Creation Research Society Quarterly 2017 54:27–37.

eKINDS Project Paper

Baraminology Classification Based on Gene Content Similarity Measurement

Jean O'Micks*

Abstract

A recent genomics-based baraminology method has been developed that measures the gene content similarity (the Jaccard Coefficient Value, or JCV) between species and assigns them to individual baramins. The method is based on the creationist assumption that genes are conserved across genomes within a baramin and represent orthological functional units. Species from the same baramin should contain many common genes and thus have a high JCV, whereas species from different baramins should have a low JCV.

This method has been further developed and estimates baramins based also on k-means clustering. The method also calculates two parameters, the pan-genome quotient (PGQ) and the completeness index (CI), both of which describe how much genome erosion via gene loss has occurred in the pan-genome of the archebaramin since the Fall. The PGQ measures the intersect/union of all genes in all species in a given baramin, while the CI measures the number of genes in all species in the baramin divided by the number of species in the baramin times the size of the union of orthologous genes.

This method has been heretofore used in the analysis of Nucleocytoplasmic large DNA viruses (NCLDVs, which bear remarkable similarities to bacteria), Archaea, and insects. The method is applied to a data set of 26 fungal species in the present paper. The algorithm predicted three putative baramins, with seven species from Pezizomycotina, three from Agaro/Ustilagomycotina, and 15 from Saccharomycotina.

Based on previous experience, there is no single JCV cutoff by which species can be assigned into the same or different baramins. For example, bacterial baramins may have a rather low mean JCV due to horizontal gene transfer (HGT). In general, gene content baraminology studies depend on the biology of the organisms under study. With more and more protein data becoming available, the JCV method appears to be a promising tool for many future baraminology studies.

* Jean O'Micks, jeanomicks@gmail.com
Accepted for publication
September 8, 2017

Introduction

In recent years, the evolutionary community has moved away from inferring species relationships by using phylogenetic trees, which are based on single-gene families, because many times they produce conflicting results (Teichmann and Michison, 1999). Instead, so-called phylogenomic approaches are being taken into consideration to construct species relationships based on whole genome data, which average out the differences between individual genes. These include superalignment, superdistance, and supertree, as well as gene-content methods. Of these, this paper will focus more on the gene-content approach.

Until recently, not many genetic baraminology methods have been developed. Wood (2003) and Shan (2009) have proposed speculative models involving transposable elements and genomic rearrangements. Wood (2013) also performed a study that showed that genetic diversity in ancient DNA falls within modern sequence diversity in the horse, dog, and cat kinds, illustrating the continuity between older and newer species of a given holobaramin. More recently, a whole-genome content comparison method has been developed, which calculates the Jaccard Coefficient Value (JCV) for gene content between all possible species pairs within a set of species (O'Micks and Lightner, 2017).

From a creation perspective, genes code for proteins, which represent functional units in an organism, such as enzymes, transcription factors, structural proteins, or ion channels. These functional units are conserved in species across all life and resist evolutionary turnover (Cserhati, 2007). Whereas gene-content comparisons may capture the species relationships between single-celled organisms well, the relationship between multicellular organisms might not be so clear-cut. This is because, first, multiple splice variants multiply the number of gene variants and hence orthologs that a given species may have. Second, since the majority of eukaryotic species are multicellular, genetic interactions between different tissues and cell types further complicate the picture.

Species that belong to the same holobaramin arguably have similar morphological, biochemical, and genetic features. Therefore, similar species should have many genes in common on a genomic level. These genes of similar function belong to the same gene cluster, or orthology group. Therefore, the first main step in determining the genetic relationship between species is to classify individual genes according to their corresponding orthology groups. For this, many gene/protein clustering tools and databases are available. The second major step is to calculate the distances between species in the group under study based on their orthology content. The third and last step involves clustering species together based on mutual similarity between species within the same holobaramin and dissimilarity between all other species in the data set.

This paper will describe the process of determining gene orthology and calculating similarity and distance values between species based on gene content, as well as clustering species into baramins based on gene-content similarity. Then the JCV algorithm will be presented and shown how it applies to several test cases.

Gene-content Comparison Methodology

Defining gene clusters/groups

Genes with similar function are called homologs, which includes several categories. Orthologs are genes that functionally correspond to each other between two species, whereas paralogs are genes that functionally correspond to each other within a species. Paralogs could have formed via gene duplication from existing genes, otherwise multiple gene copies may have been created as redundant functional elements, such as the hemoglobin cluster genes (Liu and Doran, 2006). Resolving the proper relationship between orthologs is made difficult by the presence of paralogs, since we then have to figure out which of multiple paralogs a given ortholog matches to. Between species, xenologs are orthologous genes that are a result of horizontal gene transfer (HGT). Other issues that affect correct categorization of genes to gene clusters include gene loss/absence (Daubin et al., 2003), gene order (Vishnoi et al., 2010), and e-value cutoffs, which designate whether two genes are orthologous or not (Rosenfeld et al., 2016).

In some cases, orthology clusters have already been determined by previous researchers. In such cases, orthology group definition is not necessary.

A common step in defining clusters of orthologous genes is by running an all-versus-all blast of all the genes/proteins in the genome among all subject species. The BLAST algorithm sequentially compares a query sequence either with a single target sequence or a whole database full of target sequences and measures how good the alignment is between the query and the target(s). Some widely used gene cluster databases are the COG (Cluster of Orthologous Genes) database (Tatusov et al., 2003) and the eggNOG database (Powel et al., 2014). Algorithms that define gene clusters include OrthoMCL, InParanoid (Sonnhammer and Östlund, 2015), and OMA (Roth et al., 2008). Tables 1 and 2 list some of the best-known orthology algorithms and databases.

