geneCo: A visualized comparative genomic method to analyze multiple genome structures

Jaehee Jung\textsuperscript{1}, Jong Im Kim\textsuperscript{2}, and Gangman Yi\textsuperscript{*3}

\textsuperscript{1}Department of Information and Communication Engineering, Myongji University, Yongin, Gyeonggi-do, 17058, Korea
\textsuperscript{2}Department of Biology, Chungnam National University, Daejeon, 34134, Korea
\textsuperscript{3}Department of Multimedia Engineering, Dongguk University, Seoul, 04620, Korea

1 Data

The NCBI GenBank file, with extension .gb, .bgk and .gbf is required for geneCo. For comparative genomic analysis, several GenBank files are required as input files. The users need to upload input files into the system. The order of the uploaded files is used for the comparison of genomes, and geneCo draws the figure according to this order. For \( n \) genomes, we list the genomes, and each pair of genomes in the list is considered for comparison. Thus, each step to compare the pair of GenBank files compares the left and right genome, and draws them in accordance with options. Basically, geneCo supports changing the order of files by dragging the file name on the web-based user interface; it also supports user-defined options, which means that the user can modify most option values to analyze and draw genomes on the web-based interface. Most options are composed of Y or N selections, but some options need other values. Supplementary Figure 1 shows an example of configuration. Users can get various gene map images by modifying the option values. Users can freely use this proposed tool with the accessible URL: https://bigdata.dongguk.edu/geneCo/.

2 Algorithm

Supplementary Figure 2 shows the flow of the gene alignment algorithm based on the modified dynamic programming for gene comparison. The colored boxes in Supplementary Figure 2 represent genes. The upside of gene stands for non-complementary and the lower side of gene is complementary. One genome can be compared to another genome. As shown in the Figure, the size of box represents the length of gene and each gene is annotated with numbers, from 1 to 10. As (A) in Supplementary Figure 2 shows, boxes colored in blue and green are genes matched to each other. The blue boxes are matched to just one, which is identical to the query. However, green boxes labeled as 3 are simultaneously matched twice due to the duplication. In this case, geneCo displays all aligned genes simultaneously as shown at (B) in Supplementary Figure 2. This can be modified in the configuration file.

3 Results

Basically, geneCo supplies various configurations to draw genomes. The two major functions are the construction of a genome map and genome map comparison. the construction of a genome map is to draw genome maps based on information from the GenBank files. Supplementary Figure 3 is an example of three different genomes. This figure shows a sample output of genome map at a glance. The
Supplementary Figure 1: Example of geneCo configuration file generated by the web-based interface

```
$GROUPGENES       Y
$GROUPTYPE        M
$MISMATCHEDGENESONLY N

$DEFAULTCHRBARCOLOR  cmyk.BlueViolet
$DEFAULTGENEBBOXCOLOR cmyk.SpringGreen

$SHOWLEGEND       Y
$SHOWBPINDICATOR  F
$SHOWCHRTITLE      N
$SHOWCHRLOCUS      N
$SHOWCHRBP         N
$BPINDICATORCOLOR  cmyk.SkyBlue
```

Supplementary Figure 2: Example of the proposed algorithm for the identification of matched genes between two genomes.
different functional categories contain their own colors that are pre-defined in the color map file. Genes based on location information are shown with their functional category color in each genome vertical bar. Supplementary Figure 3 is the result of genome mapping with 15 different species, and various examples by genome map comparison, which are from Supplementary Figures 5 to 8 were performed to show the combination of geneCo functions for the visualization of various outputs. The hardware specifications used in the experiments were Intel Xeon E5-2695 v4, 2.10GHz, SSD, 128GB RAM. The maximum execution time of all experiments with the 15 GenBank files was about 4 sec and the maximum memory size was about 170 MB. This result is the average time from 10 executions of only geneCo engine in the terminal, thus, actual execution time, using web interface in various environments, of different network connections, can lead to different execution time and can take more time, depending on users computational environments.

### 3.1 Multiple construction of genome Maps

Supplementary Figure 3: Example of the simple draw.

In this experiment, 15 GenBank files were used [Table 2]. The species used were gorilla, chimpanzee, dog, cat, horse, dolphin, chicken, frog, fly, yeast, green algae (Chlorella), cryptophytes (Cryptomonas), brown algae, red algae, and amoeba. These sample results may be helpful for users to draw genomes. Supplementary Figure 3 shows multiple gene maps of three different species. The figure includes the genbank accession number, length, and gene names.

