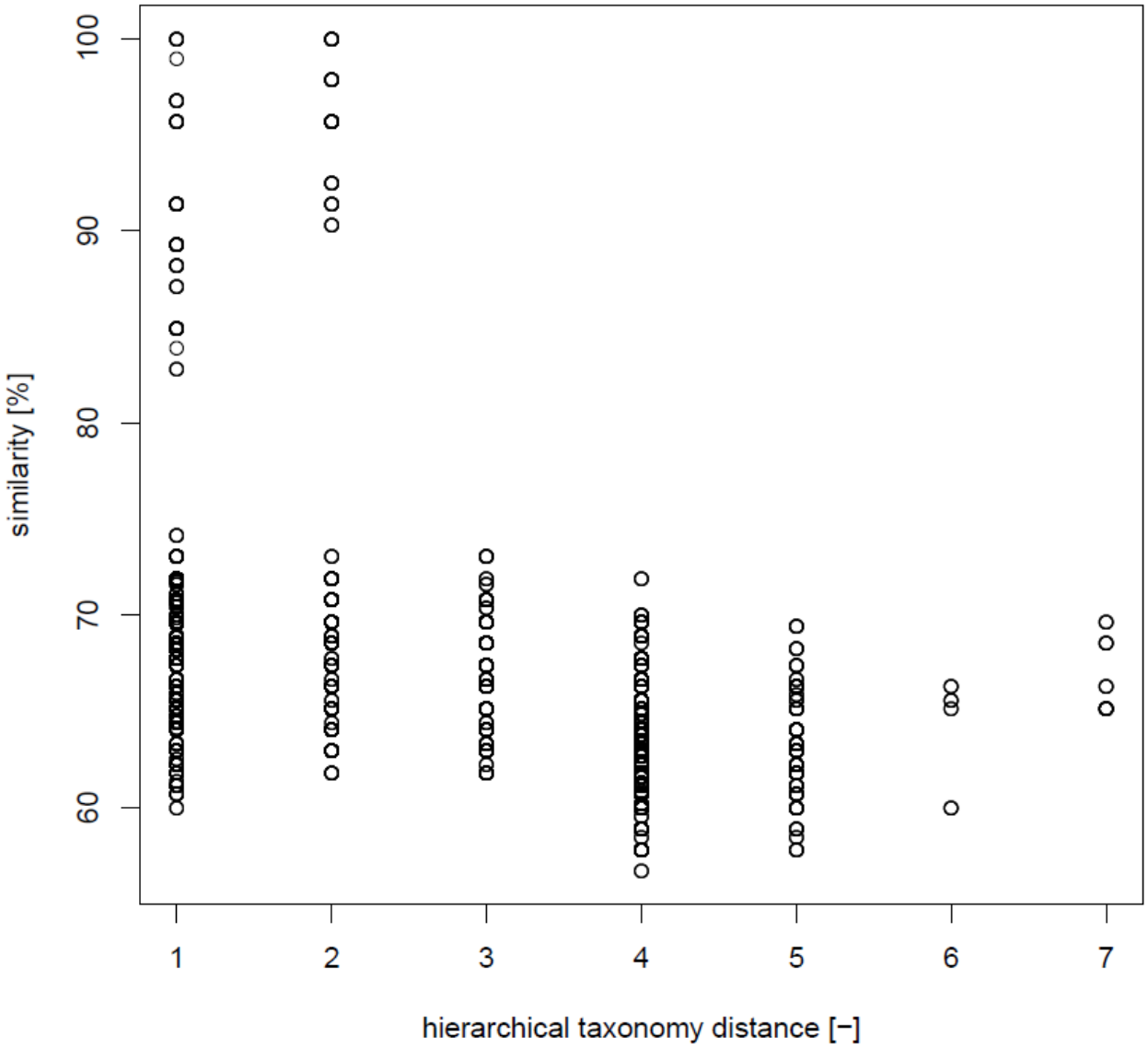

Figure 17: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1084, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

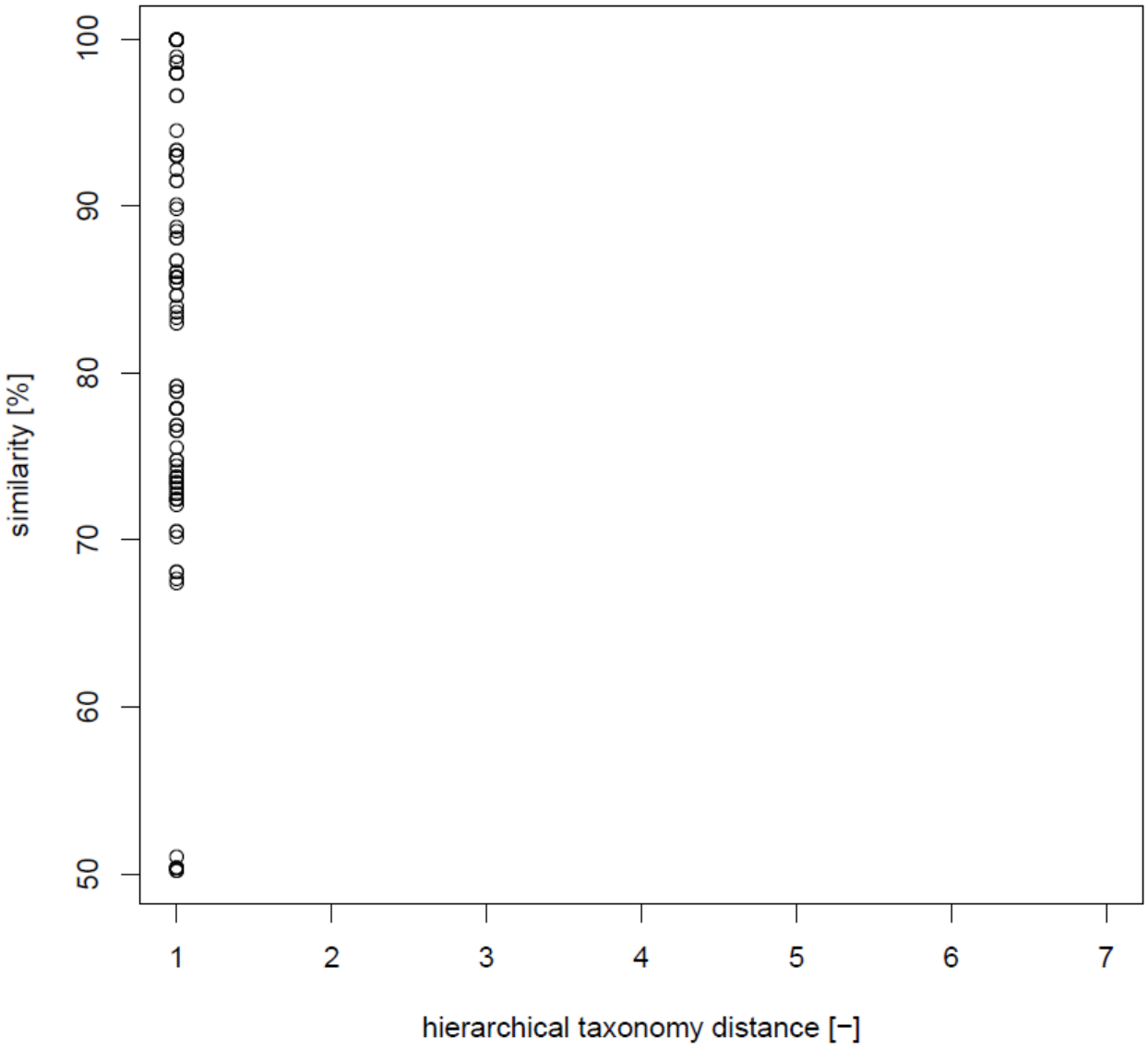

Figure 18: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1086, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

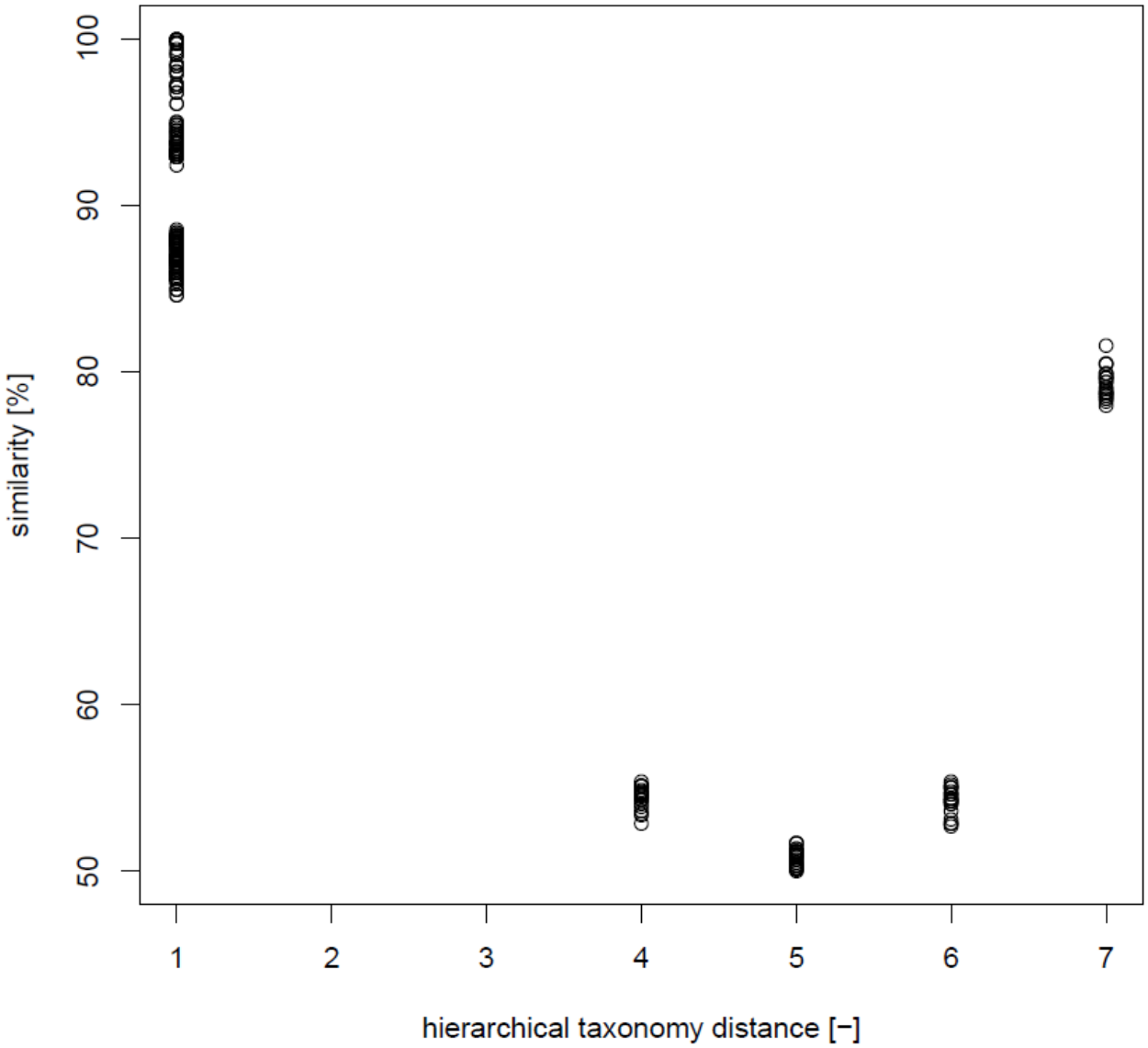

Figure 19: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1096, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

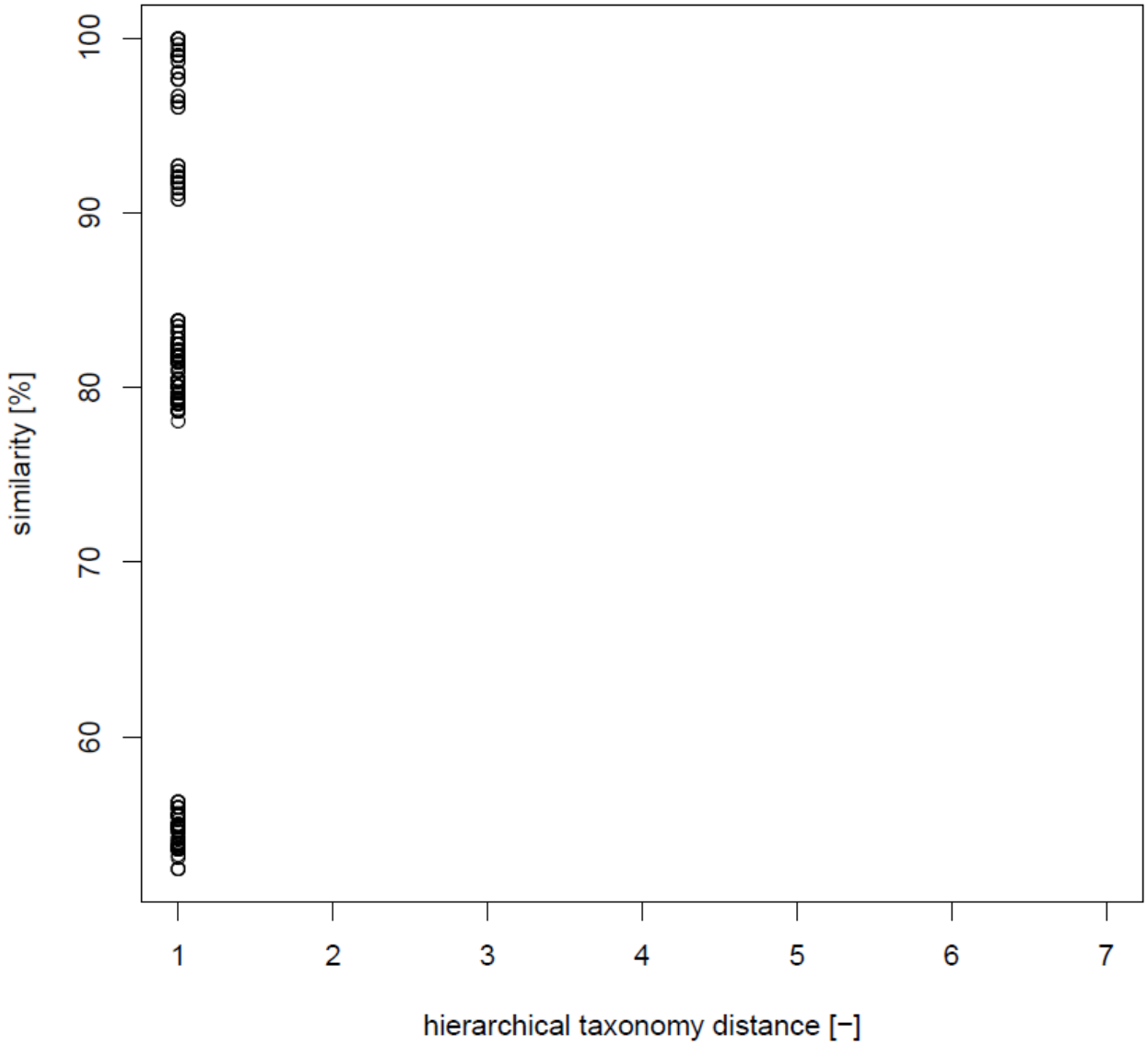

Figure 20: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1097, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

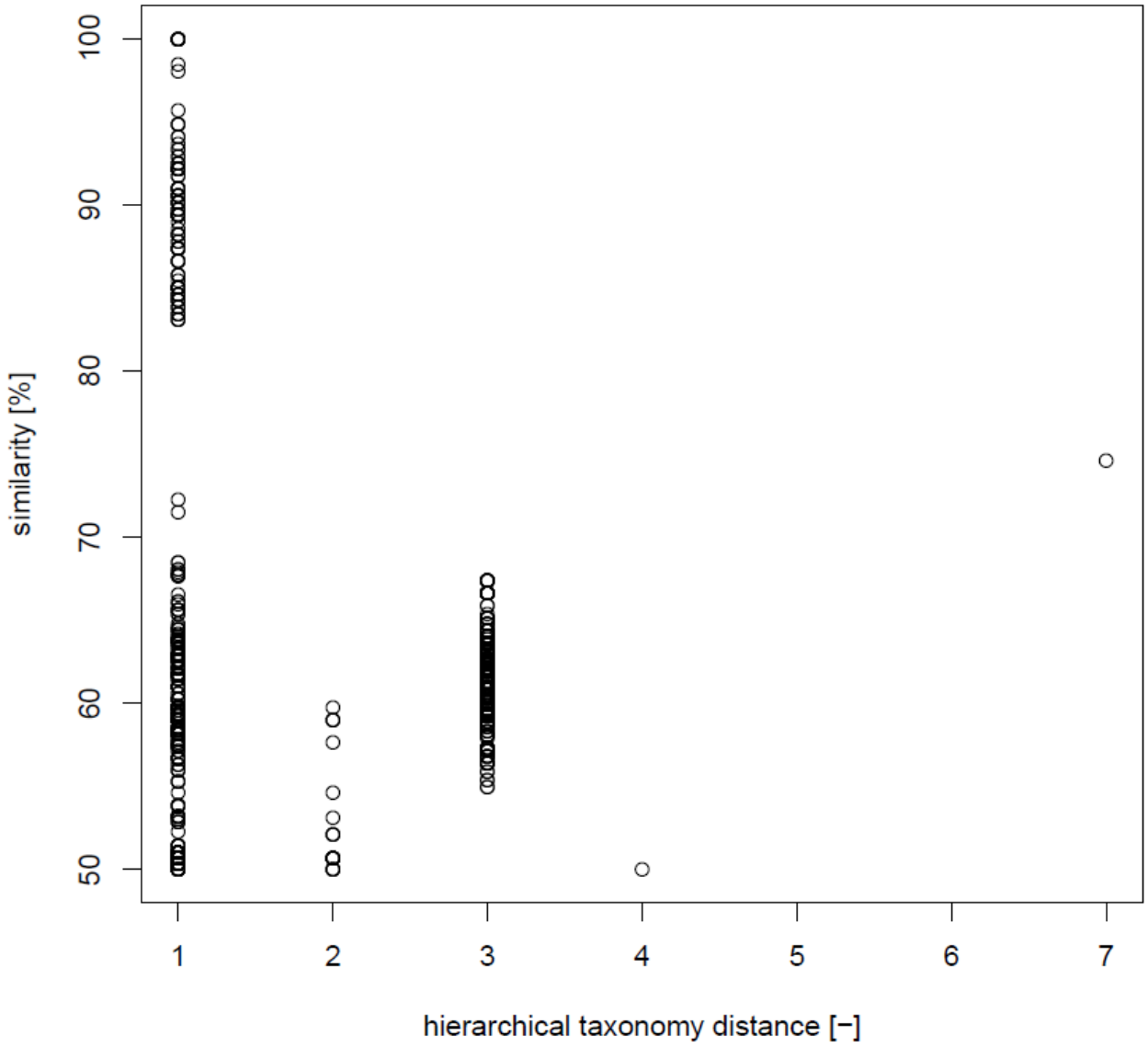

Figure 21: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1099, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