An algorithm that does all-versus-all blasts is the COGnitor program (NCBI, 2016). It was devised based on the methodology used to build the COG database (Tatusov et al., 2001). Here reciprocal best BLAST hits between two different genomes are identified as a pair of orthologs. Clusters of orthologous genes (COGs) are formed by adding ortholog pairs from at least three species (ortholog

Table 1. Several well-known gene orthology databases.

Database name	Data	Reference
COG & KOG	Clusters of orthologous groups of proteins for prokaryotes and eukaryotes	Tatusov et al., 2003
eggNOG (version 4)	Non-supervised orthologous group data for 3686 organisms	Powell et al., 2014
HomoloGene	An automated system for constructing putative homology groups from the complete gene sets of a wide range of eukaryotic species	NCBI, 2016
Inparanoid (version 8)	Ortholog groups for 273 proteomes, inferred by Inparanoid algorithm	Sonnhammer and Östlund, 2015
MBGD	Comprehensive platform for creating orthologous groups across multiple genomes	Uchiyama, 2007
OrthoDB	Hierarchical catalogs of orthologous gene groups in animals, fungi, and bacteria	Waterhouse et al., 2013
Ortholog Ontology	Ortholog database that integrates numerous types of genome and biological data	Chiba et al., 2015
OrthoMCL-DB	Stores information on orthologous groups derived by OrthoMCL algorithm	Fischer et al., 2011

triangles). The Inparanoid algorithm (Remm et al., 2001) makes provisions for removing proteins from a pairwise species comparison, which are paralogs, which are determined by giving reciprocal best within-species BLAST hits to their corresponding orthologs. The EGO algorithm extends this to multiple species comparisons (Lee et al., 2002). The OrthoMCL algorithm (Li et al., 2003), used in the method presented in this paper, also determines orthologs between species and paralogs within species and then assigns each protein sequence pair a weight, according to the $-\log_{10}$(p-value) of the BLAST hit. These weights are then represented in a symmetrical weight matrix and converted into a graph. The weights also correspond to the probability of transitioning from one protein to the next one in the MCL (Markov cluster) algorithm (Enright et al., 2002). During the application of this algorithm, the matrix is multiplied until there is little or no change in it, whereby the final matrix represents the desired protein clusters.

Different numbers of clusters of different sizes can form based on the e-value cutoff that is applied during the blast. If the cutoff is too high, then few clusters are formed, each with many members. On the other hand, if the cutoff is too low, then too many clusters form with too few members. Rosenfeld et al. (2016) found that the optimal e-value for blasts is between 1^{-50} and 1^{-100}. Otherwise, reciprocal best hits between the proteomes of different species helps determine clusters.

Another useful tool in determining orthology of a given species is comparing each protein sequence to already existing orthology databases, such as the COG database. A well-known algorithm that does this is the Inparanoid algorithm (Ostlund et al., 2010). The eggNOG database contains HMM profiles for protein families from 106 different taxonomical categories, from bacteria to insects to birds. Therefore, instead of running a BLAST search, an HMM search could be run to identify which cluster (orthology group) a given protein belongs to.

Table 2. Several well-known orthology algorithms and programs.

Algorithm	Description	Reference
COGnitor	Program that assigns protein sequence to a COG using BLASTP	Natale et al., 2000
InParanoid	Graph-based clustering of all-versus-all BLAST comparisons	Ostlund et al., 2010
OrthoMCL	Groups genes together based on sequence similarity	Li et al., 2013
OMA	Derives orthology groups based on bidirectional best hits	Roth et al., 2008

Defining baraminic distances based on gene content

When the orthology classification has been determined for each gene within a species' genome for a number of species, this data can be represented by a presence/absence matrix **M**, where $M_{i,j}$ represents the presence (1) or absence (0) of orthology group j in species i. The matrix is therefore made up of rows of zeroes and ones. Weights can be assigned to different orthology groups according to their importance within the genome. Several comparison measures have been devised that measure the difference/similarity between any given pair of species. For example, Snipen and Ussery (2010) devised the Manhattan distance, where the distance between species i and k is:

$$D_{i,k} = \left({}^{1}/_{W}\right) \sum_{j=1}^{n} w_j \left| M_{i,j} - M_{k,j} \right|$$

where n is the number of orthology groups identified between both species, and w_j is the weight of orthology group j for all groups 1 to n. Distance values ($D_{i,k}$) range from 0 to 1.

Another comparison measure is the Jaccard coefficient method, which, in contrast with the Manhattan distance, calculates the gene-content similarity between two species. The Jaccard coefficient value, or JCV is calculated by the ratio of common orthology groups divided by the total number of orthology groups present in species A and B:

$$JCV = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Three issues arise when using both of these methods. The first is whether to take mutual gene absence into account (where $M_{i,j} = M_{i,j} = 0$ for species i and k) because this might skew the result. Is it really meaningful to assign value to something that doesn't exist? In a sense, by taking mutual gene absence into account, we are presupposing what the membership of the holobaramin is going to be. This is because the presence/absence matrix defines the working set of genes for a given holobaramin, for which we wish to calculate species distance or similarity. In turn, we can define the holobaramin only by calculating these distance/similarity values between species. This problem is a tautology.

The second issue in determining species distance via the Manhattan distance measure is assigning weights to the individual orthology groups. This is perfectly valid but requires *a priori* knowledge of which groups are more or less important, which could be difficult due to the fact that the genome is made up of thousands of genes, many of which might not have been annotated yet.

A third issue is two or more genes corresponding to the same orthology group in one species as compared to only one copy in another species. This might occur if multiple paralogs in one species maps to a single gene in another species. When defining the presence/absence matrix, this means we would have to decide whether to add extra columns for the surplus orthology groups.

