A new baraminology method based on Whole Genome K-mer Signature analysis and its application to insect classification

Matthew Cserhati

A newly developed bioinformatics method called the Whole Genome K-mer Signature (WGKS) algorithm has been designed and used to analyze the whole genome sequences of 61 insect species from the genera *Aedes, Anopheles, Culex, Drosophila,* and *Glossina*. The results of this analysis have been evaluated from a baraminological viewpoint. The results have also been compared to clustering of the same genera based on whole mitochondrial genome sequence similarity and an analysis of whole proteomes by the Gene Content Method (GCM). All three analyses show that *Drosophila* (fruit flies) and *Glossina* (tsetse flies) are well-defined baramins, but the case with the mosquitoes *Aedes, Anopheles*, and *Culex* is more nuanced. The older GCM algorithm clumps all three mosquito genera into one baramin, whereas both the WGKS algorithm and the mitochondrial DNA analyses show that *Anopheles* forms its own baramin, separate from *Aedes+Culex*. The newly developed algorithm is more accurate, since it takes whole genome information into consideration, as opposed to merely the coding regions. With this new algorithm, more precise genetics-based baraminology studies can be performed by taking more genetic information into account. This new algorithm also tends to split groups, as opposed to lumping them together.

A new genetics-based bioinformatics algorithm has been developed to analyze and compare the whole genome sequences of any set of organisms.[1] This new algorithm can also be used to perform molecular baraminology studies. Compared to morphology-based baraminology algorithms, genetics-based algorithms have several advantages.

Molecular data is useful where morphological data is inconclusive.[2] Also, all heritable information is encoded in the DNA, meaning that there is much more genetic information than there is morphological data. Furthermore, biomolecules such as DNA record the life history of a species ever since its creation. Mitochondrial DNA (mtDNA) is especially useful for this, because it can recover phylogenies of closely related species.[3]

With the advent of the genomics revolution, an exponentially increasing amount of genetic data is being made available in online databases, much of which has not yet been analyzed. These sequence data include proteome sets as well as whole genomes. Whereas morphology-based data sets suffer from convergence, genetic-based algorithms do not do so as much. Morphology data sets may miss a lot of data points, especially if they were gathered from fossil specimens. Similarly, proteomics data themselves may be incomplete if one doesn't sample all the genes of an organism.[4]

In comparison to proteomics data, genome sequences are usually complete (they may still suffer from low coverage and may also have large segments of undetermined sequences). There is also an advantage to capturing information from the entire genome as opposed to just the protein-coding regions.[2]

The present algorithm is an alignment-free k-mer sequence comparison method. As such, the data are processed much faster than in alignment-based algorithms, which depend on *a priori*-defined guide trees.[5,6] The algorithm, called the Whole Genome K-mer Signature (WGKS) was first tested on 58 species from three insect genera, *Drosophila* (fruit flies), *Glossina* (tse-tse flies) and *Anopheles* (mosquitos).[1]

The mosquito genus *Anopheles* contains 485 species, 60 of which transmit the malaria vector *Plasmodium*. These species are global in their distribution and are studied mainly because of their epidemiological importance.[7] The fruit fly genus *Drosophila* includes a similarly large number of species. It is widely distributed in the northern hemisphere and is divided into five lineages.[8,9] *Drosophila melanogaster* is a well-known experimental animal due to its easy culturing, high generation time, and small body size. The genus *Glossina*, with around 20 species, is studied due to its economic and medical importance since it spreads trypanosomes; it has several subgenera.[10]

With this new baraminology algorithm we have a new tool which we can compare with existing tools currently used for baraminological analyses. This should make baraminology studies more precise.

Materials and methods

The Whole Genome K-mer Signature algorithm

The goal of the WGKS algorithm is to generate and compare the k-mer content of the genomes from all species in a given study. A k-mer is defined as a segment of DNA *k* bp long. In the Cserhati *et al*. study, *k* ranged from seven to nine (heptamers, octamers, and nonamers). K-mers of such lengths can act, for example, as transcription factor binding sites (TFBSs).[11] In baraminology we may assume that species from the same baramin will have similar genomes, since they are interrelated. Therefore, they should also have a similar k-mer content. This is because, within a baramin, individual species originating from the archebaramin (those species created during Creation Week representing a given baramin) would have undergone relatively little differential mutation after the Fall. Species from different baramins are assumed to have different genomes, even different chromosome numbers, or genome sizes. Therefore, we predict that their k-mer content should be very different. It is highly unlikely that very different organisms would have a similar k-mer content, since a high percentage of the genome does not consist of junk DNA but is functional (i.e. TFBSs, enhancers, silencers, etc.). For example, 80% of the human genome was assigned some biochemical function by the ENCODE Project in 2012, but this percentage could be even higher.[12]

An overview of the algorithm used in this study can be seen in figure 1. The algorithm is made up of three steps: 1. The generation of the WGKS for each species; 2. Comparison of WGKS between all possible species pairs and generation of similarity matrix; and 3. Visualization of similarity values on a heat map and prediction of species clusters (baramins).

Generation of WGKS

A WGKS is nothing more than a lexicographically ordered, two-column list of all possible k-mers with their corresponding score value. Since there are four DNA letters, this makes 4^k possible k-mers (i.e. $4^8 = 65{,}536$ possible octamers). A k-mer's score ($S_{k\text{-mer}}$) reflects its biological relevance. The more a given k-mer's occurrence deviates from a random distribution, the more likely it has some biological function associated with it. The random distribution is the number of occurrences of the k-mer by chance. The k-mer's score is calculated in the following way:

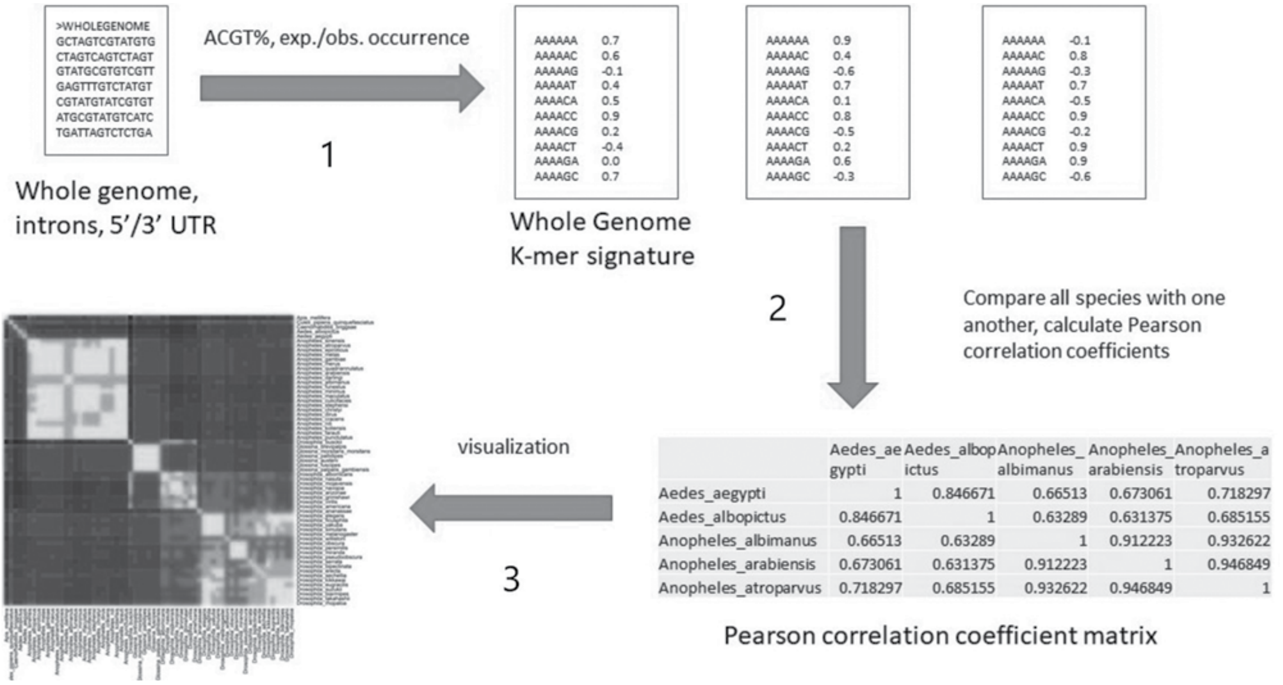
$$1.\quad S_{k\text{–}mer} = \frac{O\text{–}E}{O+E}$$

Here *O* is the number of times the k-mer is observed to occur (its actual frequency). *E* is the number of times the k-mer is expected to occur by chance. The expected occurrence *E* is:

$$2.\quad E = l_{genome} \cdot \frac{n_{1..k\text{–}1} . n_{2..k}}{n_{2..k\text{–}1}}$$

Where l_{genome} is the length of the entire genome, $n_{1..k\text{-}1}$ is the first k-1 bases of the k-mer, $n_{2..k}$ are the last k-1 bases of the k-mer, and $n_{2..k\text{-}1}$ are bases 2 to k-1 of the k-mer. The score has a value between -1 and 1.

With regards to the observed and expected



ACGT%, exp./obs. occurrence
1
Whole genome,
introns, 5'/3' UTR
Whole Genome
K-mer signature
2
Compare all species with one
another, calculate Pearson
correlation coefficients
visualization
3
Pearson correlation coefficient matrix

```
>WHOLEGENOME
GCTAGTCGTATGTG
CTAGTCAGTCTAGT
GTATGCGTGTCGTT
GAGTTTGTCTATGT
CGTATGTATCGTGT
ATGCGTATGTCATC
TGATTAGTCTCTGA
```

AAAAAA	0.7
AAAAAC	0.6
AAAAAG	-0.1
AAAAAT	0.4
AAAACA	0.5
AAAACC	0.9
AAAACG	0.2
AAAACT	-0.4
AAAAGA	0.0
AAAAGC	0.7

AAAAAA	0.9
AAAAAC	0.4
AAAAAG	-0.6
AAAAAT	0.7
AAAACA	0.1
AAAACC	0.8
AAAACG	-0.5
AAAACT	0.2
AAAAGA	0.6
AAAAGC	-0.3

AAAAAA	-0.1
AAAAAC	0.8
AAAAAG	-0.3
AAAAAT	0.7
AAAACA	-0.5
AAAACC	0.9
AAAACG	-0.2
AAAACT	0.9
AAAAGA	0.9
AAAAGC	-0.6

	Aedes_aegypti	Aedes_albopictus	Anopheles_albimanus	Anopheles_arabiensis	Anopheles_atroparvus
Aedes_aegypti	1	0.846671	0.66513	0.673061	0.718297
Aedes_albopictus	0.846671	1	0.63289	0.631375	0.685155
Anopheles_albimanus	0.66513	0.63289	1	0.912223	0.932622
Anopheles_arabiensis	0.673061	0.631375	0.912223	1	0.946849
Anopheles_atroparvus	0.718297	0.685155	0.932622	0.946849	1

Figure 1. Flow diagram for the Whole Genome K-mer Signature algorithm. The input is a genomic or sub-genomic region, such as the whole genome sequence or the 5' UTR. In the first step, the k-mer signature is derived for all species. In the second step, the k-mer signature for all species is compared to one another to get a PCC matrix. In the third step, the PCC matrix is visualized on a heat map.

occurrences of a given k-mer there are three possible trends:

3. $O \gg E : S_{k\text{–}mer} \rightarrow 1$ (over-represented k-mer)

4. $O \ll E : S_{k\text{–}mer} \rightarrow -1$ (under-represented k-mer)

5. $O = E : S_{k\text{–}mer} \rightarrow \approx 0$ (randomly occurring k-mer)

For example, a k-mer that occurs four times more frequently than expected by random chance has a score of (4-1)/(4+1) = 0.6.

The score distribution of octamers for *A. gambiae* is depicted in supplementary figure 1A. It is evident that the score values follow a bell-shaped curve. Supplementary figure 1B shows the Q-Q plot of the same values.