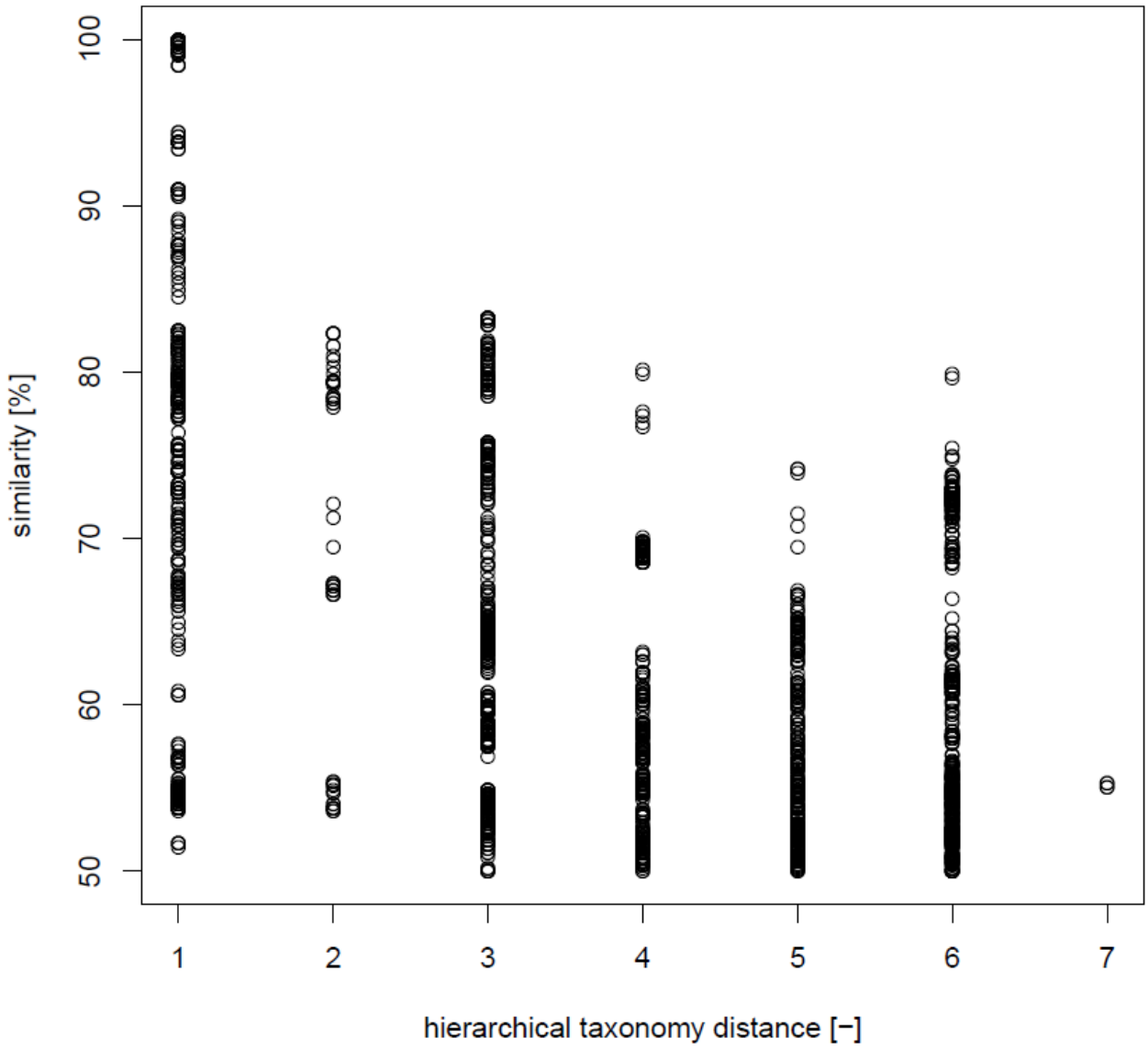

Figure 22: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 45, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

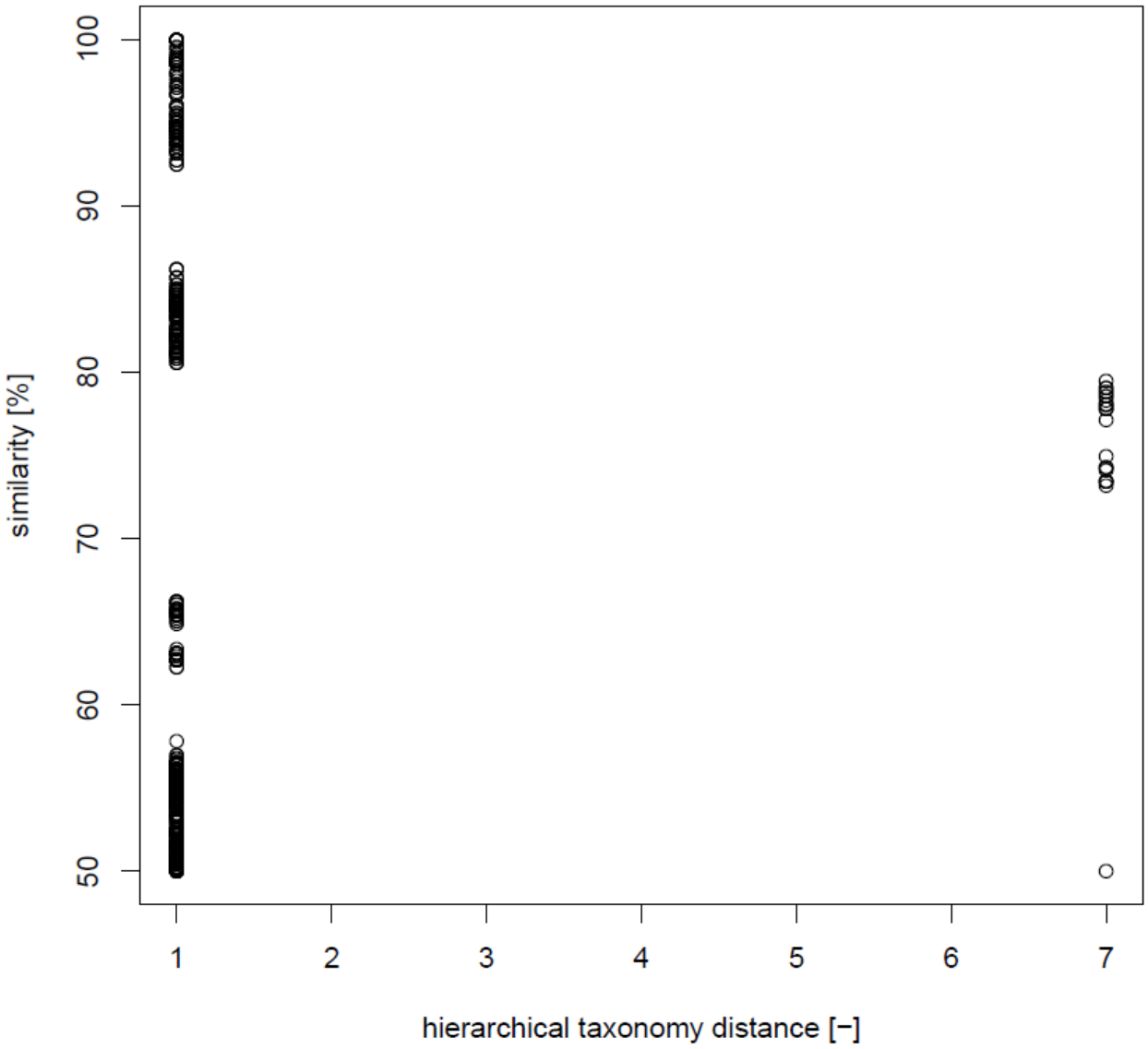

Figure 23: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 55, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

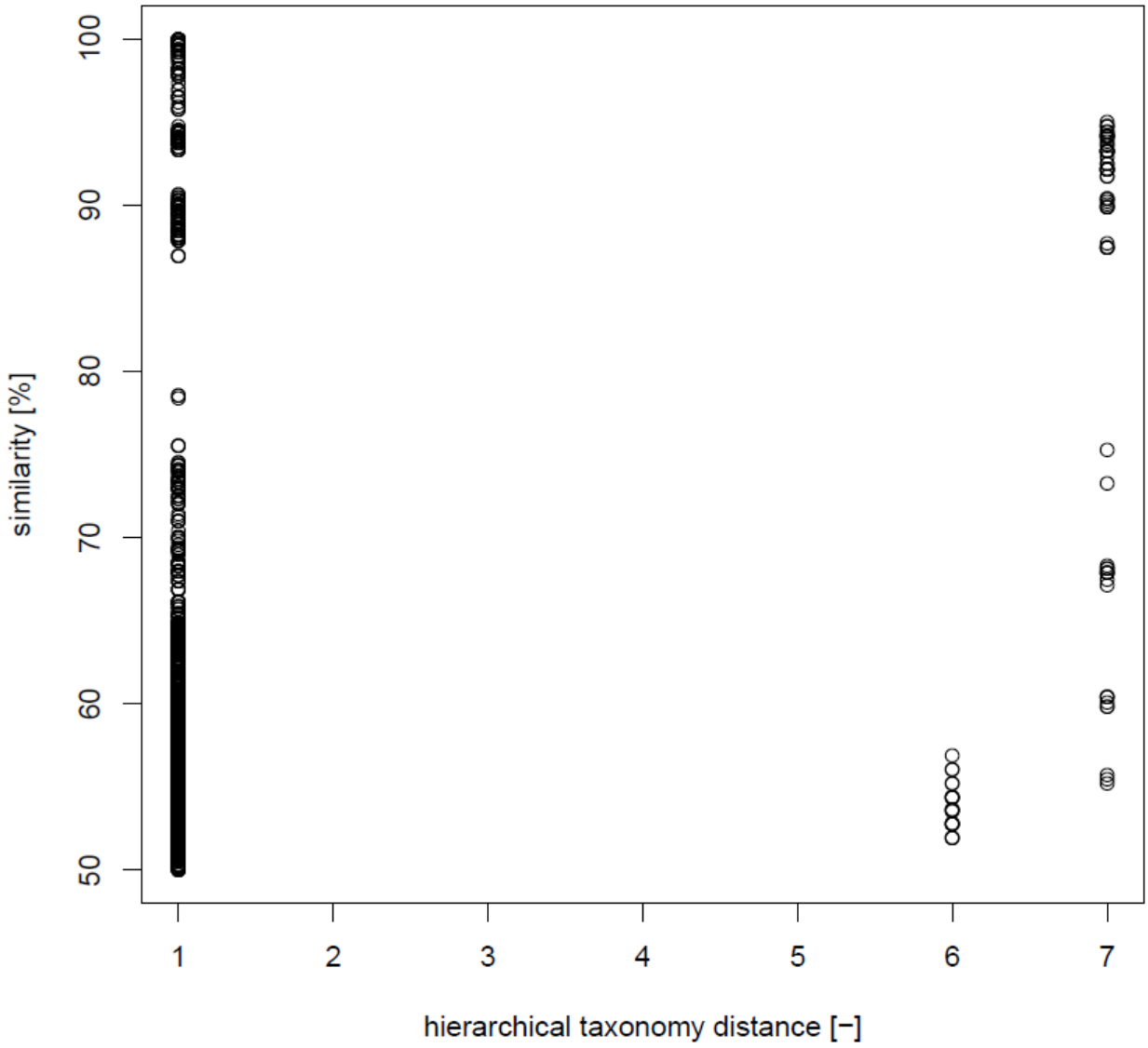

Figure 24: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 642, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

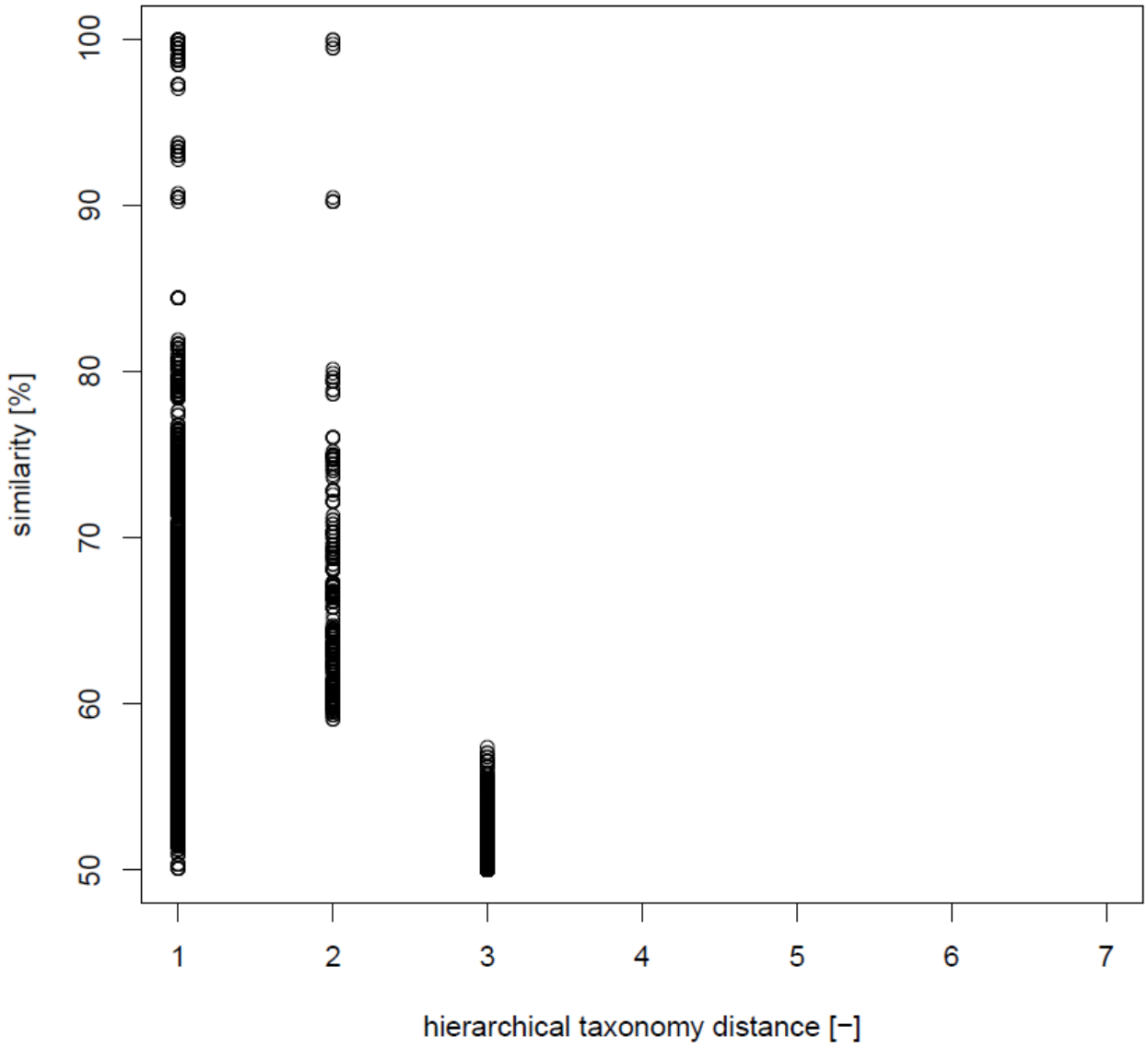

Figure 25: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 83, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