Other algorithms take gene order into consideration. This way, hypothetically, two species or strains with identical gene content may have a less than 100% similarity value, in that during speciation some of their genes may have been rearranged in their genome (i.e., chromosome segment inversion).

It would also be useful to measure the degree of gene loss within the holobaramin. The pan-genome (PG) describes the complete set of genes across all the members of a holobaramin. The pan-genome would theoretically represent the genome of the archebaramin before any gene loss subsequent to the Fall. Within the pan-genome, core genes are those genes that occur in the genome of every single species. Shell genes are genes that occur in the genomes of the majority of the species, whereas cloud genes are genes that are specific to only the genomes of a small number of species and take part in physiological processes that are specific to certain species—for example ones that produce special secondary metabolites.

Two related parameters can be calculated that measure how intact a genome is after gene loss. In other words, it also measures the degree of erosion the pan-genome has gone through after the Fall. One is the pan-genome quotient (PGQ), which is equal to the ratio of the number of core genes in the pan-genome to the total number of genes in the pan-genome:

$$PGQ = \frac{\bigcap_{i=1}^{n} G_i}{\bigcup_{i=1}^{n} G_i}$$

where n is the number of species in a given holobaramin, and G_i represents the set of genes in species i. The PGQ value can range from 0 to 1. A value of 0 means the pan-genome has completely eroded and no core genes exist in the holobaramin. A PGQ of 1 means all genes are intact and no genes have eroded from the pan-genome.

The completeness index (CI) also measures pan-genome erosion:

$$CI = \frac{\sum_{i=1}^{n} G_i}{n \cdot \bigcup_{i=1}^{n} G_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{G_i}{\bigcup_{i=1}^{n} G_i}$$

where n is the number of species in the holobaramin, and G_i represents the set of genes in species i. The numerator of the CI sums the total number genes in all species that are members of the holobaramin. The denominator is the number of species times the size of the pan-genome. Since we assume all genomes of a given holobaramin are derived from the archebaramin and the genome of the archebaramin represents the intact pan-genome before the Fall, before any gene loss had occurred, the CI measures the average genome intact-

ness within a given holobaramin. Similar to the PGQ, the CI can take a value between 0 and 1, where 0 corresponds to a pan-genome that has completely eroded with no core genes left, and 1 corresponds to a completely intact genome with no genes having been eroded from the pan-genome.

We should expect that if we already have assigned several species to a given holobaramin, the PGQ and the CI should not decrease too drastically with the addition of any new species. If it does, then it may be a sign that that species is not a member of the holobaramin.

As a result of this step, we should get a square matrix of distance/similarity values for all species pairs, from which we calculate the membership of different baramins.

Clustering species into baramins based on gene similarity

The JCV method described in this paper uses the default clustering algorithm (the "complete" method) used by the heat map function in R for defining and displaying the members of different baramins in a given data set as well as k-means clustering. Other clustering methods in R include partitioning methods, such as k-means clustering, and hierarchical agglomerative methods, such as Ward's method.

Molecular Genetic Baraminology Studies Using the JCV Method

Methodology

The JCV method (O'Micks, 2016; Yaugh, 2016; O'Micks and Lightner, 2017) implements molecular genetics-based baraminology classification in the following steps. First, the researcher downloads the complete proteome for each species under study. Each individual species has all of its protein sequences in a multifasta file. Protein sequences may be downloaded from the NCBI database; however, these proteomes may not be complete. The UniProt database (http://www.uniprot.org/proteomes/) contains proteomes with high-quality sequences for species with completely sequenced genomes but does not have information on as many species as the NCBI database.

Attention must also be paid to whether it is possible to include noncoding (that is, now inactive) protein sequences in the proteome of a given species in the analysis. Leaving out such noncoding proteins from the analysis may skew the results.

Next, the OrthoMCL algorithm (Li et al., 2003; Fischer et al., 2011) is used to assign the individual protein sequences of a proteome to existing orthology clusters. The online version of orthoMCL (http://orthomcl.org/orthomcl/proteomeUpload.do) can also be used to upload these proteome files in order to complete this step. Afterward, the result of the OrthoMCL algorithm is retrieved. All orthology group IDs for all species is then combined into a large list, which is used for the JaccardClusters script.

The R script is available at https://github.com/jeanomicks/JCV. Version 2 (JaccardClusters2.R) of the script uses k-means clustering to estimate the baramins from the data (estimates may vary per run). The new version also outputs baramin membership as well as statistics for each baramin. The output of this algorithm is the similarity matrix, a .noa and .sif file, which can be used in Cytoscape for downstream analysis, as well as a heat map that displays the JCVs for all species pairs (visualizing the similarity matrix values). Lighter colored pixels correspond to higher JCVs close to 1 (similarity between two species), whereas darker colored pixels correspond to JCVs closer to 0 (dissimilarity between two species) on the heat map. Since the heat map function uses a clustering algorithm, it visualizes different clusters of species that are similar to each other based on their JCVs.

A boxplot can be drawn comparing the mean JCV and the JCV range for all species pairs within a baramin, as well as between each member of the baramin and every other species. This illustrates the defining principle of baraminology—showing continuity within a baramin and discontinuity with all other species outside of the baramin. Ideally, these two ranges of JCVs should separate well from one another. The best-case scenario is when the intrabaraminic JCVs have a narrow range with high values and the extrabaraminic JCVs also have a narrow range with low values. The discordance between intrabaraminic JCVs and extrabaraminic JCVs can be measured by a p-value, which is the result of the Student's t-test.