Comparison of WGKS between species

Once the WGKS has been calculated for all of the organisms in the study, we can compare the species on an all-versus-all basis. The WGKS can be transformed into a vector of numbers by sorting the k-mers in alphabetical order (A…A to T…T). This gives us a list of 4^k score values. Two species vectors can be compared by calculating the Pearson Correlation Coefficient (PCC) value for them. The PCC can be calculated in the following way:

6. $$PCC = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Here x and y represent a vector of k-mer scores from two different species. They are the same length, because they cover the same 4^k k-mers. A PCC has a value between -1 and 1. The more similar the WGKS between two species (same baramin), the closer the PCC value is to 1. Species from a different baramin have a lower PCC value.

Visualization on heat maps

After computing all pair-wise WGKS values, a symmetrical square matrix can be derived that contains PCC values for all possible species pairs. The PCC matrix can then be visualized with a heat map. Brighter shades correspond to PCC values closer to 1 denoting species with a similar WGKS, belonging to the same monobaramin. Darker shades correspond to PCC values with more negative values denoting species with a different WGKS, belonging to different baramins. Heat maps were created using the heat map function in R. Clustering was done using the 'average' algorithm for the WGKS method, the 'ward.D2' algorithm for GCM, whereas the 'single' algorithm was used to depict the mitochondrial data.

Sequence data

The whole genome sequences for all 61 insect species in the WGKS study were downloaded from the NCBI Genome Database (ncbi.nlm.nih.gov/genome). Proteome sets were downloaded for 44 species of *Aedes*, *Anopheles*, *Culex*, *Drosophila*, and *Glossina* from uniprot.org/proteomes. Mitochondrial genomes for 98 species of *Aedes*, *Anopheles*, *Culex*, and *Drosophila* were downloaded from the NCBI website at ncbi.nlm.nih.gov/genome/browse#!/overview. An all-versus-all BLAST comparison was performed using the ggsearch36 command line software (version 36.3.8), downloaded from faculty.virginia.edu/wrpearson/fasta/fasta36/fasta-36.3.8h.tar.gz.[13] This software provides a faster, more accurate, and more sensitive alignment of sequences than many other aligner programs.

K-mer analysis script and plots

A python script was written to calculate the 61 insect octamer (k = 8) WGKS vectors. The script is available at github.com/csmatyi/motif_analysis. All plots were made in R version 3.4.3. These include the beeswarm, ECDF and silhouette plots, using the beeswarm, cluster, fpc, NbClust libraries, and the ecdf and eclust commands. The eclust clustering command was run for three to five clusters for the results of the mitochondrial and GCM analyses. The cutree command was used to determine clusters for the WGKS method. The beeswarm plot depicts similarity values on the y-axis. The ECDF plot shows the empirical cumulative distribution function curve. In other words, this plot shows the percentage of similarity values below a given similarity value. The silhouette plot shows the silhouette width for each individual species within a given cluster (each cluster shaded by a different colour).

Supplementary files

All supplementary data files and figures are available at github.com/csmatyi/wgks.

Results

Application of the WGKS algorithm

The WGKS for octamers for 30 *Drosophila*, 22 *Anopheles*, 6 *Glossina,* and 3 outlier (2 *Aedes* and 1 *Culex*) species (61 in total) were calculated and compared with one another. The PCC values were visualized in a heat map in figure 2. The species list, PCC values, clusters, and clustering statistics are in Supplementary File 1. The Hopkins clustering statistic is 0.85, which means that the data can be clustered well. Based on the Elbow method, four optimal clusters were predicted from the PCC values (supplementary figure 2).

These four clusters correspond to six species of *Glossina*, 22 species of Anopheles, 30 species of *Drosophila*, and 3 species of *Aedes* and *Culex*. Table 1 sums up several statistics for the four groups predicted by this analysis. The p-value describes how well the species from a given cluster separate