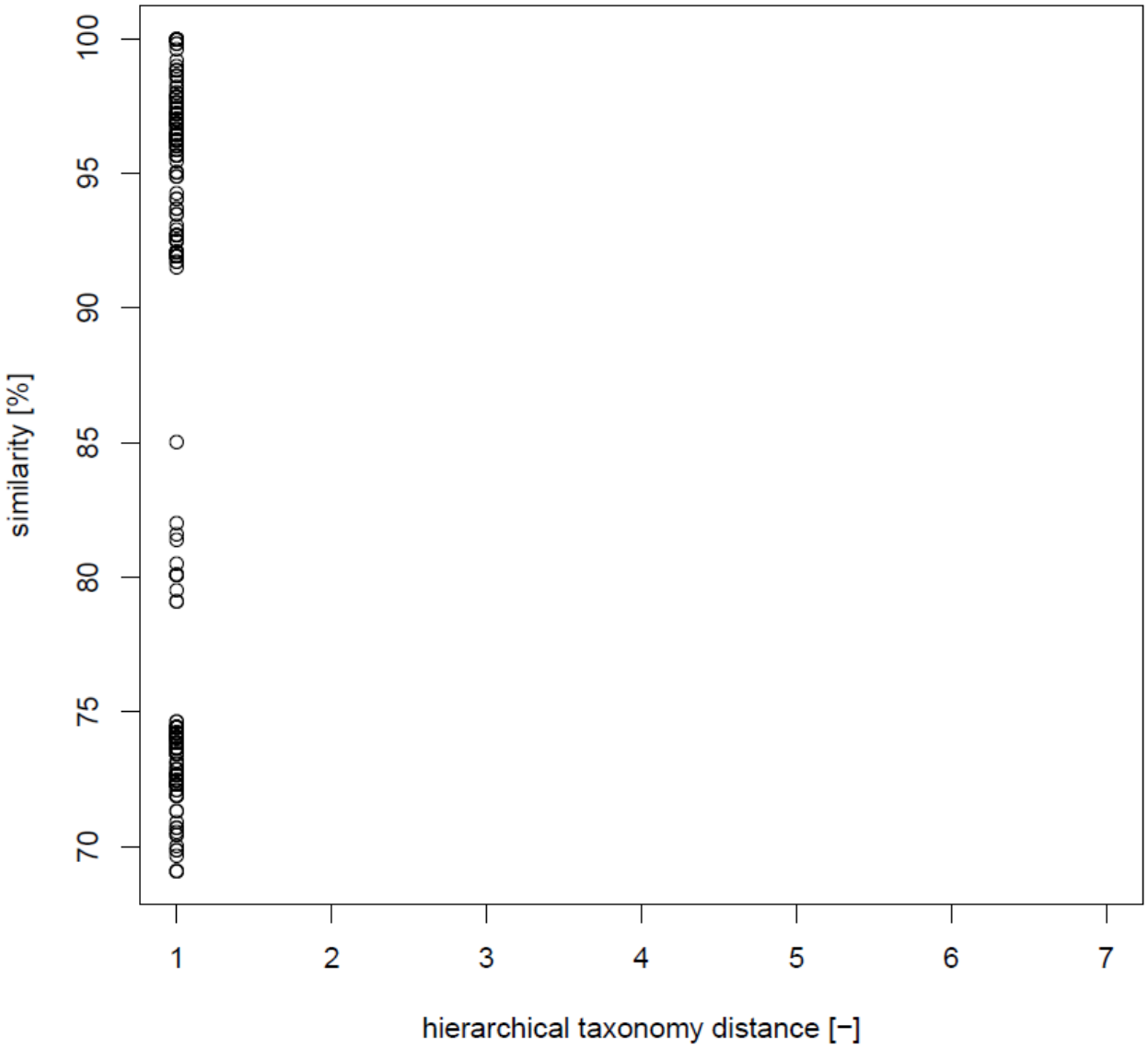

Figure 26: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 896, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

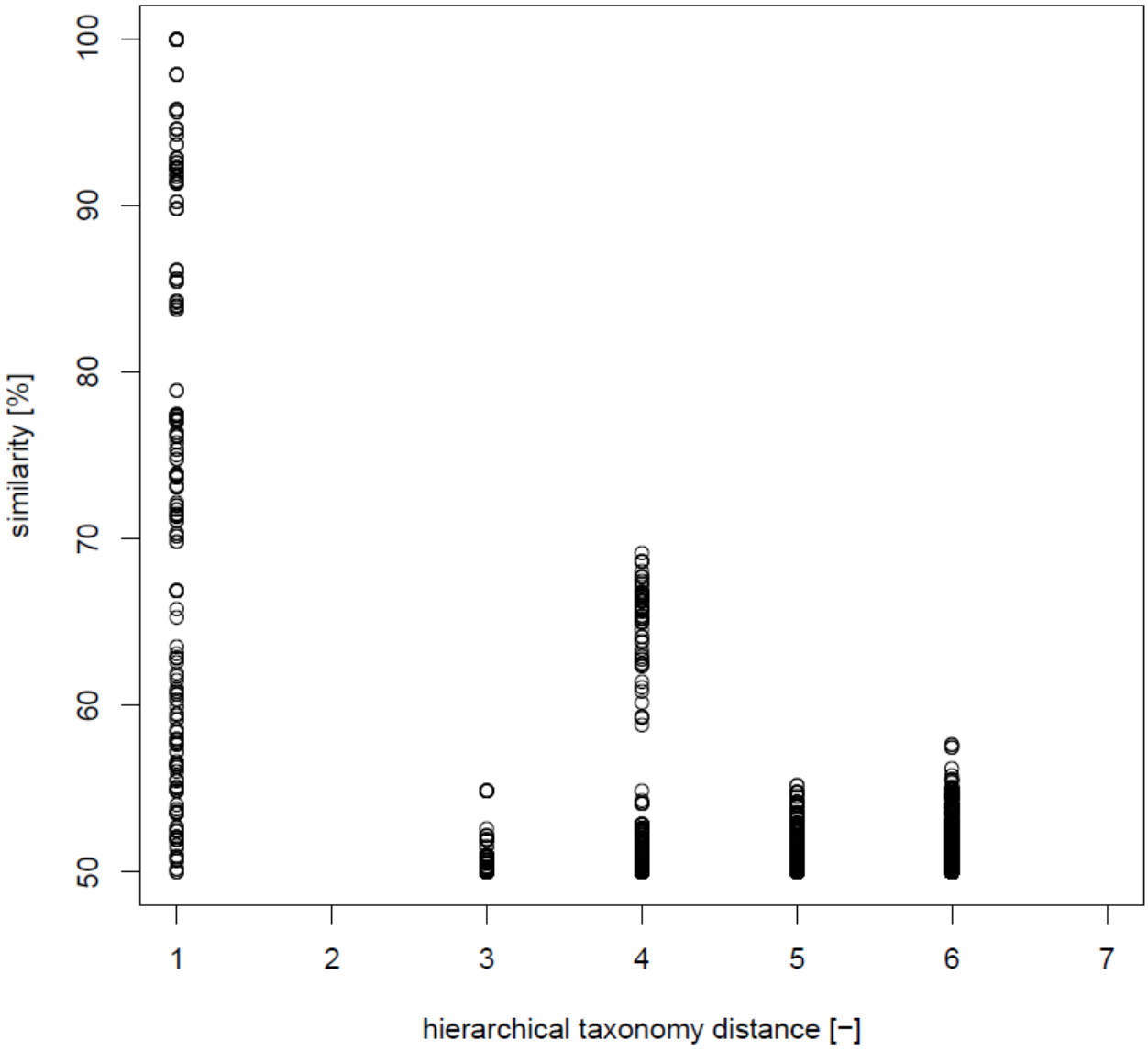

Figure 27: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 991, specific to bumble and honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

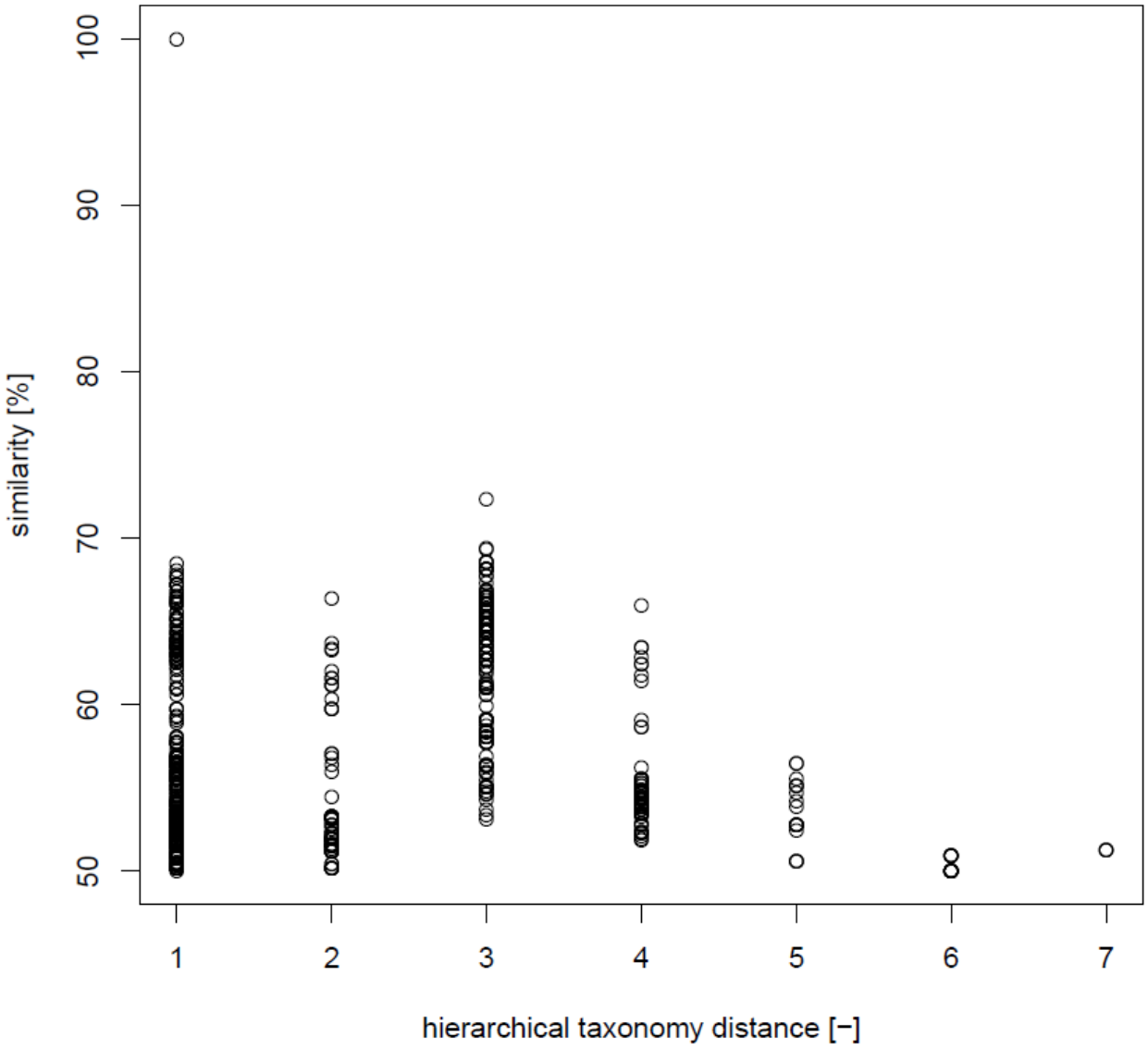

Figure 28: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1631, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

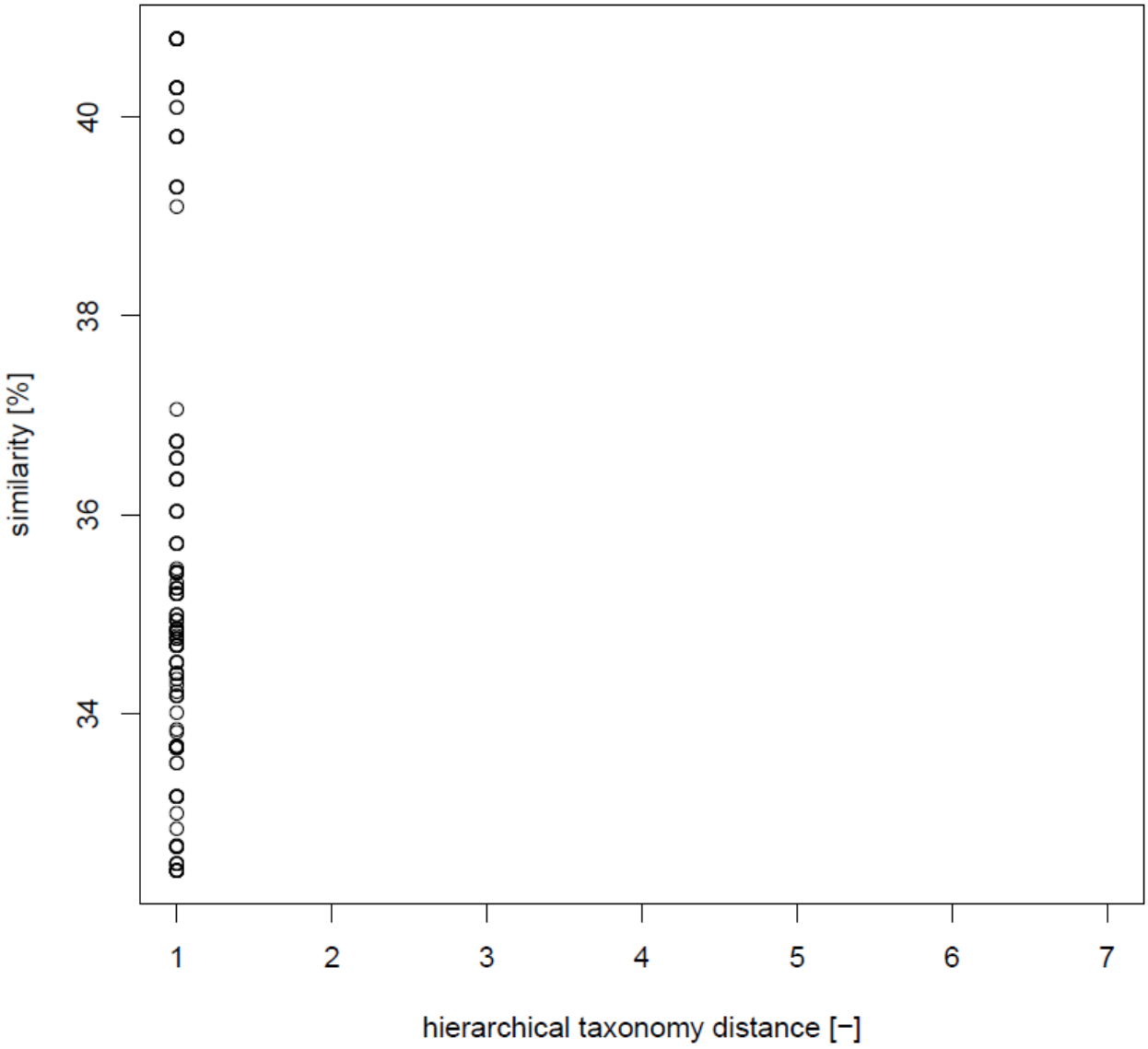

Figure 29: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1670, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

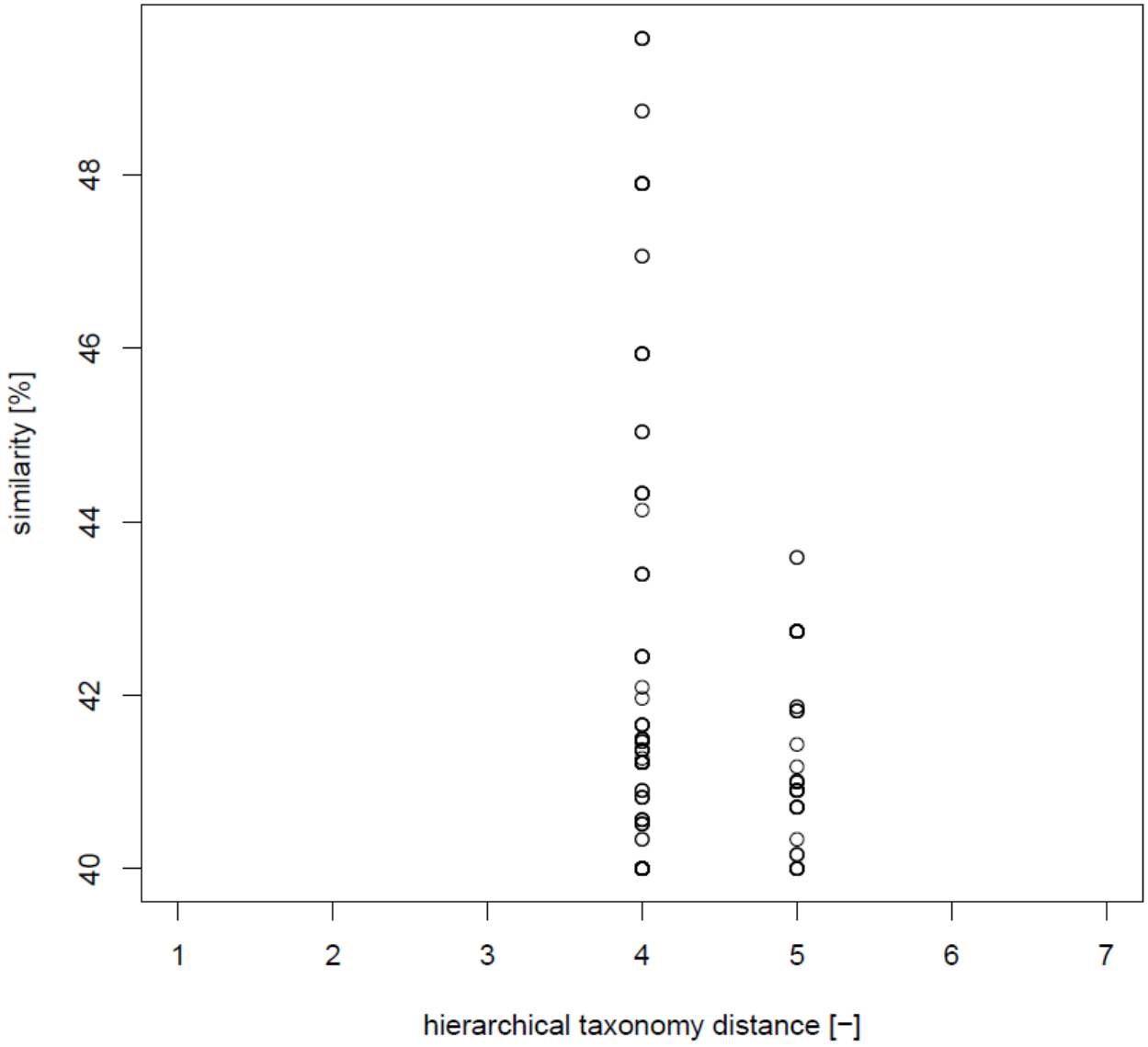

Figure 30: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1675, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