Test case on 26 fungal species

So far, the JCV method has been implemented in three baraminology studies in Archaea, Bacteria, and Insects, representing the three main domains of life. We also applied the algorithm to a set of 26 fungal species studied by Dutilh et al. (2007). These included 22 Ascomycota, 3 Basidiomycota and the Microsporidium *Encephalitozoon cuniculi* as an outlier. These species include single-celled yeast species as well as filamentous fungi. A list of these species can be seen in Table 3, along with the number of proteins they each have, the number of OrthoMCL orthology groups that these proteins were assigned to, and the putative baramin ID that they belong to.

Fungi are not specifically mentioned during the days of Creation; however, since we know that many fungal species form symbiosis with certain plant species, such as in mycorrhiza, we can speculate that fungi were created on Day 3 of Creation, along with plants (Loucks, 2009). One can also reason that if plants and animals were created according to their kind, then so were fungi. Since

fungi, along with plants do not have living souls, they do not classify as proper living beings. Fungi may have played a saprophytic role in the original created world in recycling nutrients.

These researchers used a COG-based method (Tatusov et al., 1997), which resulted in 6,488 unambiguous triangle-based triOGs. Each triOG lists all the proteins of all species that it belongs to. After transforming the data, we were able to supply the list of species and ortholog pairs to the JaccardClusters script. The heat map that displays the species relationships between the 26 fungal species can be seen in Figure 1. The intrabaraminic JCV range for all three baramins is visibly higher than that of the extrabaraminic JCV range (JCV between species from the given baramin and all other species; Figure 2).

It is interesting to observe how the size of the intersect and the union of orthologs change as more and more species are added together from a given baramin. After selecting an initial "seed" species for a given baramin, the inter-

Table 3. List of 26 fungal species from Dutilh et al., 2007, the number of proteins, orthology groups, and baraminic classification

Species	Proteins	Orthology groups	Classification 1 (heat map)	Classification 2 (k-means)
Ashbya gossypii	4720	2867	3	3b
Aspergillus fumigatus	9926	4038	1	1
Aspergillus nidulans	9541	3769	1	1
Candida albicans	11904	2399	unassigned	3b
Candida glabrata	5272	2856	3	3b
Cryptococcus neoformans	5882	1745	2	2
Debaryomyces hansenii	6896	2570	unassigned	3b
Encephalitozoon cuniculi	1918	412	outlier	outlier
Fusarium graminearum	11640	4461	1	1
Kluyveromyces lactis	5331	2976	3	3b
Kluyveromyces waltii	5230	2920	3	3b
Magnaporthe grisea	11109	3904	1	1
Neurospora crassa	10620	3858	1	1
Phanerochaete chrysosporium	11777	1942	2	2
Saccharomyces bayanus	4966	2737	3	3b
Saccharomyces castellii	4690	2656	3	3b
Saccharomyces cerevisiae	6702	3028	3	3b
Saccharomyces kluyveri	2992	1876	unassigned	3b
Saccharomyces kudriavzevii	3813	2026	unassigned	3b
Saccharomyces mikatae	3100	1640	unassigned	3b
Saccharomyces paradoxus	8955	2962	3	3b
Schizosaccharomyces pombe	4990	2098	unassigned	3b
Stagonospora nodorum	16597	4216	1	1
Trichoderma reesei	9997	4212	1	1
Ustilago maydis	6522	1701	2	2
Yarrowia lipolytica	6666	2392	unassigned	3b

sect of orthology groups common to all species slowly decreases asymptotically, until it reaches a plateau, as can be seen in Figure 3 (here the number on the x-axis shows the number of species already added to the seed species). This plateau shows the number of orthology groups that make up the core genome of a given baramin. A sharp decrease in the intersect can be observed after adding a species from baramin #2 to the seven species in baramin #1. Also, there is a sharp drop in the orthology group intersect, the PGQ, and the CI once a species from another baramin (as can be seen in Figure 4) is added to the working baramin (at eight species on the x-axis, since *Pezizomycotina* has seven species in its membership). This is because the set of core genes differs between two individual holobaramins, and their overlap is smaller than the set of core genes from either holobaramin.

There are at least three putative baramins represented on the heat map, as listed in Table 3. The statistics for these three groups are shown in Table 4. The first, stricter classification of the third baramin includes only *Ashbya gossypii*, *Candida glabrata*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces cerevisiae*, and *Saccharomyces paradoxus*. Seven other species are unassigned, namely *Candida albicans*, *Debaryomyces hansenii*, *Saccharomyces kluyveri*, *Saccharomyces kudriavzevii*, *Saccharomyces mikatae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. K-means clustering includes these seven species along with the previous eight species to form a baramin with 15 species. However, this way the mean JCV of the third baramin drops from 0.86 to 0.65 and the p-value is much higher at 7.5×10^{-42} as opposed to 3.1×10^{-66}, although statistically still rather significant. Thus there is no significant reason in not expanding the third group to include all 15 species. The third baramin, including only eight species, also has a higher PGQ and CI value than the same baramin including 15 species. However, since the p-value for 15 species within the third baramin is also statistically significant, we can state that the third baramin is made up of 15-member species.

This third, large-sized baramin corresponds to the subphylum *Saccharomycotina*, which includes the genera *Ashyba*, *Candida*, *Kluyveromyces*, and *Saccharomyces* and is considered to be monophyletic on evolutionary trees (James et al., 2006). The seven species in the first baramin belong to the subphylum *Pezizomycotina*. The three species in the second baramin belong to separate subphyla (*Agaricomycotina* and *Ustilagomycotina*). This number of species is too small to say whether they form a single holobaramin.