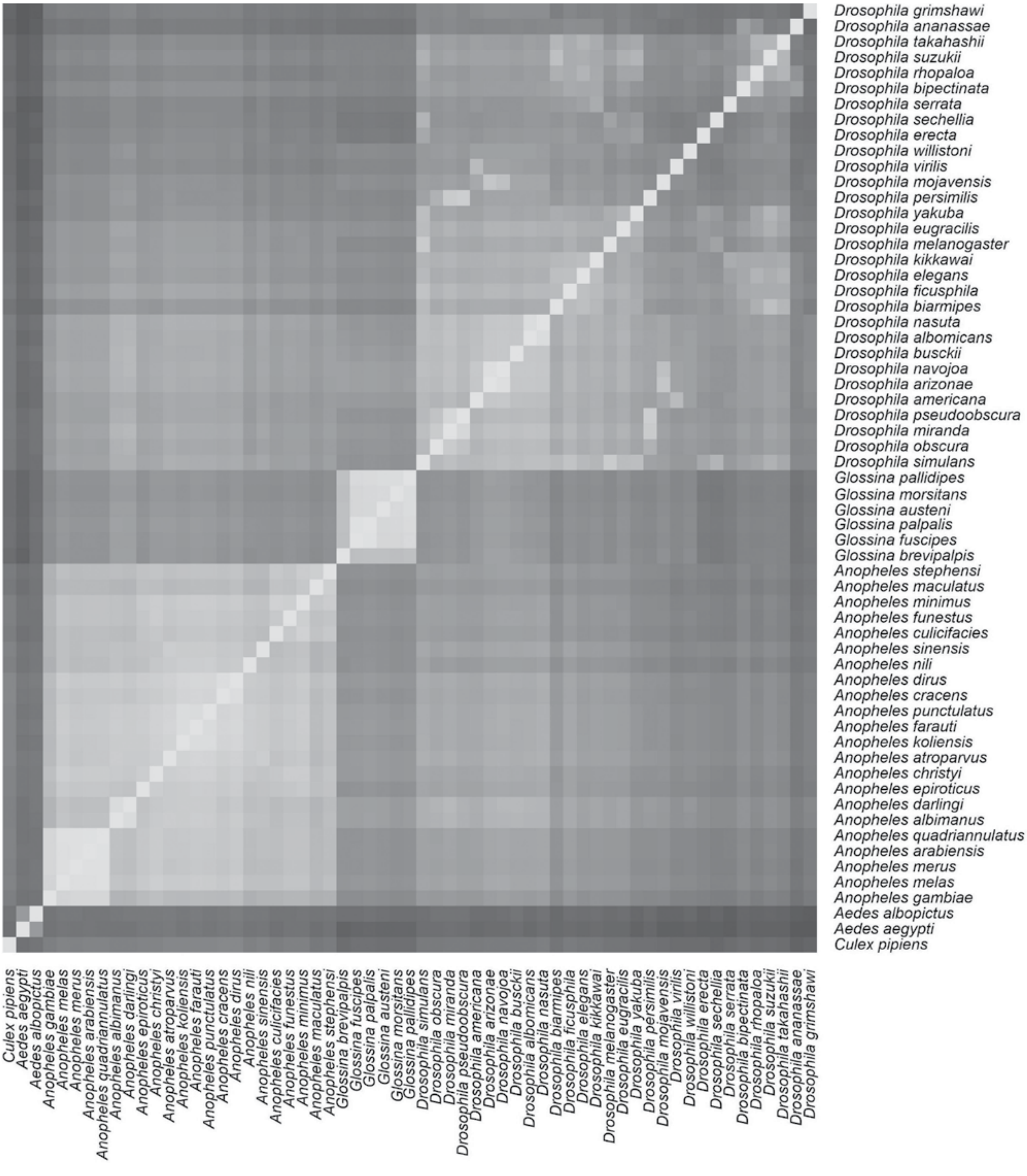


Figure 2. Heat map showing PCC values between 61 species analyzed by the WGKS algorithm. Lighter shades represent PCC values closer to 1, indicating species from the same baramin. Darker shades represent PCC values closer to 0, indicating species from different baramins.

from all other species in the study. To calculate the p-value, those PCC values calculated between species were compared from a given cluster versus those PCC values for all species pairs between a species from the given cluster and a species from the non-cluster. Three p-values are extremely low, but the p-value for *Aedes*+*Culex* is statistically insignificant (0.433). This could be because only a small number of species from these two genera were included in this analysis. This indicates that the species in this analysis for sure correspond to three putative baramins: *Drosophila*, *Anopheles,* and *Glossina*.

All PCC values were depicted in a beeswarm plot and an ECDF plot in supplementary figure 3A and B. The JCV values are largely spread out between 0 and 1.0. The ECDF plot also reflects this by taking on a sigmoidal curve with an inflection at around 0.35.

The three larger putative baramins in this study can be represented by a phylogenetic tree. The 30 *Drosophila* species in figure 3A are broken up into two monobaramins. These two monobaramins correspond to two big subgenera of the genus *Drosophila*, *Drosophila,* and *Sophophora*. *D. busckii* is the single member of the subgenus *Drosophila*. *D. grimshawi* seems to be misplaced, since this species is a member of the subgenus *Drosophila*. This species has the lowest mean PCC value within *Drosophila*, meaning that it must be an outlier species. *D. grimshawi* is endemic to the Hawaiian islands, and 32–39% of its genome is estimated to be made up of satellite sequences.[14]

Table 1. Group statistics for the four clusters discovered by the cutree algorithm from the PCC matrix using the WGKS method. St. dev. = standard deviation.

cluster	no. species	min. PCC	mean PCC	max. PCC	PCC st. dev.	p-value
Drosophila	30	0.123	0.452	0.955	0.152	1.08E-36
Anopheles	22	0.556	0.744	0.975	0.083	7.99E-214
Glossina	6	0.671	0.838	0.974	0.114	3.55E-11
Culex+Aedes	3	0.146	0.235	0.399	0.142	0.433

Table 2. Average silhouette width for three to five clusters using hierarchical clustering with the eclust R command for all three analyses

k	mtDNA analysis	GCM
3	0.58	0.64
4	0.48	0.58
5	0.49	0.5

Table 3. Group statistics for the three clusters discovered by the eclust algorithm from the mtDNA sequence similarity matrix

cluster	no. species	min. sim.	mean sim.	max. sim.	sim. st. dev.	p-value
Drosophila	13	0.864	0.9	0.999	0.03	4.91E-42
Anopheles	65	0.838	0.9	0.996	0.022	0
Culex+Aedes	20	0.880	0.936	0.997	0.033	2.23E-86

The six *Glossina* species form a tight cluster in the top right of the heat map. They have a mean PCC of 0.838, the highest of any predicted cluster. Their Neighbour Joining Tree can be seen in figure 3B. This tree follows the three subgenera: *Glossina* (*Morsitans*, including *G. m. morsitans*, *G. pallidipes,* and *G. austeni)*; *Palpalis* (including *G. palpalis*, *G. fuscipes)*; and *Fusca* (representing *G. brevipalpis*). Interestingly, the branches of the *Glossina* tree separate based on the difference in GC% (see Supplementary File 1). The Morsitans group has a GC% of 34.1%, whereas the Palpalis group has a value of 33.6%, and Fusca has a value of only 31.2%. Simply the difference in GC% might cause differences in k-mer frequencies, which lead to differences in k-mer scores, and which in turn lead to differences in CC values between species. *G. brevipalpis* is an outlier, with its low GC%; its genome size is also the smallest at 3.2 Gbp. Based on whole-genome nucleotide alignments of supercontigs and predicted coding sequences, *G. brevipalpis* is the least similar to all other species. It also differs from the other *Glossina* species in that it has the highest proportion of simple repeats and the lowest coverage of transposable elements. It also has on average less than 5,000 protein-coding genes less than the other species.[15] Based on this, it could be that *G. brevipalpis* is the archebaramin of this group.