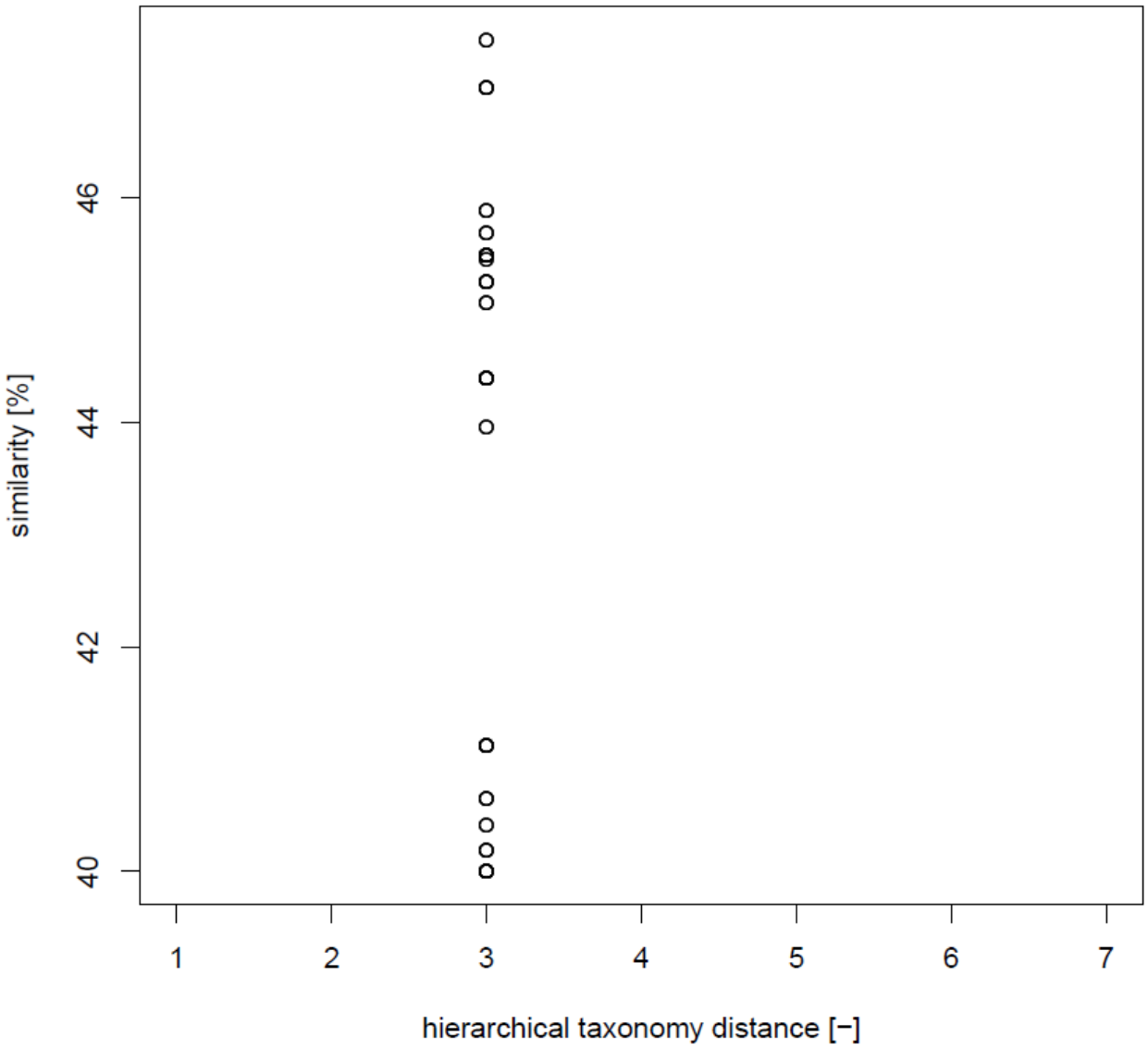

Figure 31: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1678, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

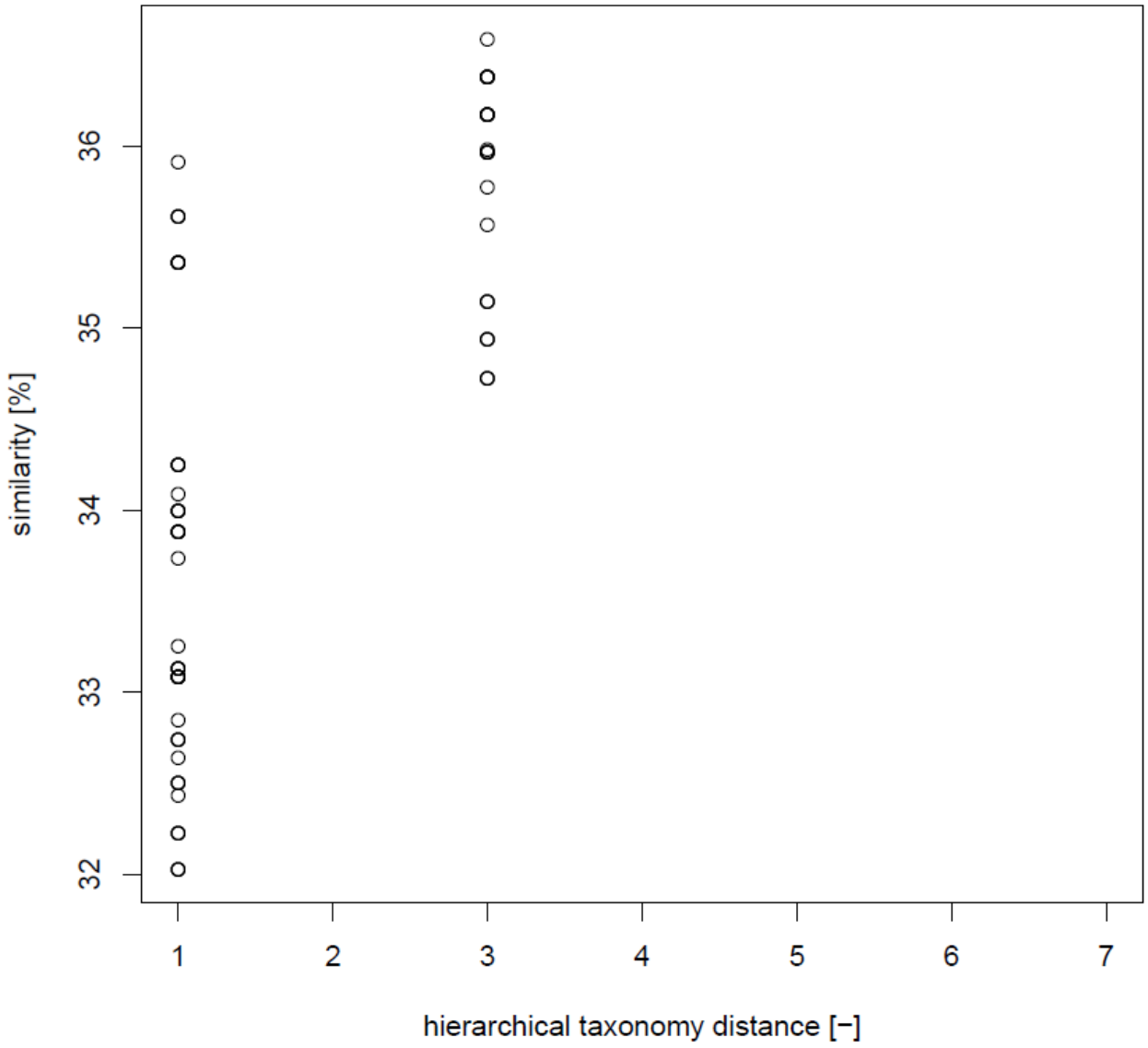

Figure 32: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1685, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

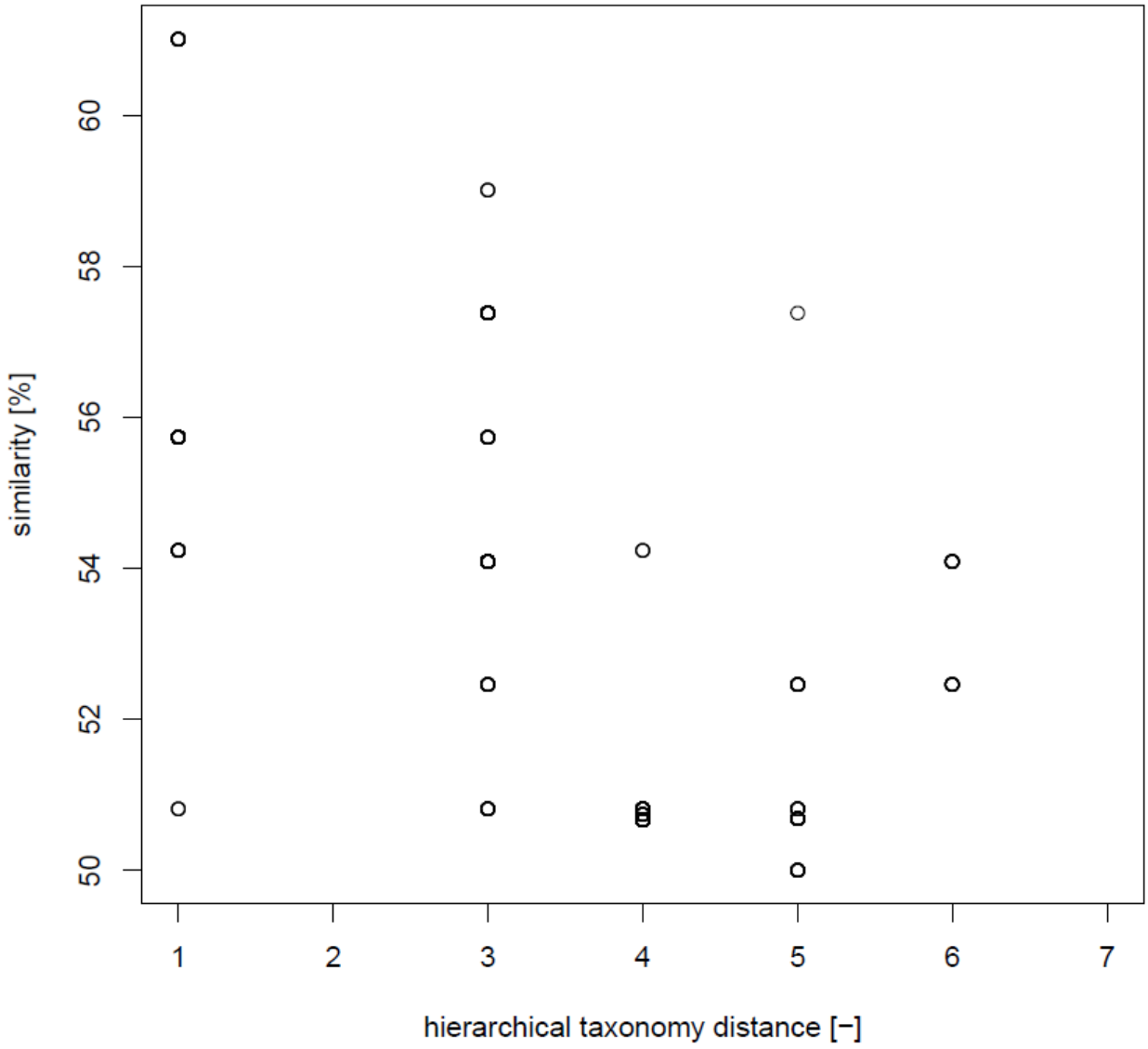

Figure 33: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1689, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

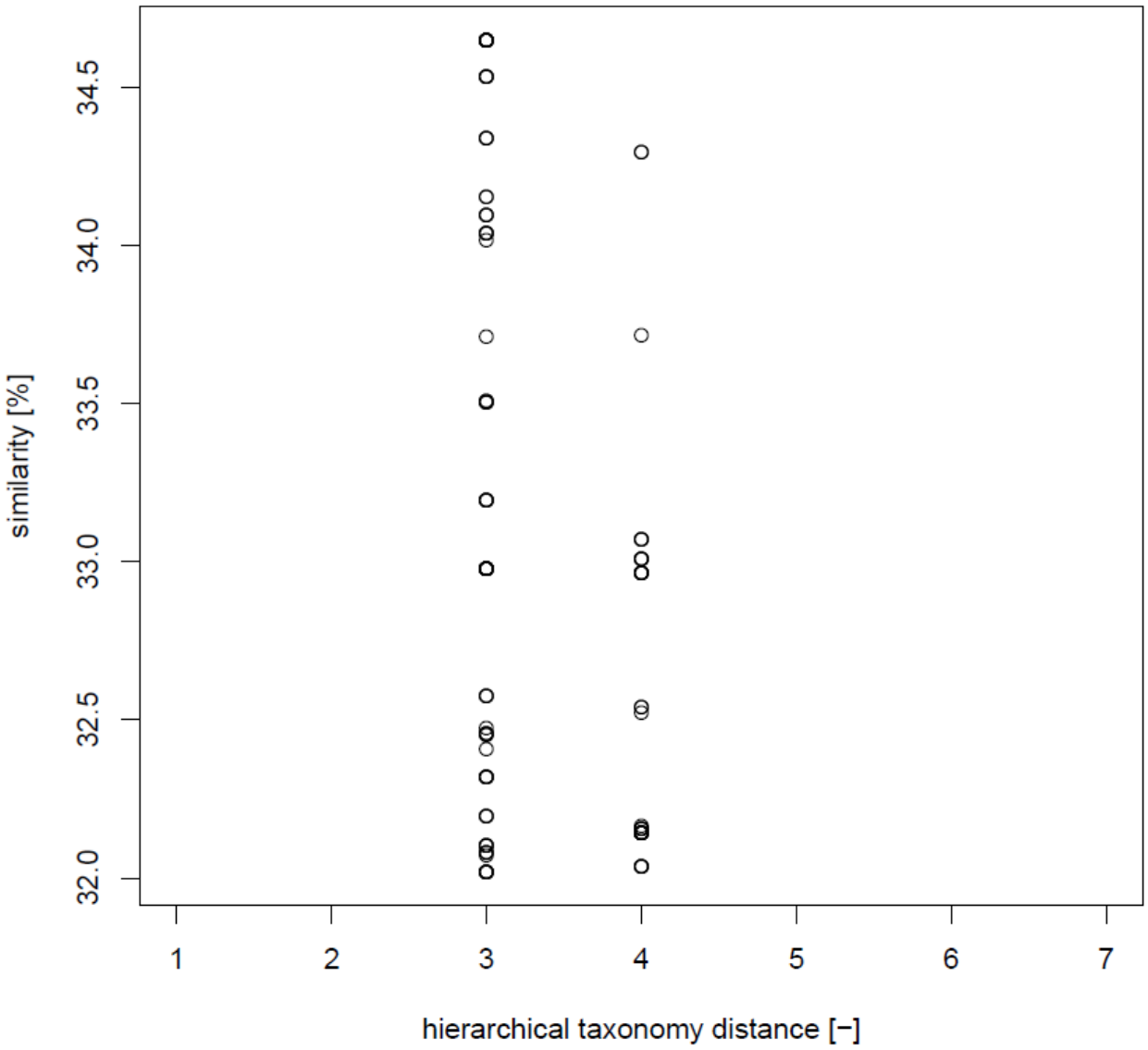

Figure 34: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1757, specific to bumble bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

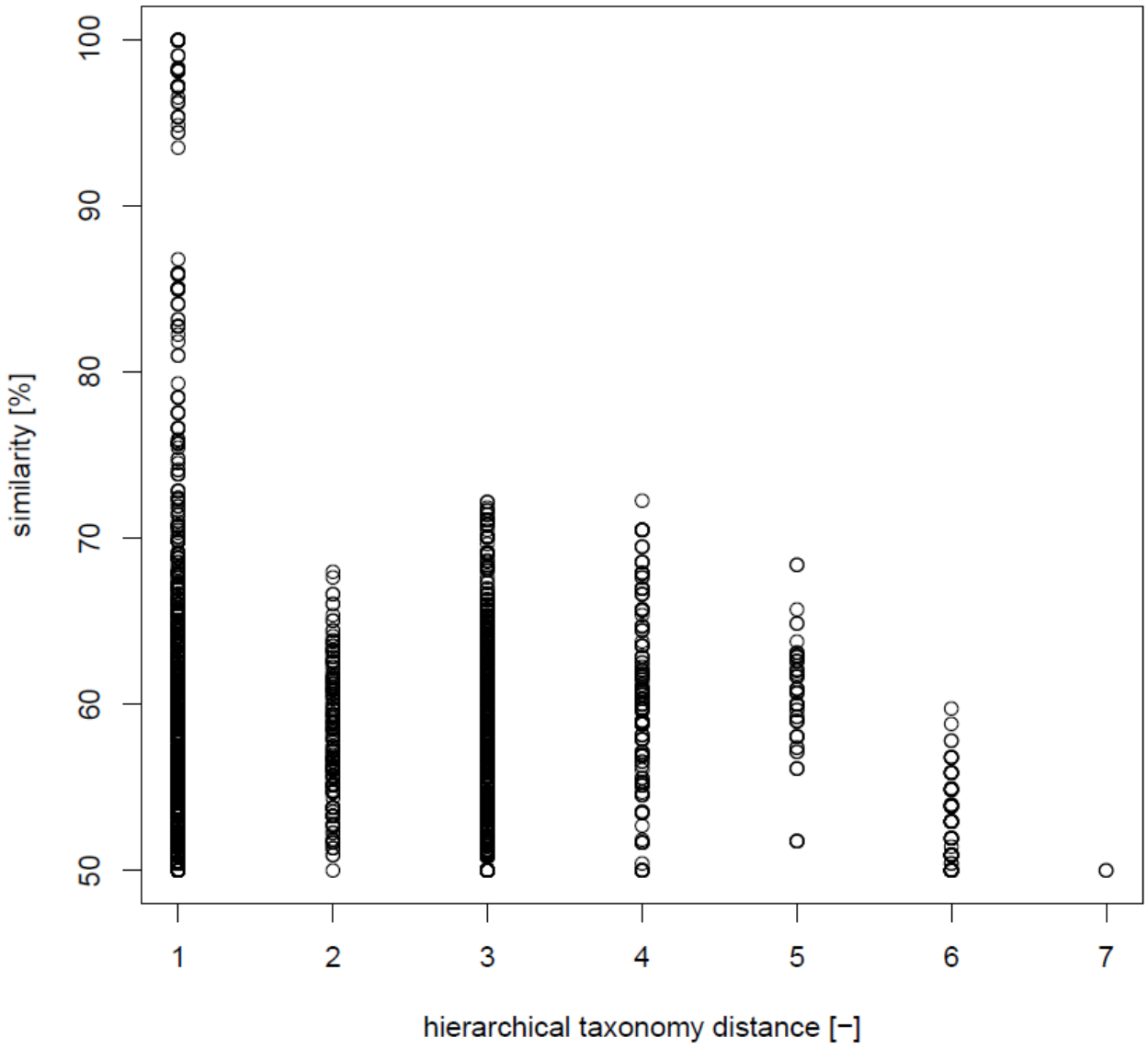

Figure 35: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1168, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