The reason that three *Saccharomyces* species, *S. kluyveri*, *S. kudriavzevii*, and *S. mikatae* show somewhat lower JCVs to the other *Saccharomyces* species is because their genomes are incomplete (Cliften et al., 2003; Scannell et al., 2006). Table 3 shows that besides the outlier species, *Encephalitozoon cuniculi*, these three species have the lowest number of proteins in their genome.

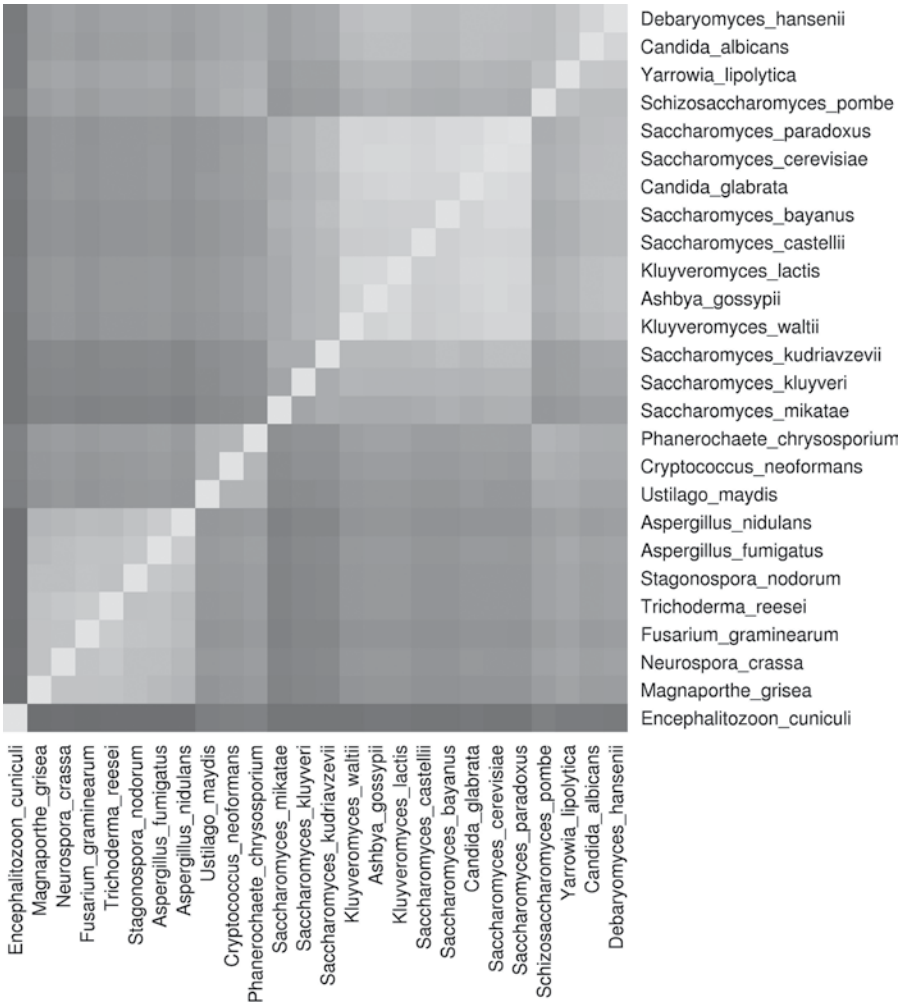

Figure 1. Heat map of JCVs for 26 species of fungi based on the Jaccard Coefficient Method. Lighter colors denote higher JCVs, closer to 1, whereas darker colors denote lower JCVs, closer to 0.

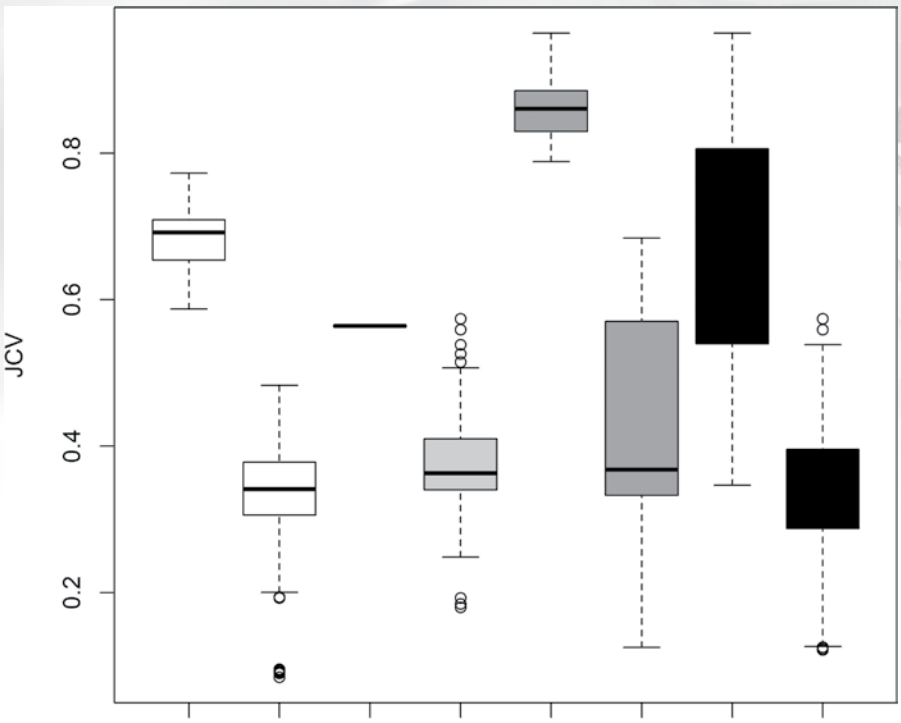

Figure 2. Box plot for four different baraminic classifications of 26 fungal species. 1: White—intrabaraminic and extrabaraminic JCVs of Pezizomycotina, 2: Light gray—intrabaraminic and extrabaraminic JCVs of Agaro/Ustilagomycotina, 3a: Dark gray—intrabaraminic and extrabaraminic JCVs of Saccharomycotina (with only 8 species), 3b: Black—intrabaraminic and extrabaraminic JCVs of Saccharomycotina (with 15 species).