Figure 3C depicts the Neighbour Joining tree for the *Anopheles* baramin. Within the baramin is a monobaramin made up of the *Anopheles gambiae* complex.[16] The average PCC value for the 22 *Anopheles* species is 0.744, which is much higher than the average value for the 30 *Drosophila* species. This could be due to the higher variation in the base background distribution in *Drosophila* than in *Anopheles*.[17] For the 22 Anopheles species the standard deviation is 0.083, whereas for the 30 *Drosophila* species this is 0.152. Compared to *Drosophila*, *Anopheles* genes also have relatively fewer introns.

Analysis of mitochondrial genomes

The whole mitochondrial genome (mtDNA) sequences of 98 species of *Anopheles*, *Culex*,

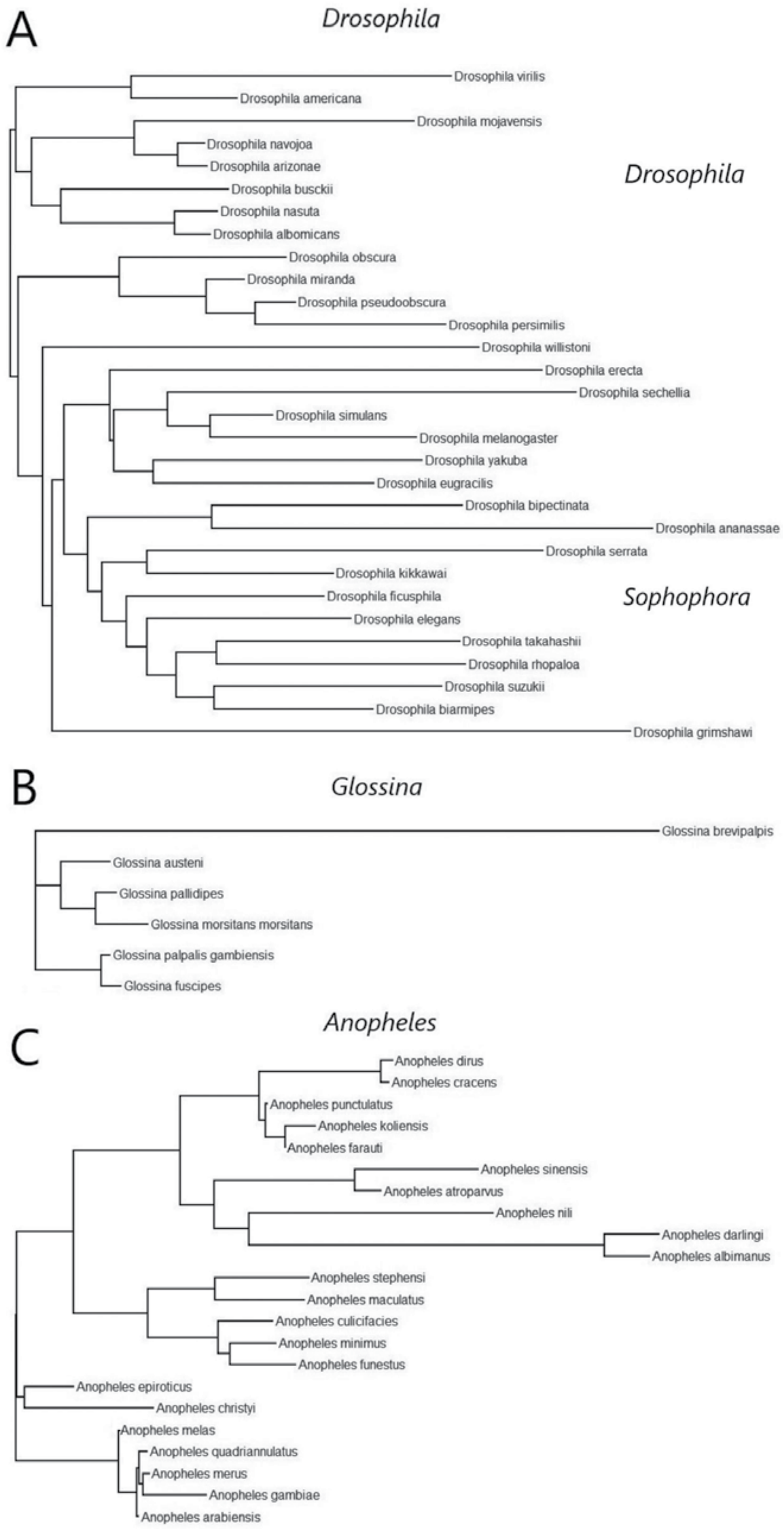


Figure 3. A. Neighbour Joining tree for the genera *Drosophila*, B. *Anopheles*, C. *Glossina*

Drosophila, and *Glossina*, were analyzed to see if there is any agreement between the mitochondrial DNA and the analysis of the whole genome sequence using the WGKS algorithm. These sequences were aligned with one another using the ggsearch36 algorithm. From this alignment a sequence identity matrix was calculated. The Hopkins clustering statistic was 0.904, which indicates that the similarity matrix is very good for clustering.

The eclust function was run on the mtDNA sequence similarity matrix. The average silhouette width was 0.58 for three clusters (see table 2 and supplementary figure 4). The species list, sequence similarity matrix, clusters, and clustering statistics can be found in Supplementary File 2. Table 3 sums up several statistics for the three groups. On the heat map in figure 4, *Drosophila* and *Culex* both form a separate cluster. Here *Drosophila* and *Anopheles* also both form a separate group, as well as *Aedes*+*Culex*.

This means that both mitochondrial and whole genome results corroborate the same baraminic classification. The beeswarm and ECDF plots can be seen in supplementary figures 5A and B. Since there are more species (98 compared to 58) under consideration, the beeswarm plot is broadened. The ECDF plot has two large and one small hump on it. The similarity values roughly above 0.825 represented by two clumps on the beeswarm plot and two humps on the ECDF plot may correspond to species pairs in the same baramin.

