


# Rapid Targeted Assembly of the Proteome Reveals Evolutionary Variation of GC Content in Avian Lice

Avery R Grant<sup>1</sup> , Kevin P Johnson<sup>2</sup>, Edward L Stanley<sup>3</sup>, James Baldwin-Brown<sup>4</sup>, Stanislav Kolenčik<sup>5</sup> and Julie M Allen<sup>6</sup>

<sup>1</sup>Department of Biology, University of Nevada, Reno, Reno, NV, USA. <sup>2</sup>Illinois Natural History Survey, Prairie Research Institute, University of Illinois at Urbana-Champaign, Champaign, IL, USA. <sup>3</sup>Department of Natural History, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. <sup>4</sup>Department of Biology, The University of Utah, Salt Lake City, UT, USA. <sup>5</sup>Faculty of Mathematics, Natural Sciences, and Information Technologies, University of Primorska, Koper, Slovenia. <sup>6</sup>Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA.

Bioinformatics and Biology Insights  
Volume 18: 1–11  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11779322241257991



**ABSTRACT:** Nucleotide base composition plays an influential role in the molecular mechanisms involved in gene function, phenotype, and amino acid composition. GC content (proportion of guanine and cytosine in DNA sequences) shows a high level of variation within and among species. Many studies measure GC content in a small number of genes, which may not be representative of genome-wide GC variation. One challenge when assembling extensive genomic data sets for these studies is the significant amount of resources (monetary and computational) associated with data processing, and many bioinformatic tools have not been optimized for resource efficiency. Using a high-performance computing (HPC) cluster, we manipulated resources provided to the targeted gene assembly program, automated target restricted assembly method (aTRAM), to determine an optimum way to run the program to maximize resource use. Using our optimum assembly approach, we assembled and measured GC content of all of the protein-coding genes of a diverse group of parasitic feather lice. Of the 499 426 genes assembled across 57 species, feather lice were GC-poor (mean GC = 42.96%) with a significant amount of variation within and between species (GC range = 19.57%–73.33%). We found a significant correlation between GC content and standard deviation per taxon for overall GC and GC<sub>3</sub>, which could indicate selection for G and C nucleotides in some species. Phylogenetic signal of GC content was detected in both GC and GC<sub>3</sub>. This research provides a large-scale investigation of GC content in parasitic lice laying the foundation for understanding the basis of variation in base composition across species.

**KEYWORDS:** Bioinformatics, computational resource efficiency, base composition, feather lice, protein-coding genes, phylogenetic signal, AT (adenine/thymine) rich

**RECEIVED:** September 25, 2023. **ACCEPTED:** May 2, 2024.

**TYPE:** Research Article

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NSF grants 1925312 to JMA, NSF DEB 1925487 and DEB 1926919 to KPJ. The computational work in this publication was made possible by a grant from the National Institute of General Medical Sciences (GM103440) from the National Institutes of Health. The authors would like to acknowledge the support of Research & Innovation and the Cyberinfrastructure

Team in the Office of Information Technology at the University of Nevada, Reno for facilitation and access to the Pronghorn High-Performance Computing Cluster.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Avery R Grant, Department of Biology, University of Nevada, Reno, Reno, NV 89557, USA. Email: [averygrant@unr.edu](mailto:averygrant@unr.edu)

## Introduction

Genomic characteristics, such as base composition, play an important role in the evolution and ecology of organisms. These features can be influential in molecular mechanisms involved in gene function, phenotype, and amino acid composition.<sup>1–3</sup> Base composition is typically measured as GC content (proportion of guanine and cytosine in DNA), which has been directly linked to amino acid composition.<sup>4</sup> As amino acids are the building blocks of proteins, variation in amino acid composition is a critical component of protein evolution.<sup>4</sup> Measuring GC content within and among species may be the first step for understanding adaptation at a molecular level. For example, identifying adaptive alleles, which may be indirectly constrained based on GC content,<sup>5,6</sup> in threatened populations is imperative for species and ecosystem conservation.<sup>7</sup> GC content has been found to be highly variable among and within species as well as at different organizational levels (ie, proteome vs genome<sup>8,9</sup>). Here, we mainly focus on the GC

content of protein-coding genes, and it is important that comparisons between studies be made using analogous data sets. Even with this large-scale variability, some patterns stand out, such as higher recombination rates in GC-rich genes<sup>10</sup> and a negative relationship between GC content of third codon positions and chromosome length.<sup>11,12</sup> The patterns of GC content variation across organisms are thought to be linked to genomic characteristics such as methylation throughout the genome,<sup>13</sup> expression levels of coding genes<sup>14,15</sup>, and genome-wide gene conversion.<sup>16</sup> Many hypotheses have been suggested to explain variation in GC content across different regions of the genome, such as molecular mechanics, environmental factors, natural selection, or a combination thereof.<sup>16–18</sup> Before any of these mechanisms can be investigated, determining GC content across an organism's genes and comparing the variation found among closely related species is needed to understand how base composition has influenced diversification and adaptation.<sup>19</sup>