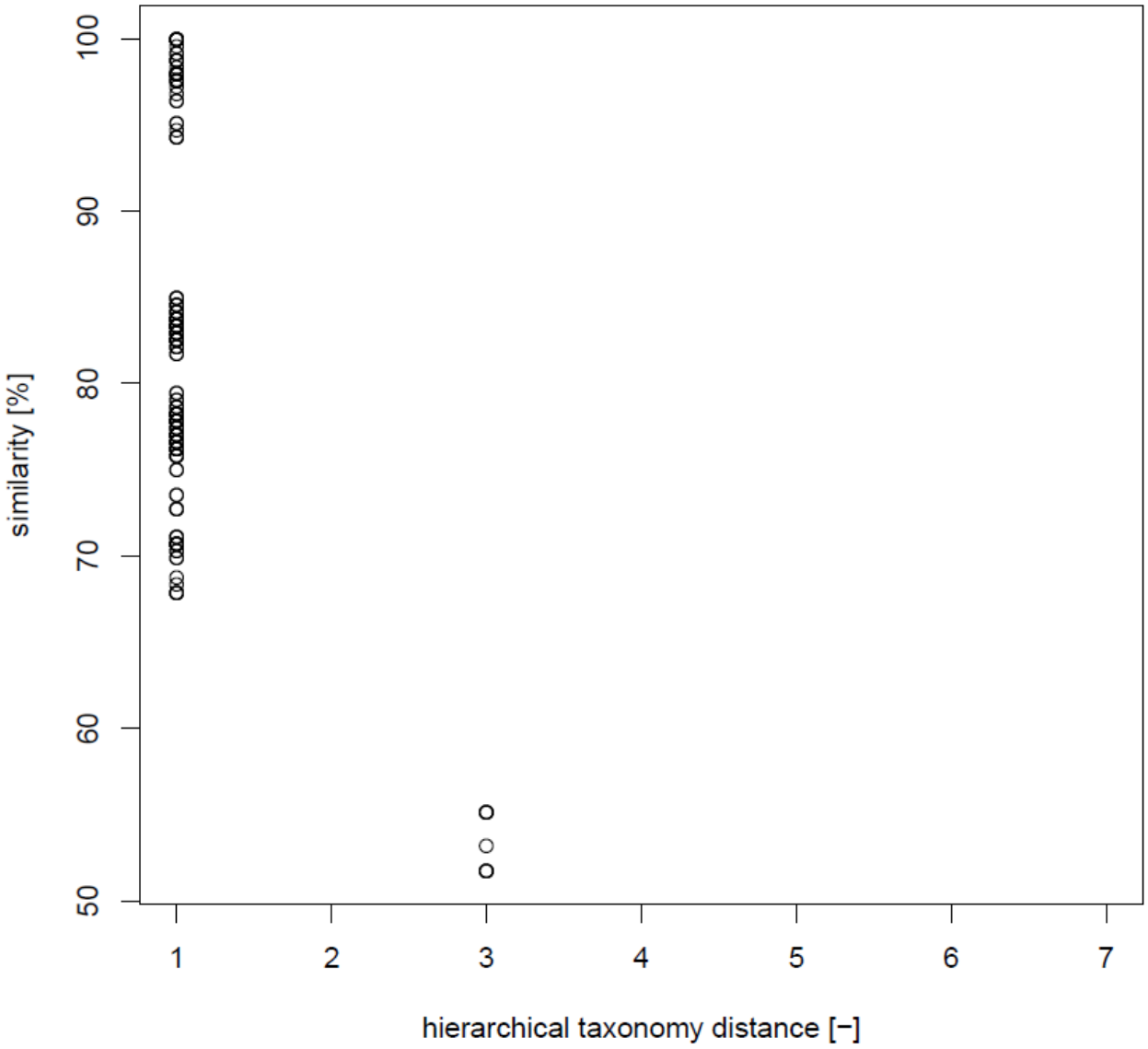

Figure 36: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1364, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

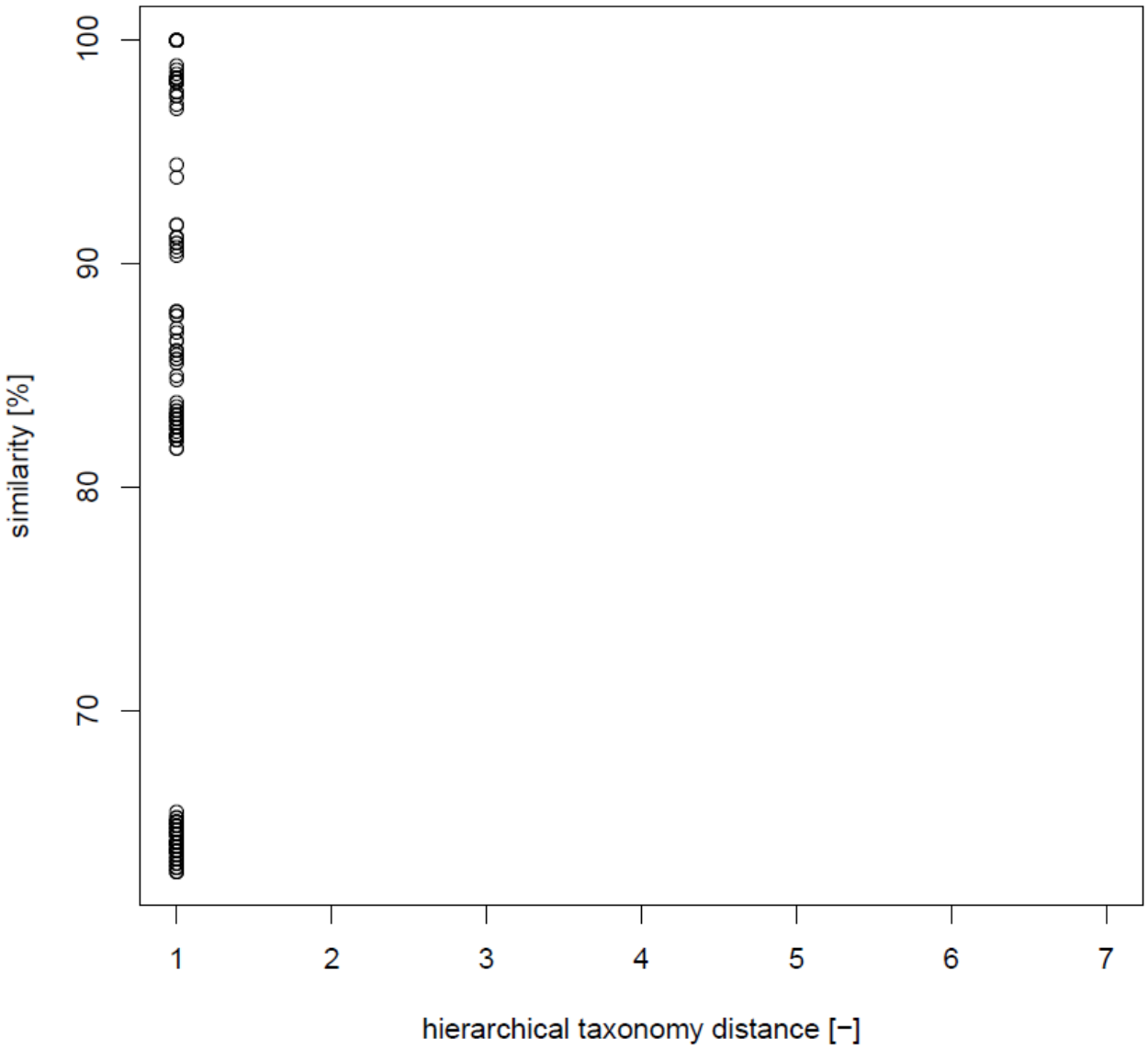

Figure 37: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1378, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

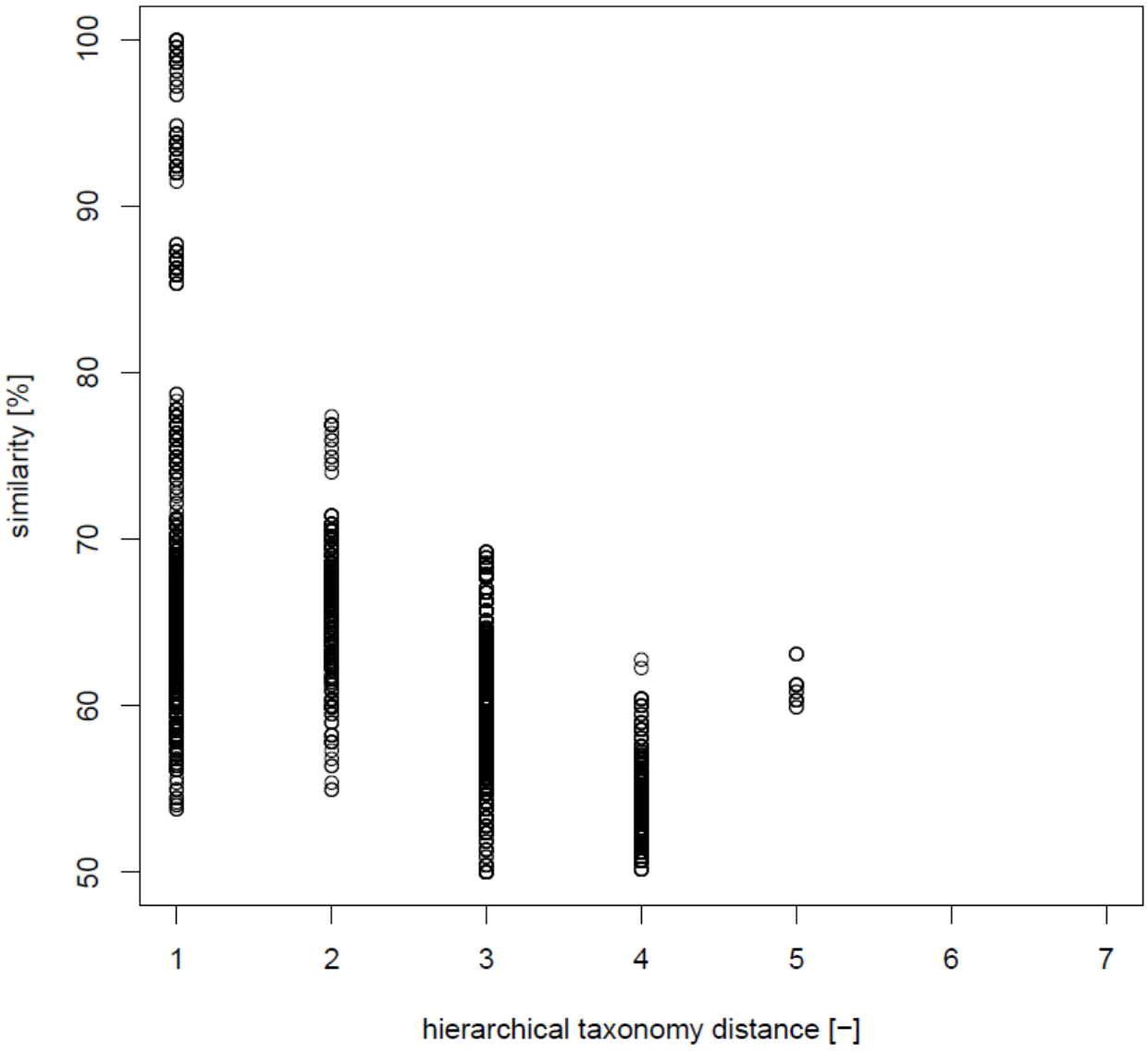

Figure 38: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1380, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

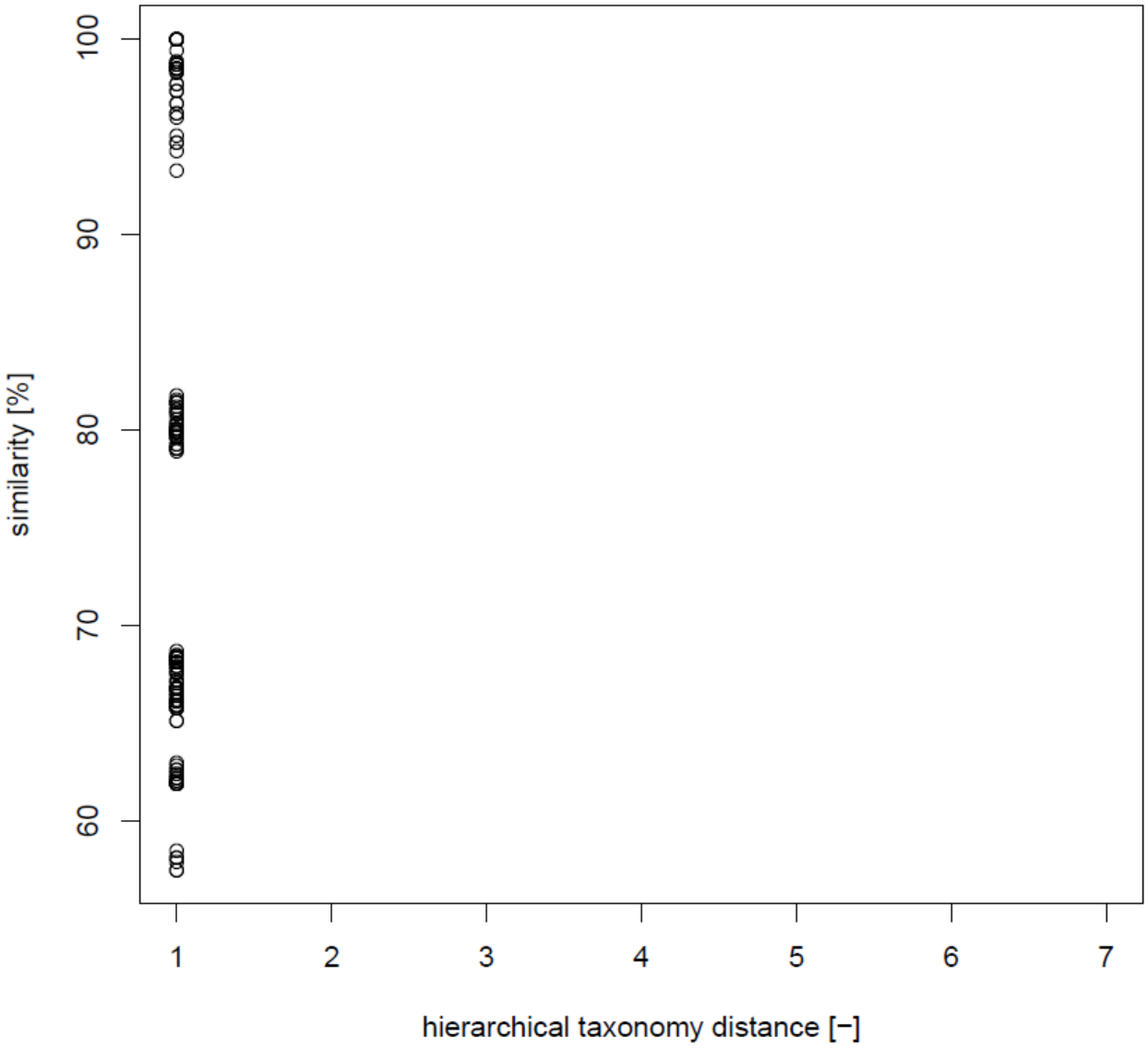

Figure 39: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1385, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