Application of the GCM algorithm

In order to measure how well the WGKS algorithm classifies individual species, the GCM algorithm[18] was applied on 44 species of *Aedes*, *Anopheles*, *Culex*, *Drosophila*, and *Glossina*. For this, the whole proteomes for these species were downloaded from UniProt database,

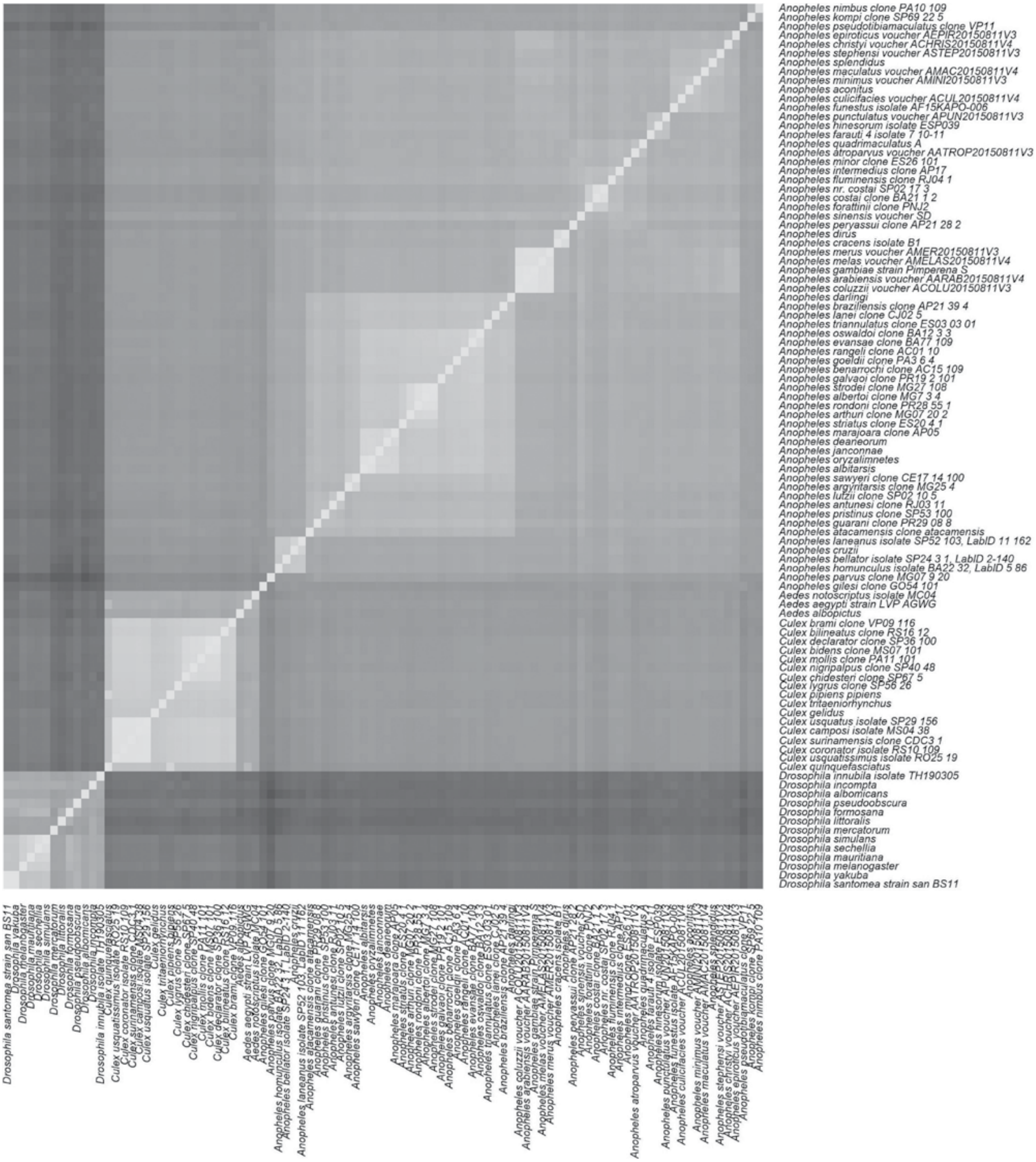


Figure 4. Heat map showing sequence similarity values between 98 species coming from the alignment of mtDNA whole genome sequences. Lighter shades represent PCC values closer to 1, indicating species from the same baramin. Darker shades represent PCC values closer to 0, indicating species from different baramins.

and run according to the protocol described in O'Micks, 2017. The Hopkins clustering statistic is 0.833, which is very high. Based on the eclust function, three clusters were predicted with an average silhouette width of 0.64 (supplementary figure 6). The species used in this analysis as well as the JCV matrix, the clustering, and the clustering statistics are available in Supplementary File 3. Table 4 sums up several statistics for the three groups.

The JCV values for each pair of species can be seen in the heat map in figure 5. In this figure, we can see that all mosquitos, *Drosophila* and *Glossina,* all separate well from one another. A larger clump of JCV values can be seen in the beeswarm plot in supplementary figure 7A at a value around 0.67, which corresponds to the large hump on the ECDF plot in figure 6B. Several smaller clumps can be observed above the first clump in the beeswarm plot. These clumps seem to spread out in the beeswarm plot and correspond to a gradual rise in the ECDF plot.

Gene synteny is highly conserved between *Drosophila* species, but it is much weaker between *D. melanogaster* and *A. gambiae*.[19] This indicates the separate baraminic status of these two insect genera. In this analysis *D. busckii* fits well within the *Drosophila* cluster. This could be because the GCM only deals with the presence or absence of genes, and not the gene order. In contrast, the WGKS method analyzes the whole genome sequence, taking all genomic information into account, and is thus a more fine-grained method.