As an outlier, *Encephalitozoon cuniculi* belongs to a different phylum than *Ascomycota* and *Basidiomycota*. The classification of fungi is in constant flux. Apparently, in fungi, the holobaramin may even reach the level of subphylum, as opposed to animals and plants, but further study is needed.

Discussion

This work has reviewed a recent statistical phylogenomics-based baraminology method that has been used on three data sets covering bacterial, Archaea, and eukaryotic species, and newly on a fungal data set. This method could potentially be used in more baraminology studies in a variety of different species. The method apparently works both in unicellular and multicellular organisms. However, comparisons of results from genetic and morphology-based baraminology studies must be made to validate our results and increase our confidence that our classification is correct.

One of the main arguments against using genetic data in baraminology is that it does not accurately translate to morphological characterizations (Wood and Murray, 2003). However, since the genotype determines the phenotype, it is precisely changes in the genome of the archebaramin that leads to speciation within a baramin. In other words, mutations in the genome of the archebaramin lead to speciation during the life history of the holobaramin. Furthermore, with quickly accumulating genomics and proteomics data, species sets with available proteomics data of any kind may be put together, provided that a complete data set is provided.

One challenge with the method is providing enough, high quality data. Based on a baraminology study of 107 insects (O'Micks and Lightner, 2017), if the number of proteins within a group of species varies too much, some species might not fall within their proper baramin. Another problem is if there are many short protein sequences for a given species, the ortholog algorithm might reject them (as is the case in the OrthoMCL algorithm), thus fewer orthologs are identified for a given species and its JCVs with other members of its baramin will be low. Therefore the species will not cluster well enough. In addition, different orthology algorithms might give different ortholog classifications. Furthermore, the present algorithm counts multiple members of the same orthology group only once.

An important issue is the selection of a cutoff JCV that could serve as an indicator as to whether two species belong to the same baramin. One would think, intuitively, that species within a baramin would have a high number of common genes. A commonsense speculation could be that at least half of all genes should be common between any two species within a baramin (JCV ≥ 0.5). However, as in the case of bacteria and Archaea (O'Micks, 2016; Yaugh, 2017), the average JCV within a baramin can be quite low, even as low as 0.19 as in certain *Ascoviridae*, and 0.45 in some methanogenic Archaea. This is due to the high rate of HGT in these single-cell organisms. Therefore, as of yet, no certain JCV cutoff can be set in stone to differentiate between species; rather, the cutoff depends on the biology of the given group of species under study. A good way to determine baramin membership is by monitoring the gene intersect and the PGQ and CI values, which gives us a picture of the size of the core set of genes (the

Table 4. Characteristics of the three fungal holobaramins from Dutilh et al., 2007.

Baramin	Species	Mean JCV±stdev	JCV range	PGQ	CI	p-value
Pezizomycotina	7	0.69±0.05	0.59–0.79	0.38 (2001/5295)	0.77 (28458/37065)	$5.1x10^{-27}$
Agaro/Ustilagomycotina	3	0.57±0.02	0.56–0.6	0.42 (1070/2525)	0.71 (5388/7575)	$5.2x10^{-6}$
Saccharomycotina	8	0.86±0.04	0.79–0.96	0.62 (2022/3243)	0.89 (23002/25944)	$3.1x10^{-66}$
Saccharomycotina	15	0.65±0.15	0.35–0.96	0.11 (399/3694)	0.64 (35611/55410)	$7.5x10^{-42}$

pan-genome) of a given baramin. For example, in the case of the *Pezizomycotina*, we see that the gene intersect, PGQ, and CI decreases asymptotically, which implies the presence of a core set of genes within the pan-genome of the holobaramin. This is because since these genes are present in all species in the baramin, this set of core genes cannot decrease any farther. When core gene sets from two different baramins are compared with each other, the overlap between these two gene sets is smaller than either core gene set. This is because these sets of core genes are responsible for the bauplan for two different kinds of organisms. However, after adding a seventh species from the *Agaro/Ustilagomycotina* baramin, this trend breaks. The intersect drops from 2001 genes to 1050, almost half (Figure 3). The drop in PGQ is from 0.38 to 0.2, also almost half. The drop in CI is less pronounced; from 0.77 to 0.71 (Figure 4). PGQ and CI values were also calculated for the five holobaramins that were predicted in the study on insects (O'Micks and Lightner, 2017). As we can see, similar PGQ and CI values occur for both fungi and insects.