### 3.2 Genome map comparison

Another major function in geneCo is Map comparison that shows identical or different genes between genomes. The function of the genome map comparison is constructed by two functional categories: aligning genes to identify matched genes between genome that can be further conserved genes and finding genes that are not matched with each other between genomes that can be further non-conserved genes. With various configuration to draw outputs, geneCo supports various options such as the color palette, range to display, object size, etc.
3.2.1 Comparison of Common Genes

It is difficult to check all GenBank files if we want to identify similarities or differences of genomic features, but geneCo supports the identification of identical or different gene names of different genomes. We used the modified sequence alignment algorithm based on dynamic programming algorithm. Supplementary Figure 6 shows matched genes of two different concatenated genomes. The line color between matched genes is to ensure that users can easily recognize that they are similar. Supplementary Figure 7 is the same result with Supplementary Figure 6, but the difference is that it displays matched gene names only. If the simple figure is required to identify regions that show identical or different genes, the configuration can be adjusted. Supplementary Figure 8 is a good example of this. geneCo can adjust the configuration for most objects in the result such as gene names, bp indicator, etc.

3.2.2 Comparison of Non-Common Genes

Supplementary Figure 8 shows the result of the alignment of mismatched genes with mismatched gene names, while Supplementary Figure 9 presents the lower gap value between genomes and those without gene names. We can adjust the display option for the interval length of genomes.
Supplementary Figure 7: Alignment of matched genes without gene names

Supplementary Figure 8: Alignment of mismatched gene names between the pair of two different genomes

Supplementary Table 1: The functional comparison of five different applications.

| Feature                                | Circos | Mauve | BRIG | ACT | geneCo |
|----------------------------------------|--------|-------|------|-----|--------|
| Multiple sequence analysis             | o      | o     | o    | o   | o      |
| GenBank format                         | o      | o     | o    | o   | o      |
| Web-based interface                    |        |       |      |     |        |
| User-defined configuration             | o      | o     |      | o   | o      |
| Output file in vector file type        |        |       |      |     |        |
| Zoom in the specific gene region       |        |       |      |     |        |
| Mismatched (non-conserved) genes       |        |       |      |     |        |
Supplementary Figure 9: Alignment of mismatched gene without name between two different genomes with lower gap value

4 Evaluation

Supplementary Table 1 is the comparison of five different software tools. The important key feature of geneCo, compared to existing software, is that it supports the identification of mismatched genes. As we described in Section 3.2, geneCo can identify non-conserved genes and conserved genes based on user options. geneCo not only supports web-based user interface, but also provides user configurations such as Supplementary Figure 1.

Supplementary Table 3 presents nucleomorph, chromosome 1 and plastid genomes. Supplementary Table 5 shows output figures that were generated by the data of nucleomorph, chromosome 1, between the first and the fourth rows in Supplementary Table 3. Supplementary Table 5 shows those tested with plastid examples, from the fifth to the eighth row. In two comparison tables, options used in geneCo configuration were mismatched genes = Yes and Display name of mismatched genes only = Yes. Mauve and ACT shows matched regions only. ACT and geneCo support the zoom-in function to investigate the specific regions.

In the BRIG result on Supplementary Table 4, we can see many regions colored in white, which means that many mismatched genes exist. However, geneCo finds all mismatched genes and names them in each position. The color of BRIG in Supplementary Table 5 is darker than in Supplementary Table 4 because the plastid length is usually smaller than nucleomorph, chromosome 1.
### Supplementary Table 2: Representative mitochondrion data set to evaluate geneCo.

| Species Description          | GenBank Accession | Length | Reference |
|-----------------------------|-------------------|--------|-----------|
| *Gorilla Descriptio*        | NC_001645         | 16364  | 4         |
| *Canis lupus familiaris*    | NC_002008         | 16727  | 10        |
| *Felis catus*               | NC_001700         | 17009  | 15        |
| *Equus caballus*            | NC_001640         | 16660  | 21        |
| *Delphinus capensis*        | NC_012061         | 16385  | 20        |
| *Gallus gallus*             | NC_001323         | 16775  | 18        |
| *Xenopus borealis*          | NC_018776         | 17474  | 14        |
| *Drosophila melanogaster*   | NC_024511         | 19524  | 11        |
| *Saccharomyces cerevisiae*  | isolate NCYC3594  |        |           |
| *Chlorella variabilis*      | isolate NC64A     |        |           |
| *Cryptomonas curvata*       | strain CNUKR      |        |           |
| *Undaria pinnatifida*       |                   |        |           |
| *Pyropia yezoensis*         |                   |        |           |
| *Naegleria fowleri*         |                   |        |           |