Table 4. Group statistics for the three clusters discovered by the eclust algorithm from the JCV matrix using the GCM

cluster	no. species	min. JCV	mean JCV	max. JCV	JCV st. dev.	p-value
Drosophila	15	0.83	0.910	0.978	0.029	5.64E-182
mosquitos	22	0.787	0.819	0.859	0.022	3.97E-23
Glossina	6	0.625	0.782	0.917	0.067	3.24E-94

Table 5. Classification of different insect genera based on the three different analyses

	Aedes	*Anopheles*	*Culex*	*Drosophila*	*Glossina*
WGKS	C	B	C	A	D
mtDNA	C	B	C	A	-
GCM	C	C	C	A	D

In the JCV matrix, *D. pseudoobscura* had the absolute lowest mean JCV compared to all other species (0.487). In comparison, the average JCV value within *Drosophila* is 0.91. This species also has the smallest number of proteins mapped to orthology groups in the OrthoMCL database. In a comparison between the gene content of *D. melanogaster* and *D. pseudoobscura*, TBLASTN (a program which compares protein sequences against a dynamically translated nucleotide database) discovered 12,179 putative ortholog regions between the two genomes. Of these, only 9,946 genes (81.7%) had a reciprocal best protein hit with a protein from *D. melanogaster*.[15] For this reason *D. pseudoobscura* was removed from the analysis.

Comparative genome analyses show that chromosomal regions do not match up between *Aedes aegypti* and *D. melanogaster*. Instead, synteny is much more highly conserved between *Ae. aegypti* and *A. gambiae*.[20] The JCV between *Ae. aegypti* and *A. albopictus* is 0.799. Between the two *Aedes* species and species from *Anopheles* the mean JCV is 0.692. The mean JCV between the two *Aedes* species and *Drosophila* species is 0.593. The mean JCV between the *Aedes* species and the *Glossina* species is 0.569. This may indicate that *Aedes* could be part of the same baramin as *Anopheles*, as they are both genera from the family Culicidae. This could also possibly mean that mosquitos form a single holobaramin, but this conflicts with results from the WGKS method. Since the GCM only takes the coding regions into account, the results from the WGKS method might be more accurate.

Discussion

Comparing evidence from three studies including five genera of Dipteran insects, a new molecular baraminology method can provide us with new insight into the classification of different species into given baramins. Table 5 shows a comparison of the clustering results from the three analyses. The GCM classifies all mosquitos into the same holobaramin, whereas the WGKS method separates *Anopheles* from *Aedes*+*Culex*. The results from the mtDNA study give the same results as the WGKS algorithm. This could be because both studies analyze whole sequence data, whether it is from the whole genome, or from the mitochondrial genome. Summarizing the results for all three methods, we can say that *Drosophila*, *Glossina*, *Anopheles*, and *Aedes*+*Culex* belong to separate holobaramins.

Both the mtDNA study and the application of the GCM cover only a small portion (only 1–2%) of the entire genome. The GCM is based on gene content similarity, whereas the mtDNA study is based on overall sequence similarity. Compared to the mtDNA analysis, the WGKS results are based on the correlation of motif content, but in a way which conveys information from the whole genome. Furthermore, it appears that the WGKS algorithm tends to split groups in contrast to algorithms which lump species together (e.g. the GCM).

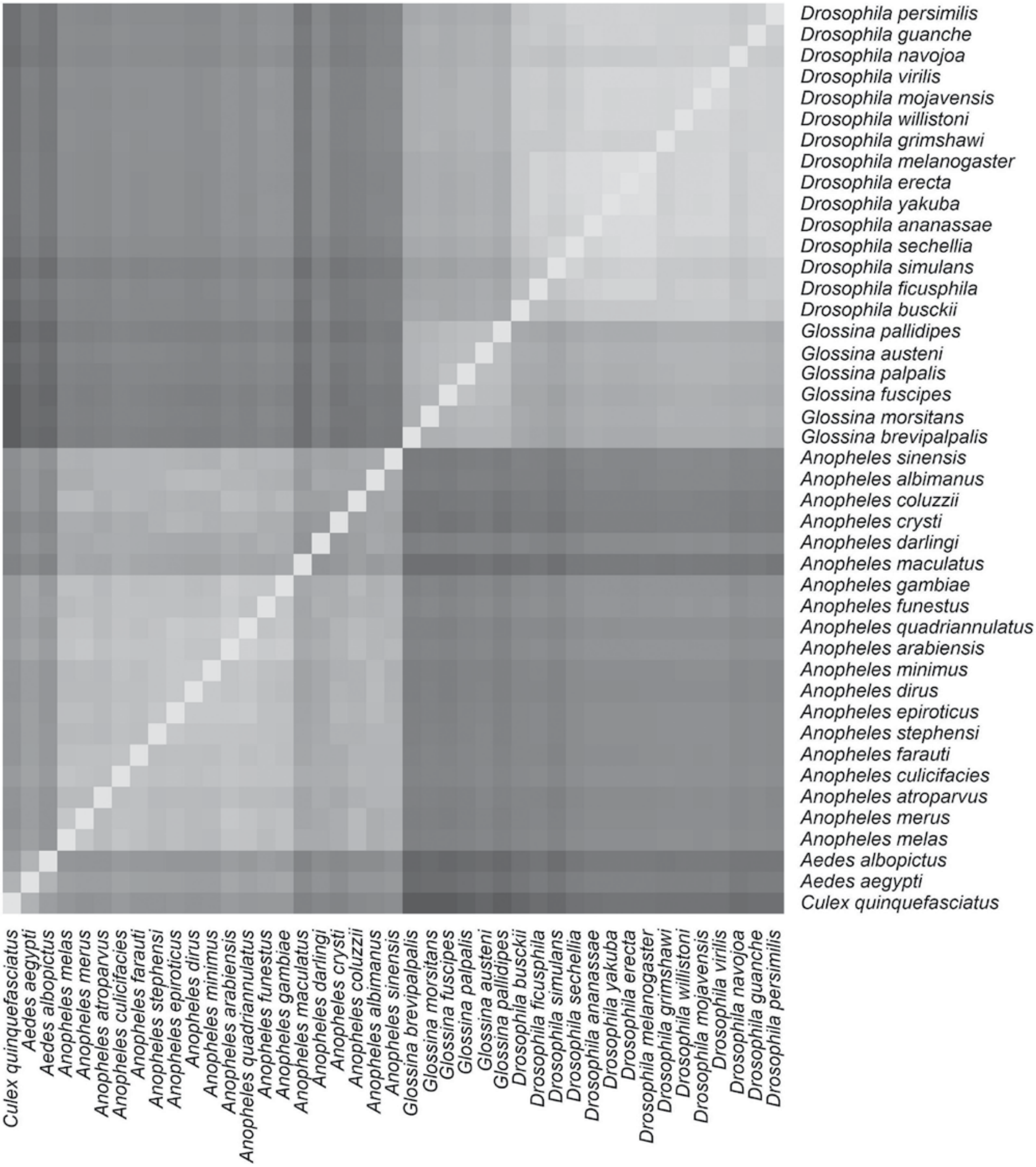


Figure 5. Heat map showing sequence similarity values between 44 species analyzed by the GCM. Lighter shades represent PCC values closer to 1, indicating species from the same baramin. Darker shades represent PCC values closer to 0, indicating species from different baramins.