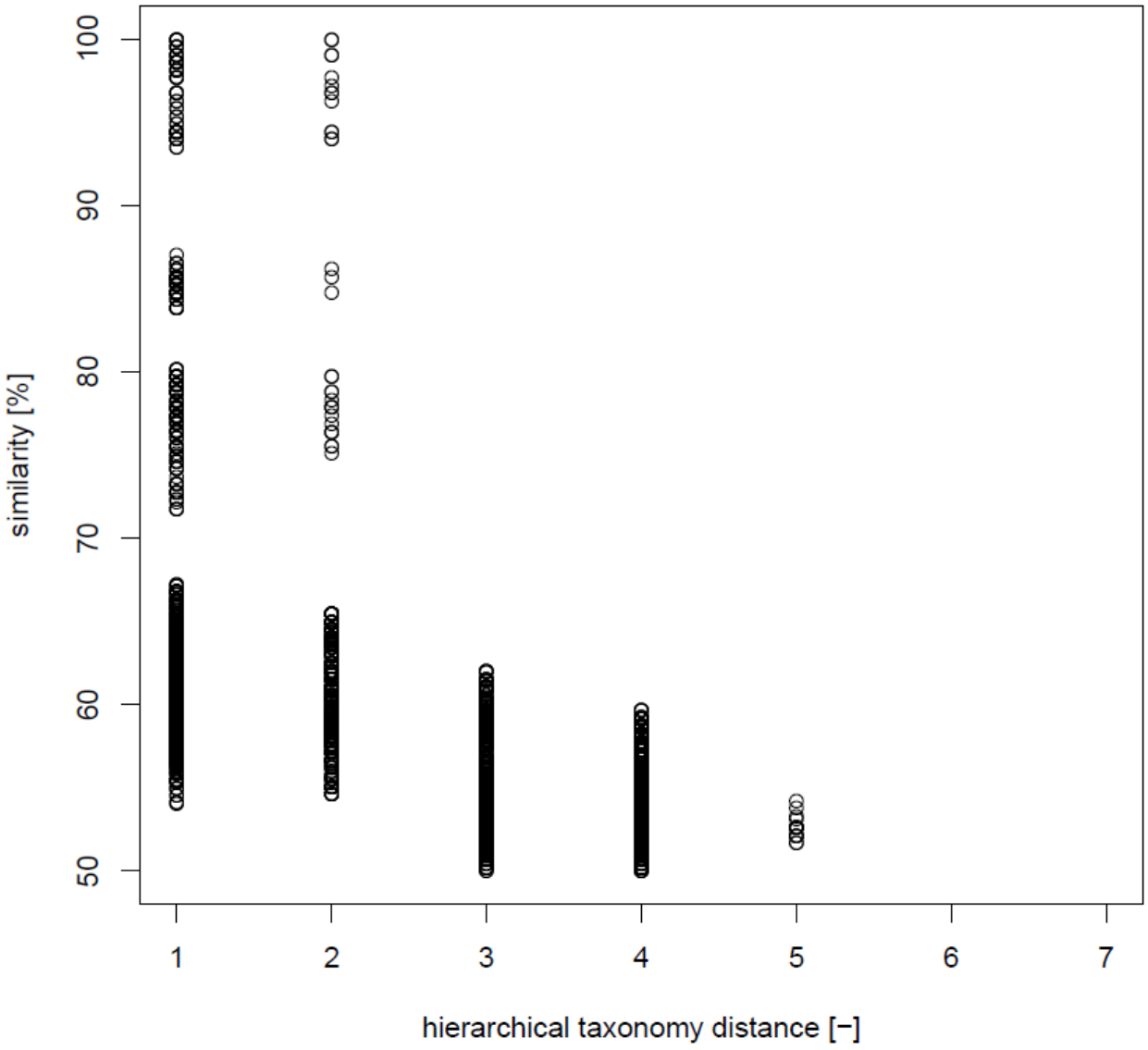

Figure 40: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1393, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

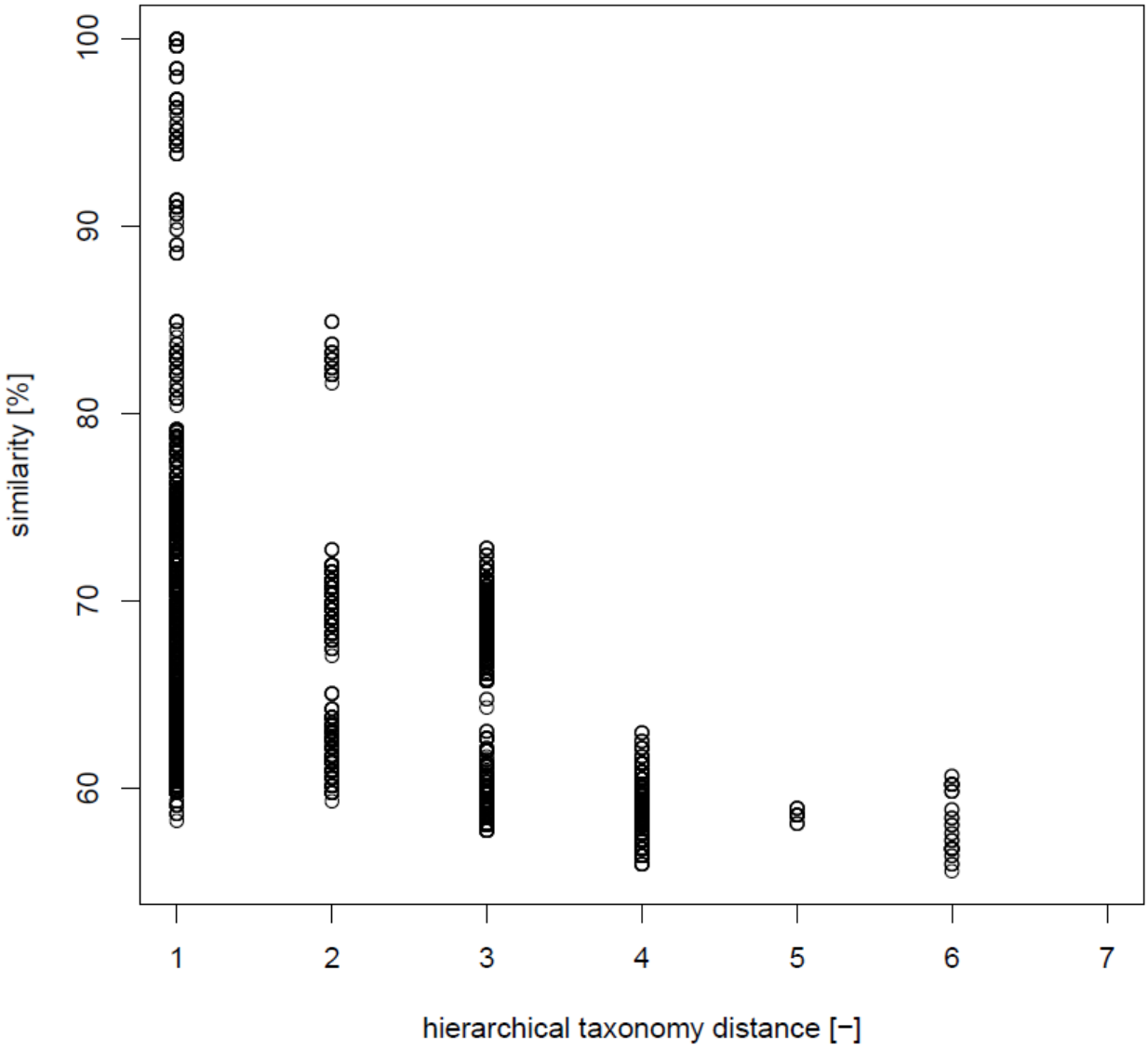

Figure 41: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1394, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

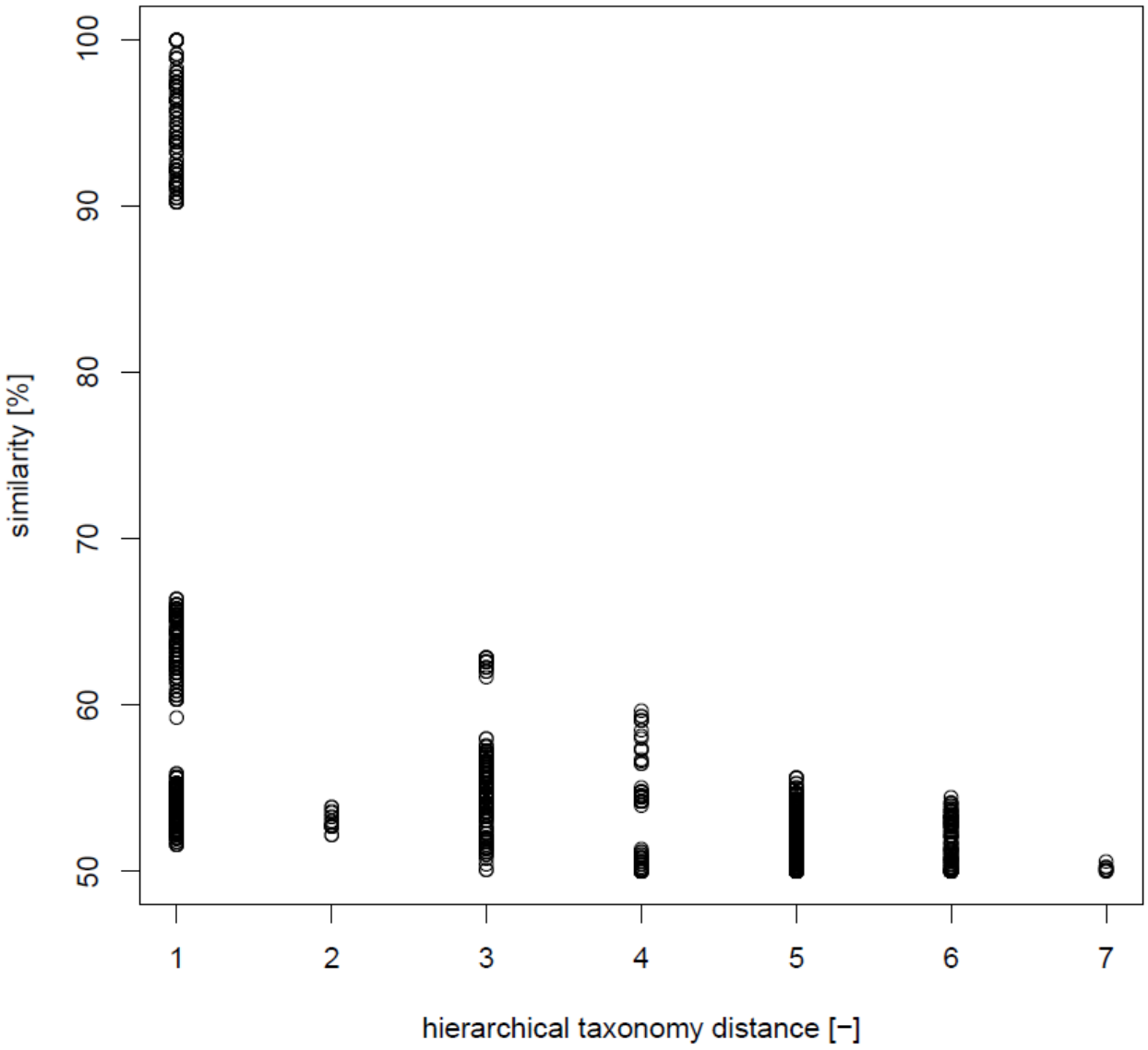

Figure 42: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1396, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

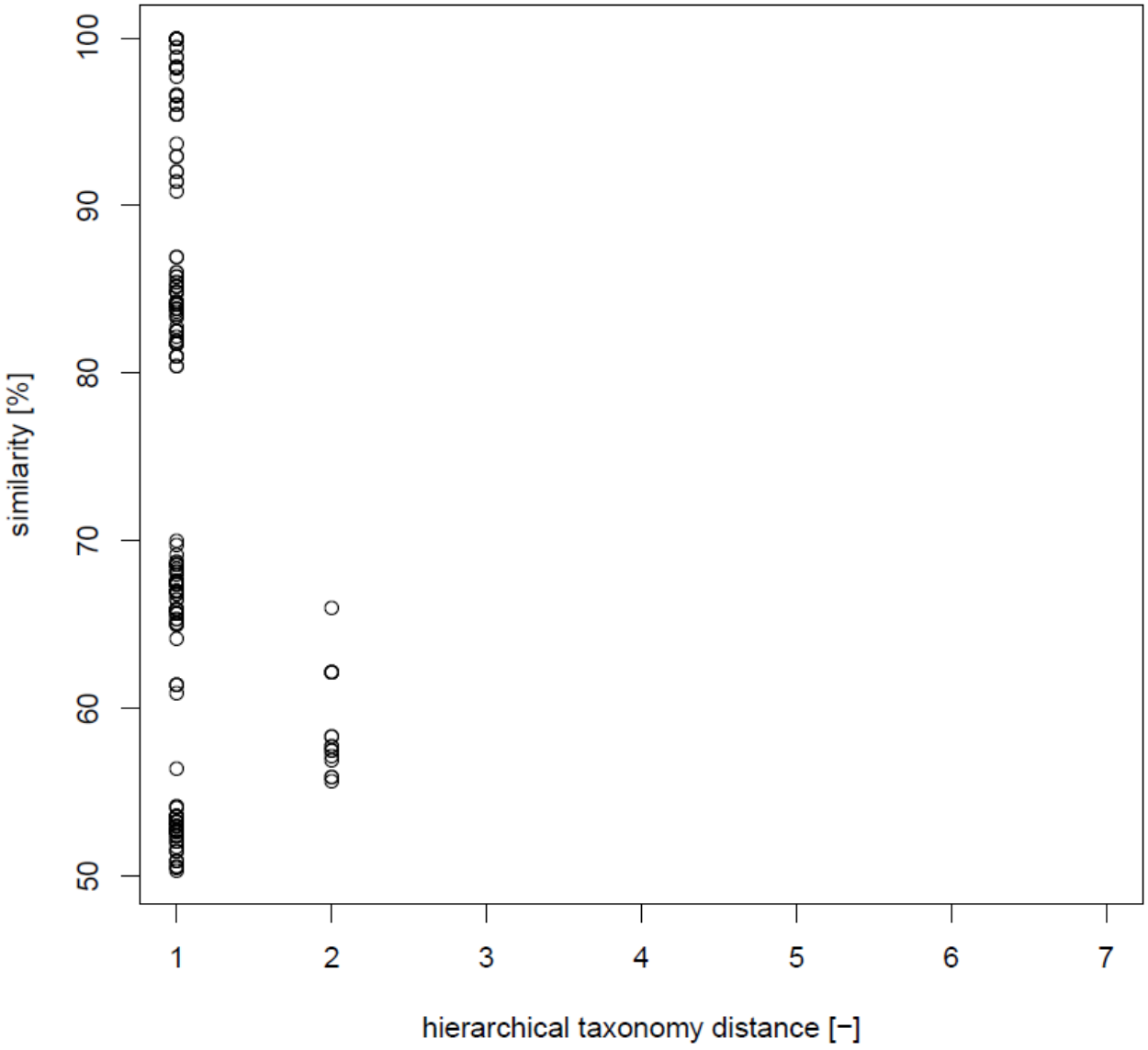

Figure 43: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1399, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

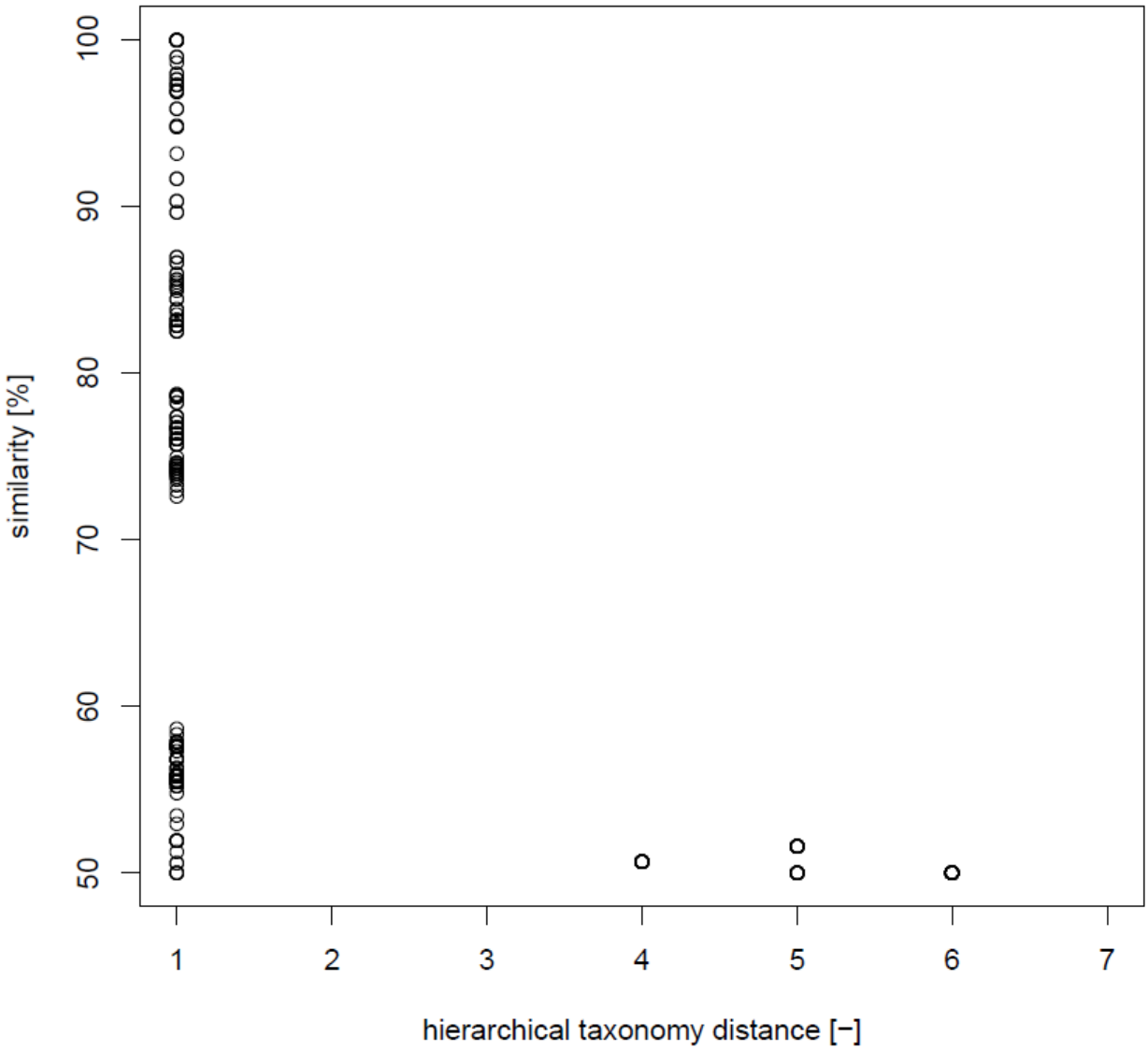

Figure 44: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1400, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