### Supplementary Table 3: Cryptomonas paramecium, nucleomorph, chromosome 1 Data sets used for four different applications.

| Genome | GenBank Accession | Length     | Reference | Figure |
|--------|-------------------|------------|-----------|--------|
| *Cryptomonas paramecium*, nucleomorph, chromosome 1 | nucleomorph, chromosome 1 | NC_015329 | 177,338 bp | 17 | Table 4 |
| *Chroomonas mesostigmatica*, nucleomorph, chromosome 1 | nucleomorph, chromosome 1 | CP003680 | 243,993 bp | 16 | Table 4 |
| *Hemiselmis andersenii*, nucleomorph, chromosome 1 | nucleomorph, chromosome 1 | NC_009977.1 | 207,524 bp | 12 | Table 4 |
| *Guillardia theta* nucleomorph, chromosome 1 | nucleomorph, chromosome 1 | AF165818 | 196,216 bp | 22 | Table 4 |
| *Rhodomonas salina* plastid | plastid | NC_009573 | 135,854 bp | 6 | Table 3 |
| *Teleaulax amphioxeia* plastid | plastid | NC_027589 | 129,772 bp | 8 | Table 3 |
| *Guillardia theta* plastid | plastid | NC_000926 | 121,524 bp | 2 | Table 3 |
| *Chroomonas placoidea* plastid | plastid | NC_035721 | 139,432 bp | 7 | Table 3 |
Supplementary Table 4: Output for the gene comparison in four applications with nucleomorph, chromosome 1 genome data.

| Mauve          | ACT          |
|----------------|--------------|
| ![Mauve Image] | ![ACT Image] |

| BRIG          | geneCo       |
|---------------|--------------|
| ![BRIG Image] | ![geneCo Image] |
Supplementary Table 5: Output for the gene comparison in four applications with plastid genome data.

| Mauve | ACT |
|-------|-----|
| ![Mauve Diagram](image1.png) | ![ACT Diagram](image2.png) |

| BRIG | geneCo |
|------|--------|
| ![BRIG Diagram](image3.png) | ![geneCo Diagram](image4.png) |
References

[1] S. E. Celniker, D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S. P. Dugan, E. Frise, A. Hodgson, R. A. George, R. A. Hoskins, T. Laverty, D. M. Muzny, C. R. Nelson, J. M. Pacleb, S. Park, B. D. Pfeiffer, S. Richards, E. J. Sodergren, R. Svirskas, P. E. Tabor, K. Wan, M. Stapleton, G. G. Sutton, C. Venter, G. Weinstock, S. E. Scherer, E. W. Myers, R. A. Gibbs, and G. M. Rubin. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome biology*, 3(12), 2002.

[2] S. E. Douglas and S. L. Penny. The Plastid Genome of the Cryptophyte Alga, *Guillardia theta*: Complete Sequence and Conserved Synten groups Confirm Its Common Ancestry with Red Algae. *Journal of Molecular Evolution*, 48:236–244, Feb. 1999.

[3] W. Fan, W. Guo, J. L. Van Etten, and J. P. Mower. Multiple origins of endosymbionts in chlorellaceae with no reductive effects on the plastid or mitochondrial genomes. *Scientific Reports*, 7(1):10101, 2017.

[4] D. R. Foran, J. E. Hixson, and W. M. Brown. Comparisons of ape and human sequences that regulate mitochondrial dna transcription and d-loop dna synthesis. *Nucleic Acids Res*, 16(13):5841–5861, Jul 1988.

[5] E. K. Herman, A. L. Greninger, G. S. Visvesvara, F. Marciano-Cabral, J. B. Dacks, and C. Y. Chiu. The mitochondrial genome and a 60-kb nuclear DNA segment from *Naegleria fowleri*, the causative agent of primary amoebic meningoencephalitis. *J Eukaryot Microbiol*, 60(2):179–191, Mar 2013.

[6] H. Khan, N. Parks, C. Kozera, B. A. Curtis, B. J. Parsons, S. Bowman, and J. M. Archibald. Plastid Genome Sequence of the Cryptophyte Alga *Rhodomonas salina* CCMP1319: Lateral Transfer of Putative DNA Replication Machinery and a Test of Chromist Plastid Phylogeny. *Molecular Biology and Evolution*, 24(8):1832–1842, 05 2007.