The PCC values within a given baramin will vary according to the size of the baramin. The more species there are, the larger the variation, and the more diverse the life history of a given baramin, such as *Drosophila*. In contrast, smaller baramins such as *Glossina* have less variation and a less diverse baramin life history. The background base distribution might also be a factor in species and genome sequence diversity within baramins.

We may speculate that species from the archebaramin have similar genome sequences, made during Creation Week.

Species from the same baramin should generally be able to hybridize with one another. This should be all possible, since during Creation Week there would have been no mutations to obstruct this. One of the main results of the ENCODE project is that more than 80% of the entire human genome is made up of functional units (i.e. enhancers, transcription factor binding sites, repeat elements), which are active in at least one cell type.[12] If two species from the same baramin had a different chromosome number or a very different genome structure, this would not be possible. Many of these elements have a spatial restriction to them, meaning that different genetic elements must line up in a generally linear manner (synteny). For example, enhancer elements may be located very far from one another in the genome, but they have an effect on other genetic elements.

Synteny is widespread and helps understand the life history of a given baramin. Researchers have noted a significant difference in gene order on the X chromosome between *Drosophila* and *Anopheles*.[14] It also happens to be the case that *A. gambiae* has heteromorphic sex chromosomes showing no signs of recombination, whereas *Ae. aegypti* has homomorphic sex chromosomes.[21] This supports the separation of *Anopheles* and *Aedes* species into separate baramins.

The present algorithm is more precise and more rigorous than the GCM. This new molecular baraminology method can be used in addition to existing methods to classify species into baramins.

References

1. Cserhati, M., Xiao, P., and Guda, C., K-mer based motif analysis in insect species across *Anopheles*, *Drosophila* and *Glossina* genera and its application to species classification, *Computational and Mathematical Methods in Medicine* **2019**:4259479, 2019.
2. Patterson, C., Williams, D.M., and Humphries, C.J., Congruence between molecular and morphological phylogenies, *Annu. Rev. Ecol. Syst.* **24**:153–188, 1993.
3. Hillis, D.M., Molecular versus morphological approaches to systematics, *Ann. Rev. Ecol. Syst.* **18**:23–42, 1987.
4. Cserhati, M. and Tay, J., Comparison of morphology-based and genomics-based baraminology methods, *J. Creation* **33**(2):10–15, 2019.
5. Pollard, D.A., Iyer, V.N., Moses, A.M., and Eisen, M.B., Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting, *PLoS Genet.* **2**(10):e173, 2006.
6. Yang, K. and Zhang, L., Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction, *Nucleic Acids Res.* **36**(5):e33, 2008.
7. Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S.E., and Sachs, J., A global index representing the stability of malaria transmission, *Am. J. Trop. Med. Hyg.* **70**(5):486–98, 2004.
8. Yassin, A. and Orgogozo, V., Coevolution between male and female genitalia in the *Drosophila melanogaster* species subgroup, *PLoS One* **8**(2):e57158, 2013.
9. Izumitani, H.F., Kusaka, Y., Koshikawa, S., Toda, M.J., and Katoh, T., Phylogeography of the subgenus *Drosophila* (Diptera: Drosophilidae): evolutionary history of faunal divergence between the Old and the New Worlds, *PLoS One* **11**(7):e0160051, 2016.
10. Krafsur, E.S., Tsetse flies: genetics, evolution, and role as vectors, *Infect. Genet. Evol.* **9**(1):124–41, 2009.
11. Sinha, S. and Tompa, M., Discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Res.* **30**(24):5549–5560, 2002.
12. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature* **489**(7414): 57–74, 2012.
13. Pearson, W.R., Finding protein and nucleotide similarities with FASTA, *Curr. Protoc. Bioinformatics* **53**:3.9.1–3.9.25, 2016.
14. Craddock, E.M., Gall, J.G., and Jonas, M., Hawaiian *Drosophila* genomes: size variation and evolutionary expansions, *Genetica* **144**(1):107–24, 2016.
15. Attardo, G.M., Abd-Alla, A.M.M., Acosta-Serrano, A., Allen, J.E., Bateta, R., Benoit, J.B., *et al.*, Comparative genomic analysis of six *Glossina* genomes, vectors of African trypanosomes, *Genome Biol*. **20**(1):187, 2019.
16. Bass, C., Williamson, M.S., Wilding, C.S., Donnelly, M.J., and Field, L.M., Identification of the main malaria vectors in the Anopheles gambiae species complex using a TaqMan real-time PCR assay, *Malar. J.* **6**:155. 2007.
17. Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., *et al.*, Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes, *Science* **347**(6217):1258522, 2015.
18. O'Micks, J., Baraminology classification based on gene content similarity measurement, *CRSQ* **54**(1):27–37, 2017.
19. Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., *et al.*, Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution, *Genome Research* **15**(1):1–18, 2005.
20. Severson, D.W., DeBruyn, B., Lovin, D.D., Brown, S.E., Knudson, D.L., and Morlais, I., Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*, *J. Hered.* **95**(2):103–13, 2004.
21. Toups, M.A., and Hahn, M.W., Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes, *Genetics* **186**(2):763–6, 2010.

Matthew Cserhati *has a Ph.D. in biology. He has been an active creationist for 19 years and takes a great interest in molecular biology. He has published a number of articles in* Journal of Creation. *Matthew is currently studying at Greenville Presbyterian Theological Seminary. He works for CMI-US.*