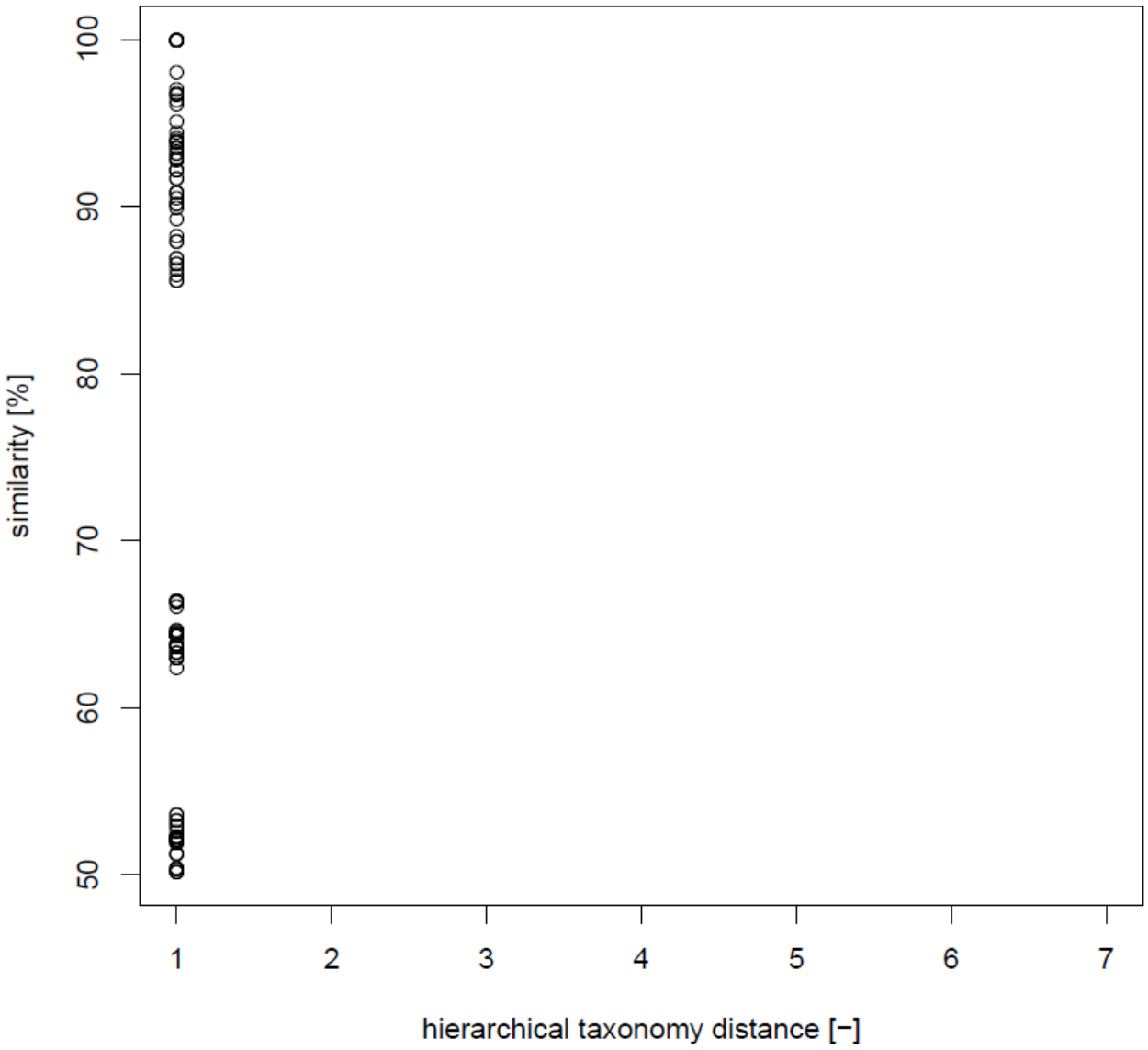

Figure 45: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1401, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

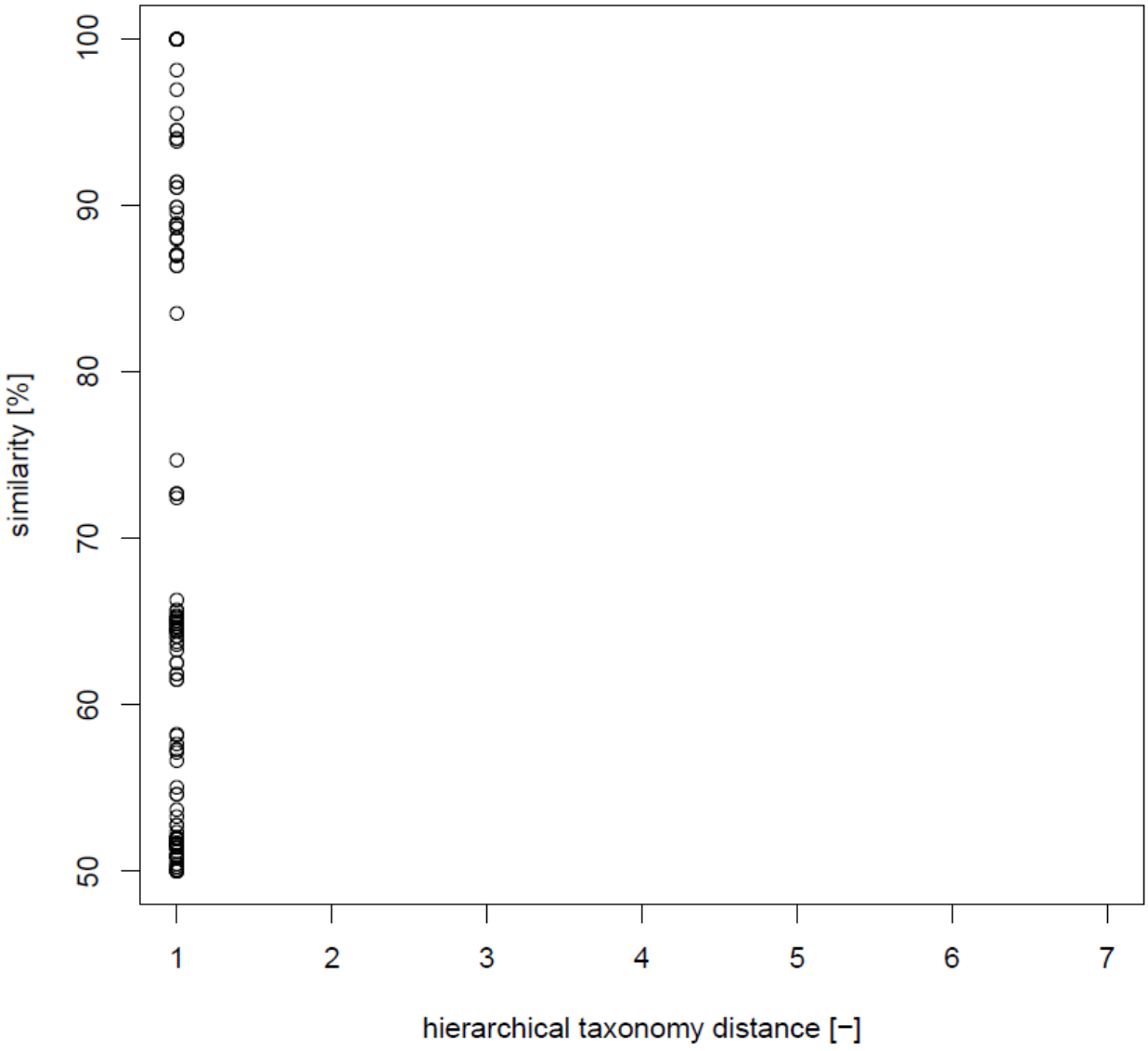

Figure 46: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1422, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

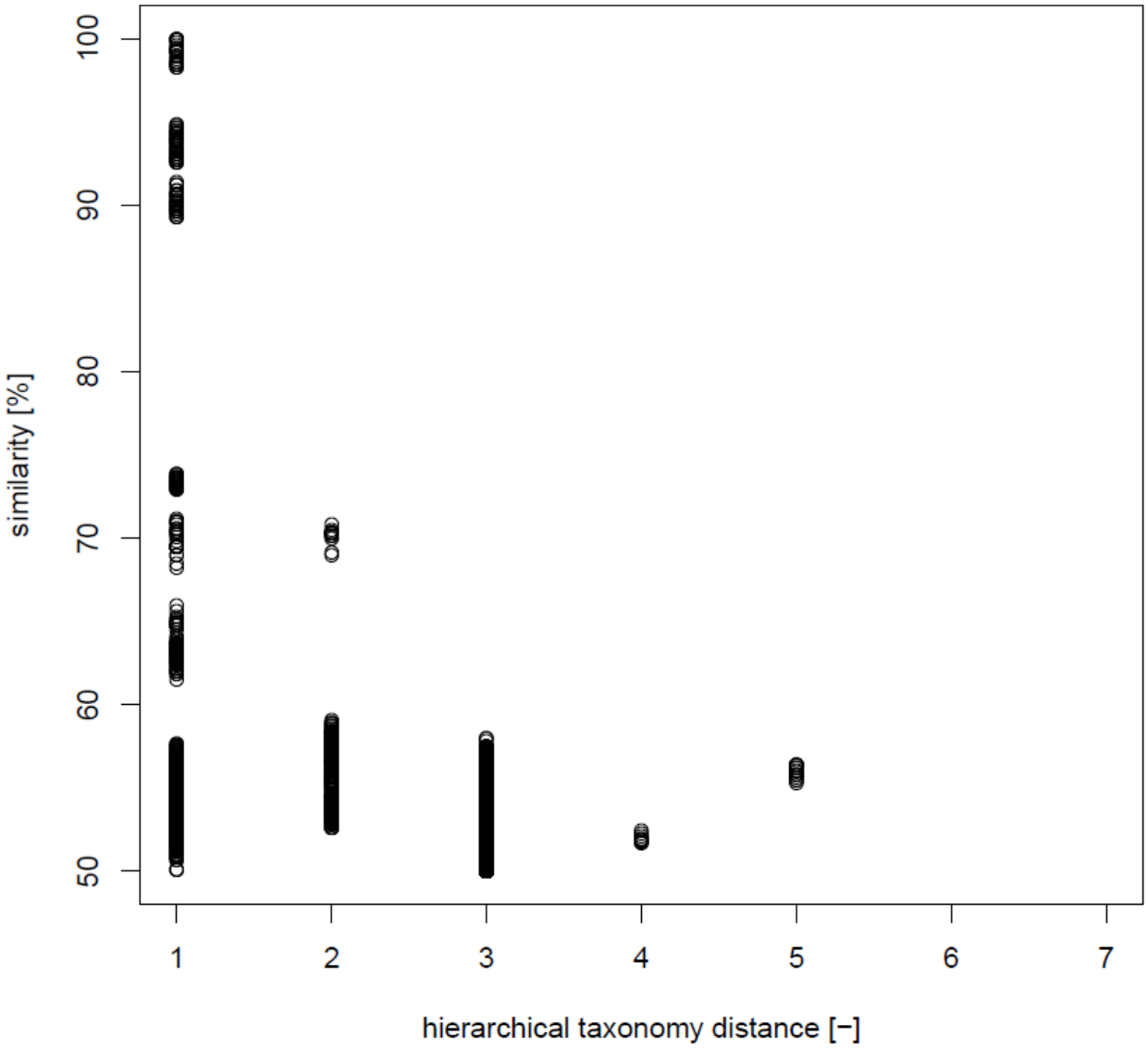

Figure 47: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1428, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

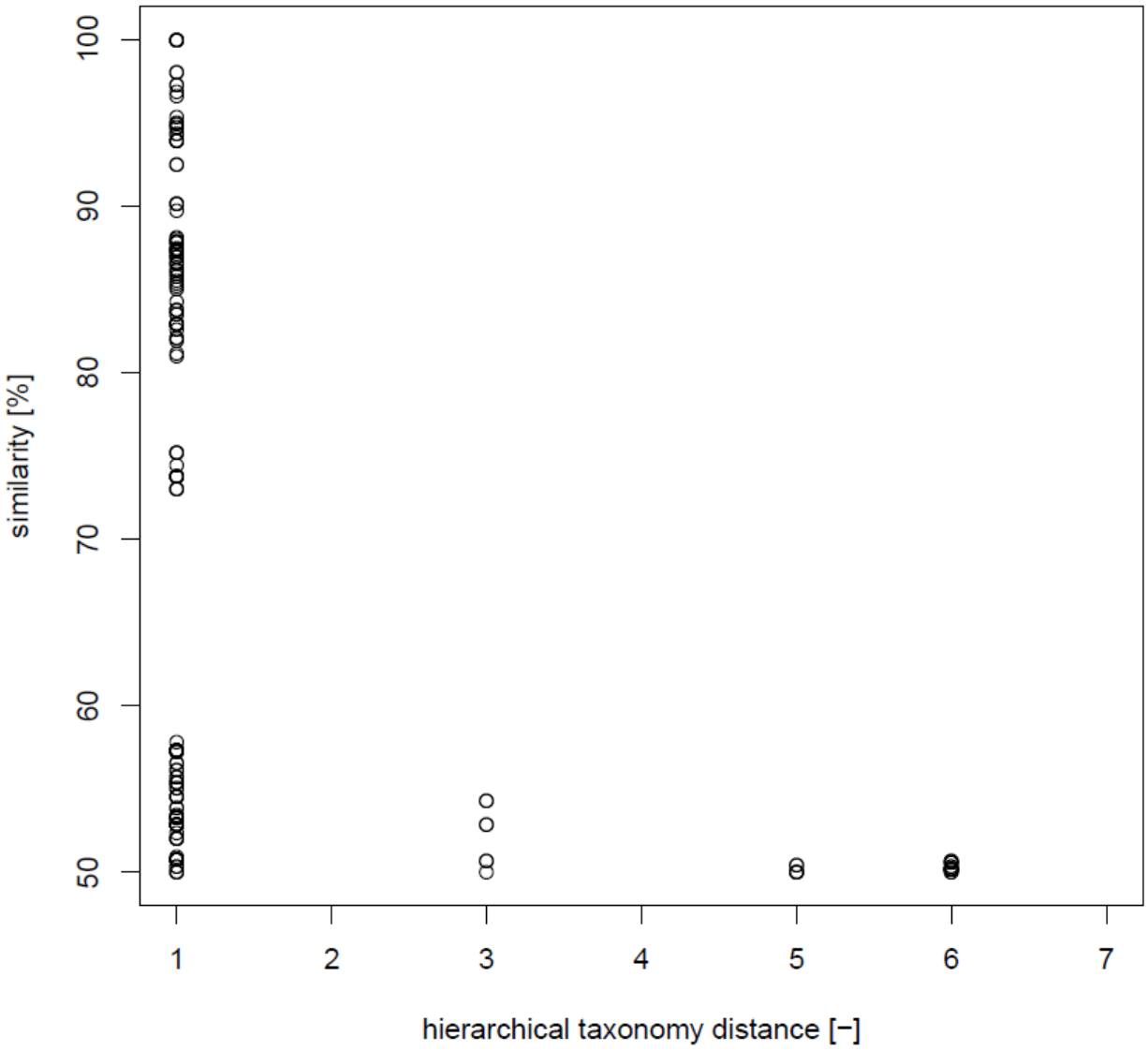

Figure 48: ***Scatter plot of sequences similarity in function of the taxonomical distance for the gene family 1458, specific to honey bees.*** *The different distance levels are: 1. Lactobacillus, 2. Lactobacillaceae, 3. Lactobacillales, 4. Bacilli, 5. Firmicutes, 6. Bacteria, 7. None (Archae, Eukaryota, contaminations).*

S2: putative HGT phylogeny inference

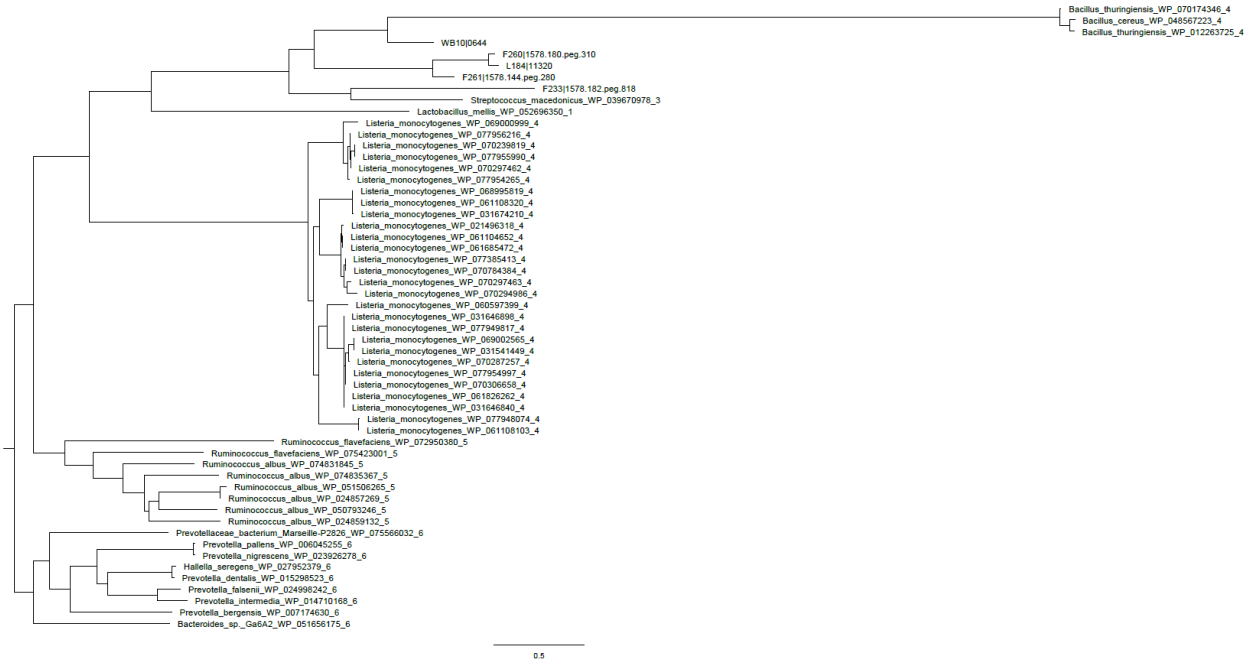

Figure 49: ***Cladogram of gene family 991 specific to bumble and honey bees, inferred by WAG model.*** *50 random BLAST hits and 5 reference genomes were used for this analysis. Each random BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