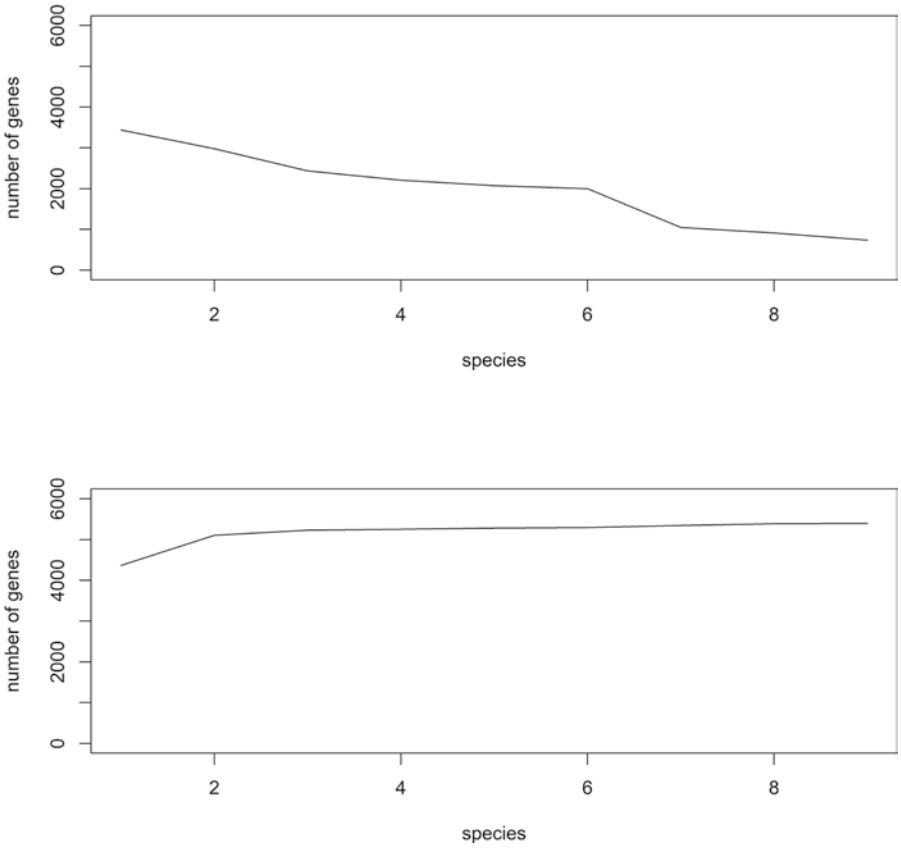

Figure 3. Changes in the number of genes in the intersect (above) and union (below) of increasing numbers of added species from the Pezizomycotina baramin, after which species from the Agaro/Ustilagomycotina baramin were added at n=6 species.

As with other baraminology methods, selecting the proper set of species is also important. If the number of species being examined is either too small, or the number of sampled holobaramins is too large, holobaramins might go undetected or might not be able to be clearly defined with too small a membership. This was the case in cluster #2 in the fungal baraminology analysis in this paper, which had only three members, or several small groups of NCLDVs with three members or less (O'Micks, 2016). In such cases, if a holobaramin is represented by only a small number of species (two or less), then it is impossible to know whether these two species belong to the same baramin or different ones. The addition of extra species is needed to determine this, whether the additional species all form one group, or whether they group with one or the other (initial, seed) species.

Table 5. Characteristics of the five insect holobaramins from O'Micks and Lightner, 2017.

Baramin	Species	Mean JCV ± stdev	JCV range	PGQ	CI	p-value
Diptera I (mosquitos)	6	0.74±0.05	0.66-0.84	0.48 (6030/12597)	0.76 (57388/75582)	$1.5x10^{-8}$
Diptera II (flies)	33	0.89±0.06	0.69-0.98	0.22 (2725/12564)	0.7 289797/414612)	0.0
Auchenorrhyncha	7	0.76±0.04	0.71-0.84	0.5 (5627/11366)	0.72 (57127/79562)	$9.8x10^{-12}$
Hymenoptera	42	0.86±0.03	0.75-0.94	0.37 (4751/12849)	0.62 (335369/539658)	0.0
Lepidopterans	10	0.71±0.16	0.45-0.9	0.48 (6030/12597)	0.76 (57388/75582)	$5.2x10^{-5}$

Materials and Methods

The JaccardClusters2.R script was further developed in R version 3.2.2. The script is available at github.com/jeanomicks/JCV. Figures 2–4 were made in R version 3.2.2. The orthology data set for 26 fungal species was downloaded from the supplementary data from Dutilh et al. (2007).

Acknowledgments

The author would like to thank Dr. Michael Price for his insight into the classification of fungal species and also Dr. Jean K. Lightner for her advice in how to draft the manuscript.