[7] J. I. Kim, C. E. Moore, J. M. Archibald, D. Bhattacharya, G. Yi, H. S. Yoon, and W. Shin. Evolutionary dynamics of cryptophyte plastid genomes. *Genome Biol Evol*, 9(7):1859–1872, Jul 2017.

[8] J. I. Kim, H. S. Yoon, G. Yi, H. S. Kim, W. Yih, and W. Shin. The Plastid Genome of the Cryptomonad *Teleaulax amphioxea*. *PLOS ONE*, 10(6):1–17, 06 2015.

[9] J. I. Kim, H. S. Yoon, G. Yi, W. Shin, and J. M. Archibald. Comparative mitochondrial genomics of cryptophyte algae: gene shuffling and dynamic mobile genetic elements. *BMC Genomics*, 19(1):275, 2018.

[10] K. S. Kim, S. E. Lee, H. W. Jeong, and J. H. Ha. The complete nucleotide sequence of the domestic dog (canis familiaris) mitochondrial genome. *Molecular Phylogenetics and Evolution*, 10(2):210 – 220, 1998.

[11] F. Kong, P. Sun, M. Cao, L. Wang, and Y. Mao. Complete mitochondrial genome of *Pyropia yezoensis*: reasserting the revision of genus *Porphyra*. *Mitochondrial DNA*, 25(5):335–336, 2014. PMID: 23841614.

[12] C. E. Lane, K. van den Heuvel, C. Kozera, B. A. Curtis, B. J. Parsons, S. Bowman, and J. M. Archibald. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci U S A*, 104(50):19908–19913, Dec 2007.

[13] T.-Y. Li, J.-Q. Qu, Y.-J. Feng, C. Liu, S. Chi, and T. Liu. Complete mitochondrial genome of *Undaria pinnatifida* (Alariaceae, Laminariales, Phaeophyceae). *Mitochondrial DNA*, 26(6):953–954, 2015. PMID: 24409911.

[14] R. E. Lloyd, P. G. Foster, M. Guille, and D. T. J. Littlewood. Next generation sequencing and comparative analyses of *Xenopus* mitogenomes. *BMC Genomics*, 13(1):496, Sep 2012.
[15] J. V. Lopez, S. Cevario, and S. J. O’Brien. Complete Nucleotide Sequences of the Domestic Cat (Felis catus) Mitochondrial Genome and a Transposed mtDNA Tandem Repeat (Numt) in the Nuclear Genome. *Genomics*, 33(2):229 – 246, 1996.

[16] C. E. Moore, B. Curtis, T. Mills, G. Tanifuji, and J. M. Archibald. Nucleomorph Genome Sequence of the Cryptophyte Alga *Chroomonas mesostigmatica* CCMP1168 Reveals Lineage-Specific Gene Loss and Genome Complexity. *Genome Biology and Evolution*, 4(11):1162–1175, 10 2012.

[17] G. Tanifuji, M. Dlutek, N. T. Onodera, N. Donaher, J. M. Archibald, and T. J. Wheeler. Complete Nucleomorph Genome Sequence of the Nonphotosynthetic Alga *Cryptomonas paramecium* Reveals a Core Nucleomorph Gene Set. *Genome Biology and Evolution*, 3:44–54, 12 2010.

[18] J. R. Valverde, R. Marco, and R. Garesse. A conserved heptamer motif for ribosomal rna transcription termination in animal mitochondria. *Proc Natl Acad Sci U S A*, 91(12):5368–5371, Jun 1994.

[19] J. F. Wolters, K. Chiu, and H. L. Fiumera. Population structure of mitochondrial genomes in *Saccharomyces cerevisiae*. *BMC Genomics*, 16(1):451, 2015.

[20] Y. Xiong, M. C. Brandley, S. Xu, K. Zhou, and G. Yang. Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. *BMC Evolutionary Biology*, 9(1):20, Jan 2009.

[21] X. Xiufeng and Ú. Árnason. The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region. *Gene*, 148(2):357 – 362, 1994.

[22] S. Zauner, M. Fraunholz, J. Wastl, S. Penny, M. Beaton, T. Cavalier-Smith, U.-G. Maier, and S. Douglas. Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. *Proceedings of the National Academy of Sciences*, 97(1):200–205, 2000.