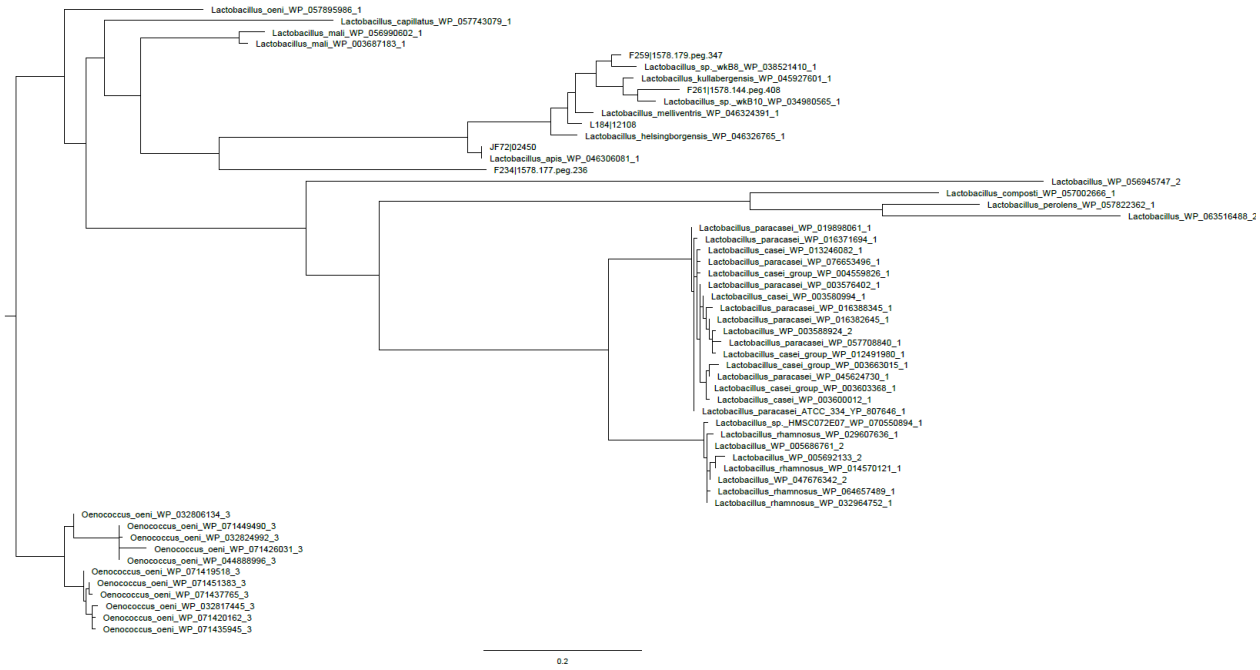

Figure 50: ***Cladogram of gene family 1099 specific to bumble and honey bees, inferred by JTT model.*** *50 random BLAST hits and 5 reference genomes were used for this analysis. Each random BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

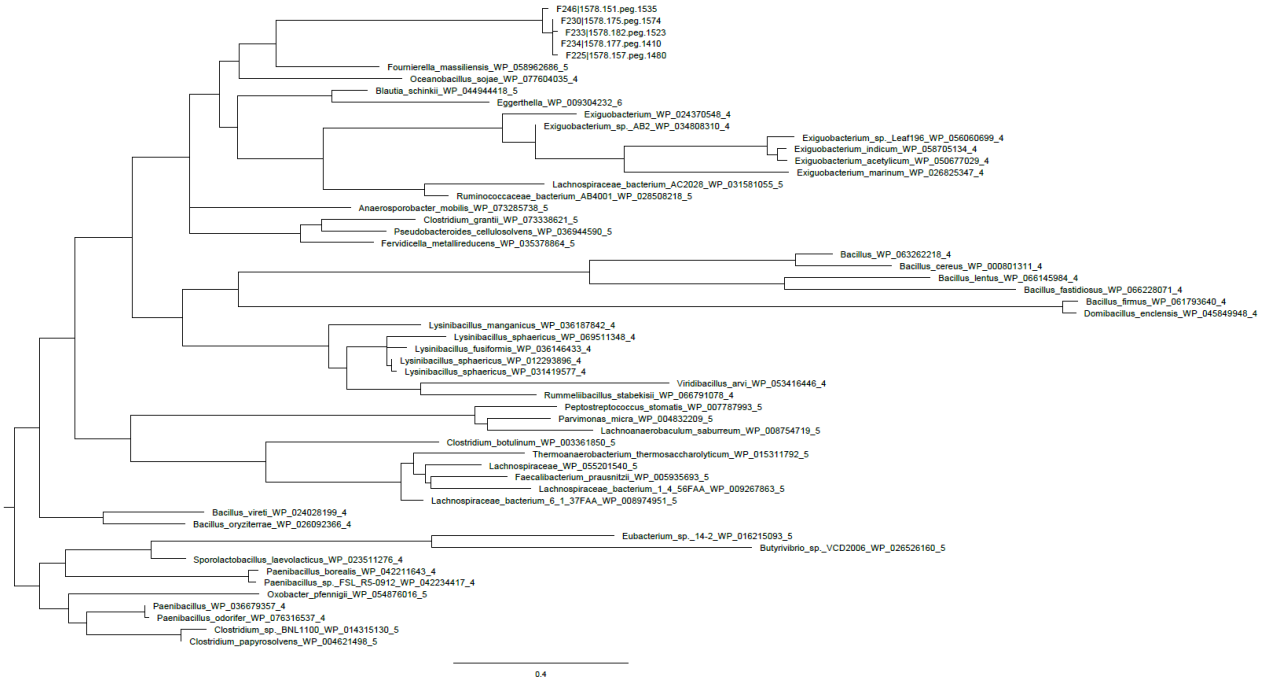

Figure 51: ***Cladogram of gene family 1674 specific to bumble bees, inferred by LG model.*** *50 random BLAST hits and 5 reference genomes were used for this analysis. Each random BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

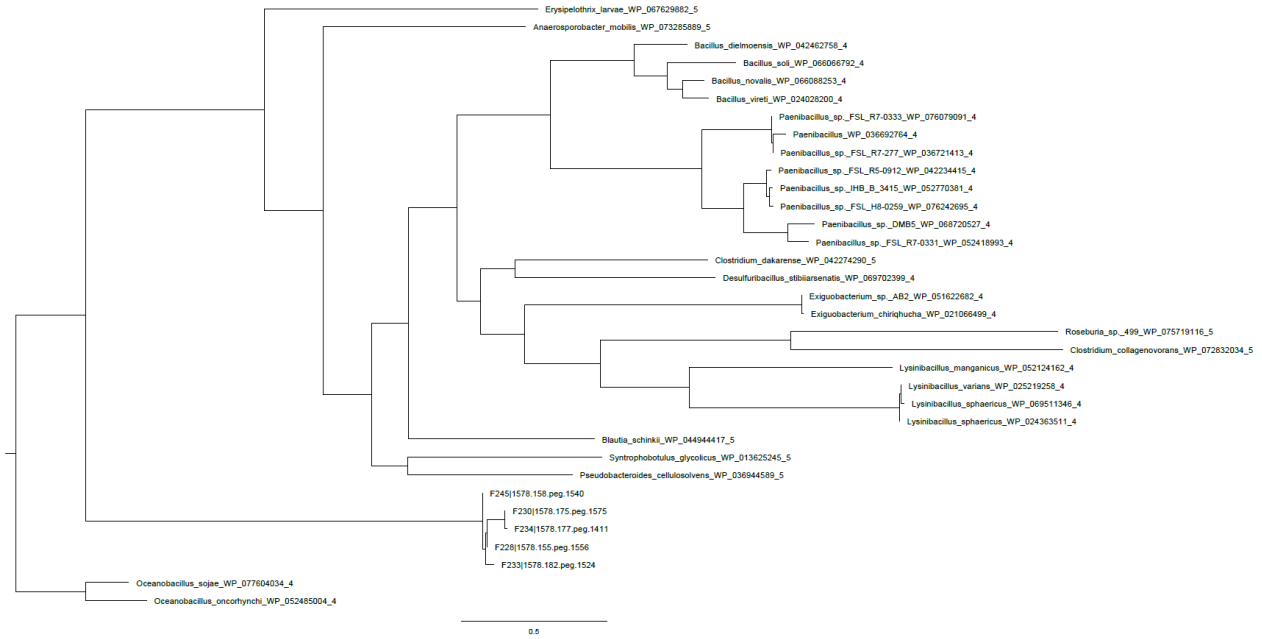

Figure 52: ***Cladogram of gene family 1675 specific to bumble bees, inferred by LG model.*** *29 BLAST hits and 5 reference genomes were used for this analysis. Each BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

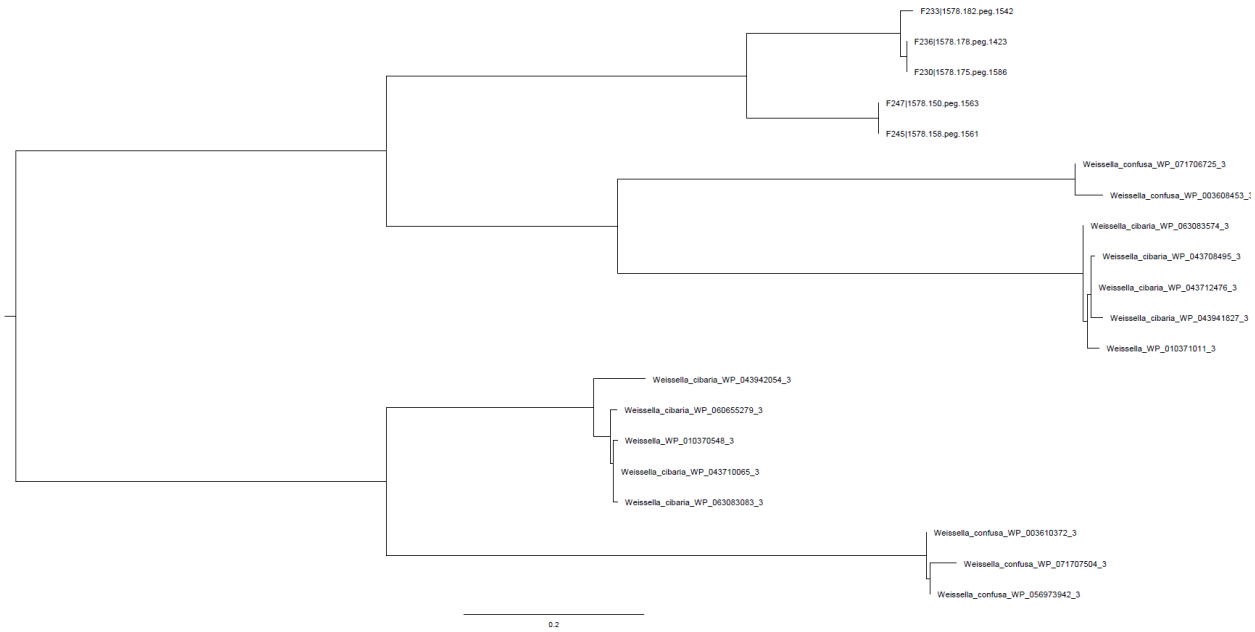

Figure 53: ***Cladogram of gene family 1678 specific to bumble bees, inferred by LG model.*** *15 BLAST hits and 5 reference genomes were used for this analysis. Each BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

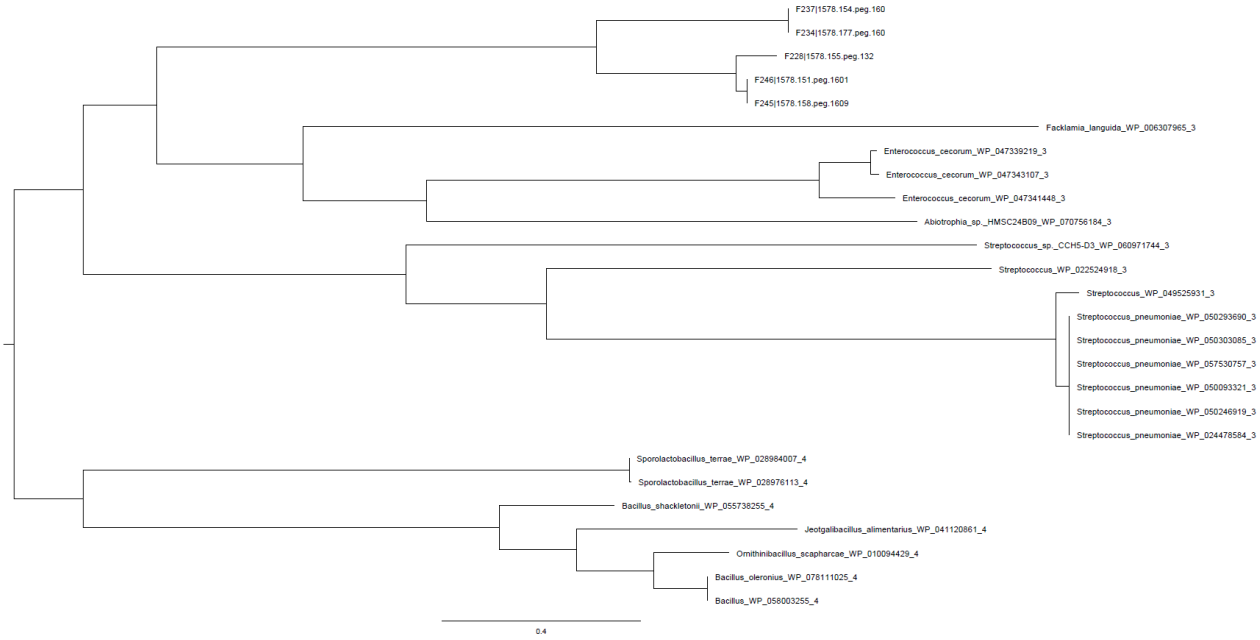

Figure 4: ***Cladogram of gene family 1757 specific to bumble bees, inferred by LG model.*** *21 BLAST hits and 5 reference genomes were used for this analysis. Each BLAST hits are presented by the name of the species, followed by the protein identifier and the hierarchical taxonomy distance. Reference genomes contain only the strain and the protein identifier.*

S3: map gene family into genomic island

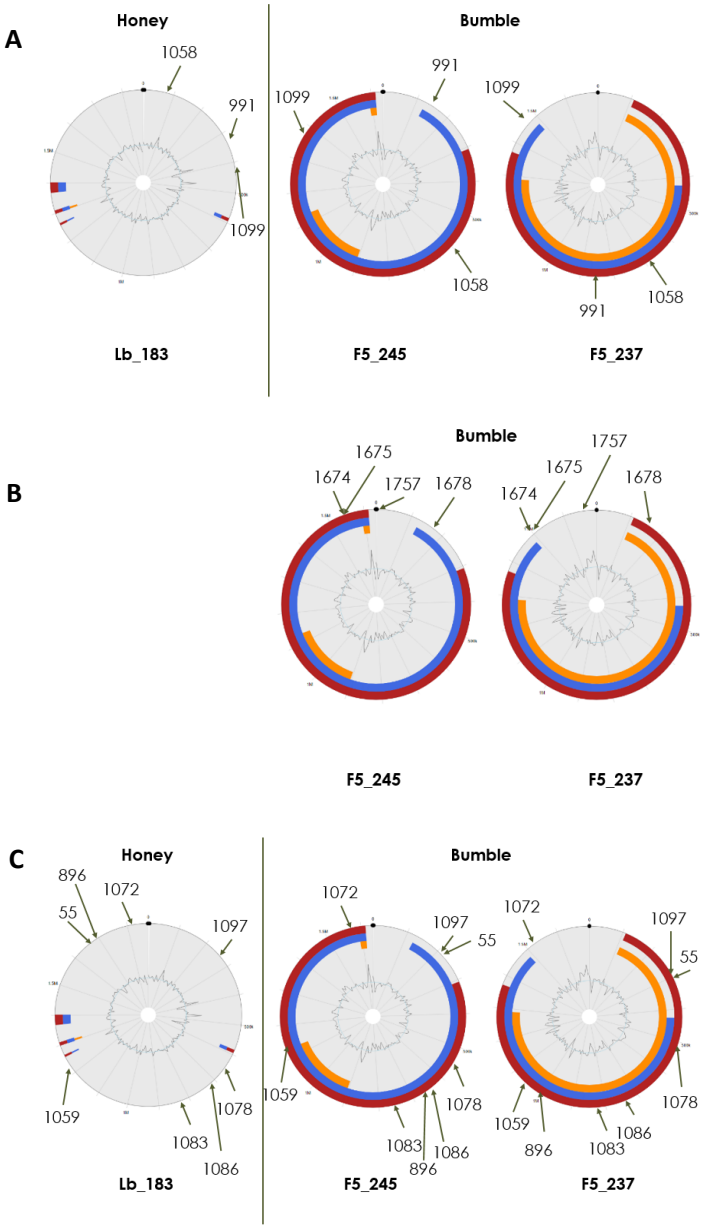

Figure 55: ***Putative gene family acquired by HGT map into reference genomes.*** *Genomic islands are predicted using IslandViewer. Blue line corresponds to the prediction method based on abnormal sequence composition or genes that are functionally related to mobile elements. Orange line corresponds to the prediction method based on codon usage bias. Red line summaries both blue and orange prediction methods. The representative species is indicated above each reference genome, while the strain is indicated below. Gene families are indicated by an arrow showing the ortholog position onto its reference genome. A: HGT specific to Firm-5, B: gene families specific to bumble and C: genes families specific to Firm-5 group.*