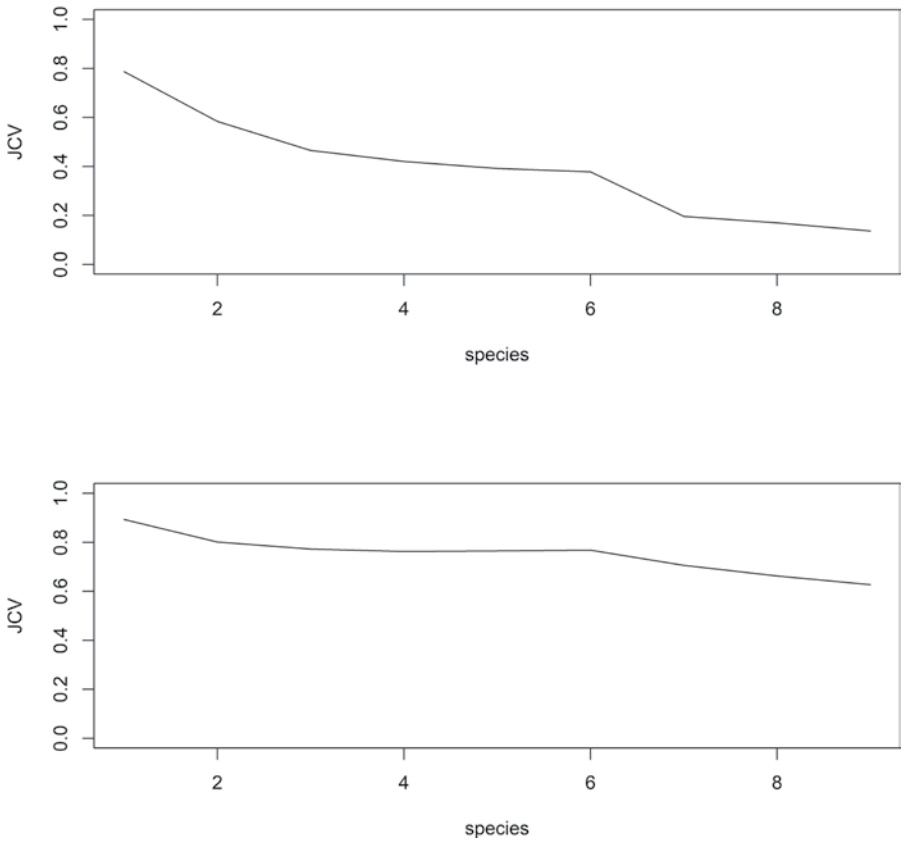

Figure 4. Changes in the PGQ (above) and CI (below) parameters of increasing numbers of added species from the Pezizomycotina baramin, after which species from the Agaro/Ustilagomycotina baramin were added at n=6 species.

References

Chiba, H., H. Nishide, and I. Uchiyama. 2015. Construction of an ortholog database using the semantic web technology for integrative analysis of genomic data. *PLoS One.* 10(4):e0122802.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, et al. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science.* 301(5629):71–76.

Cserhati, M. 2007. Creation aspects of non-coding sequences. *Journal of Creation.* 21(2):101–108.

Daubin, V., N.A. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science.* 301(5634):829–832.

Dutilh, B.E., V. van Noort, R.T. van der Heijden, T. Boekhout, B. Snel, et al. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics.* 23(7):815–824.

Enright, A.J., S. van Dongen, and C.A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30(7):1575–1584.

Fischer, S., B.P. Brunk, F. Chen, X. Gao, O.S. Harb, et al. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current Protocols*

in Bioinformatics. Chapter 6: Unit 6.12.1–19.

James, T.Y., F. Kauff, C.L. Schoch, P.B. Matheny, V. Hofstetter, et al. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*. 443(7113):818–822.

Lee, Y., R. Sultana, G. Pertea, J. Cho, S. Karamycheva, et al. 2002. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research* 12(3):493–502.

Li, L., C.J. Stoeckert Jr., and D.S. Roos. 2013. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13(9):2178–2189.

Liu, Y., and D. Moran. 2006. Do new functions arise by gene duplication? *Journal of Creation* 20(2):82–89.

Loucks, I.S. 2009. Fungi from the biblical perspective: design and purpose in the original creation. *Answers Research Journal* 2:123–131.

Natale, D.A., U.T. Shankavaram, M.Y. Galperin, Y.I. Wolf, L. Aravind, et al. 2000. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biology* 1(5):RESEARCH0009.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 44(D1):D7–19.

O'Micks, J. 2016. Baraminic analysis of nucleocytoplasmic large DNA viruses. *Journal of Creation* 31(1):73–79.

O'Micks, J., and J.K. Lightner. 2017. An initial estimate of created kinds within four insect orders (Diptera, Hemiptera, Hymenoptera, and Lepidoptera) using molecular data. *Creation Research Society Quarterly* (in press).

Ostlund, G., T. Schmitt, K. Forslund, T. Köstler, D.N. Messina, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* 38(Database issue):D196–203.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research* 42(Database issue):D231–239.

Remm, M., C.E. Storm, and E.L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* 314(5):1041–1052.

Rosenfeld, J. A., J. Foox, and R. DeSalle. 2016. Insect genome content phylogeny and functional annotation of core insect genomes. *Molecular Phylogenetics and Evolution*. 97:224–32.

Roth, A.C., G.H. Gonnet, and C. Dessimoz. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9:518.

Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*. 440(7082):341–345.

Shan, E.L. 2009. Transposon amplification in rapid intrabaraminic diversification. *Journal of Creation*, 23(2):110–117.

Snipen, L., and D.W. Ussery. 2010. Standard operating procedure for computing pangenome trees. *Standards in Genomic Science* 2(1):135–141.

Sonnhammer, E.L., and G. Östlund. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* 43(Database issue):D234–239.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* 278(5338):631–637.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29: 22–28.

Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Teichmann, S.A., and G. Mitchison. 1999. Is there a phylogenetic signal in prokaryote proteins? *Journal of Molecular Evolution* 49(1):98–107.

Uchiyama I. 2007. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Research* 35(Database issue):D343–346.

Vishnoi, A., R. Roy, H.K. Prasad, and A. Bhattacharya. 2010. Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationship among closely related microorganisms. *PLoS One*. 5(11):e14159.

Waterhouse, R.M., F. Tegenfeldt, J. Li, E.M. Zdobnov, and E.V. Kriventseva. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Ressearch* 41(Database issue):D358–365.

Wood, T.C. 2003a. Perspectives on AGEing, a young-earth creation diversification model. In Ivey, R.L (editor), *Proceedings of the Fifth International Conference on Creationism*, pp. 479–489. Creation Science Fellowship, Pittsburgh, PA.

Wood, T.C., and M.J. Murray 2003b. *Understanding the Pattern of Life: Origins and Organization of the Species*. Broadman & Holman, Nashville, TN.

Wood, T.C. 2013. Mitochondrial DNA analysis of three terrestrial mammal baramins (Equidae, Felidae, and Canidae) implies an accelerated mutation rate near the time of the Flood. In Horstmeyer, M. (editor), *Proceedings of the Seventh International Conference on Creationism*. Creation Science Fellowship, Pittsburgh, PA.

Yaugh, A. 2017. Baraminological analysis of a set of archaea species based on genomic data. *Creation Research Society Quarterly* 53:140–154.