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

A disproportionate amount of genomic research has focused on vertebrate groups, particularly mammals and birds.<sup>20,21</sup> Insects have been given much less attention yet are the most diverse group of animals in the world<sup>22</sup> and are facing catastrophic declines.<sup>23</sup> Parasitic lice (Phthiraptera) are of particular interest due to their global distribution, fast rate of evolution, and high level of diversification,<sup>24–28</sup> providing an ideal system to study GC content and protein evolution in a closely related group of organisms<sup>29</sup>. Lice have among the smallest genomes within insects<sup>30</sup>; however, only a few studies have examined GC content of coding genes in this group with conflicting results. A small number of genes were found to be GC-rich, whereas a much larger set were GC-poor.<sup>31,32</sup> Many insects have low GC content overall,<sup>33–36</sup> which is consistent with the results of GC content of coding genes in parasitic lice found by Virrueta Herrera et al,<sup>32</sup> even though the mtDNA of some parasitic lice genes is GC-rich compared with other insects<sup>31,37</sup>. A large-scale genome-wide investigation would provide a better understanding of the patterns in base composition of parasitic lice and allow for more comprehensive comparisons with other organisms.

These genomic studies often use gene assembly programs run on a high-performance computing (HPC) cluster. A significant amount of computational resources are necessary for large-scale data, and these resources are not always available or accessible.<sup>38</sup> Researchers often pay for an allotted amount of resources and are charged for these resources even if they are idle, leading to wasted time and money.<sup>39</sup> Given that gene assemblies are generally resource-intensive, inefficient resource use can quickly become wasteful. A critical component of continuing the advancement and accessibility of molecular studies is the development of programs that can efficiently prepare and analyze these genome-scale data sets.

Automated target restricted assembly method (aTRAM)<sup>40,41</sup> is a targeted gene assembly program, which assembles specific loci from unassembled sequences using a closely or distantly related locus as the reference.<sup>42</sup> Automated target restricted assembly method begins by creating a library from the unassembled sequence reads, which consists of BLAST formatted databases split into multiple groups of paired-end sequences and a relational database to associate read-pairs. Next, a target sequence is blasted against these groups to identify homologous reads which are then assembled into a contiguous piece of DNA, termed contig. This process is repeated using the newly assembled contig as the query sequence and so on for multiple iterations until the target locus is assembled. Through this process, aTRAM breaks up large tasks (eg, Basic Local Alignment Search Tool (BLAST)) into multiple small tasks, using central processing units (CPUs). Automated target restricted assembly method allows the user to determine the number of CPUs in each run and uses them to blast the groups of paired-end reads within a library in parallel, increasing the overall rate of gene assembly. However, it is unclear if increasing the number of

CPUs results in the most efficient use of resources. Adding more CPUs might be lowering the computational efficiency as more of these resources might be sitting idle.

Here, we assembled all nuclear protein-coding genes for several feather louse (Ischnocera) genera. We used a reference set of genes from the recently annotated pigeon louse (*Columbicola columbae*) genome<sup>43</sup> to measure GC content across the proteome of this diverse group of ectoparasites. We compared GC content among genes and across species and estimated the phylogenetic signal of GC content. Our aim was to gain an understanding of the level of variation in base composition found within this group of insects and to provide the data needed to continue investigating protein evolution.

We used aTRAM, which was designed to maximize resource use by allowing the user to incorporate as many cores as available during the assembly process. However, based on the time and resources used to assemble a large number of genes from a single taxon, it is unclear if aTRAM is using those resources efficiently, or if there are more optimal ways to use resources (eg, in parallel) to maximize efficiency. Before assembling the genes for our full set of taxa, we first investigated aTRAM resource efficiency given different amounts of computational resources to see how the available resources are being used during the assembly process. Focusing on 2 of the most commonly manipulated computational resources (CPUs and tasks), we measured the rate of and computational efficiency of gene assemblies using a varying number of resources. Our goal is to maximize computational efficiency whereas optimizing the rate at which genes can be assembled.

## Methods

Our data set included 57 species in 54 genera of parasitic feather lice. Raw data were obtained from the NCBI (National Center for Biotechnology Information) Sequence Read Archive (SRA) (see Supplemental Data for SRA). All paired-end reads were trimmed using Trimmomatic v0.39 in paired-end mode to remove areas of low quality and to clip adapters (Illumina universal adapter).<sup>44</sup> We used a sliding window of 4 base pairs with a minimum quality of 20 (Phred + 33), and all reads shorter than 100 base pairs were dropped. All genes were assembled with aTRAM v2.4.3<sup>42</sup> using the 13 362 annotated genes from the pigeon louse genome<sup>43</sup> as a reference.

### *aTRAM*

The software aTRAM was designed to maximize computational resources. For example, running BLAST searches on each group of paired-end reads can be done in parallel using many CPUs or sequentially using a single CPU. It is unknown which of the following methods would improve gene assembly rate and resource efficiency: a) adding more CPUs linearly or b) running multiple instances of aTRAM in parallel with fewer CPUs. Although HPC clusters generally have multiple options

for resource manipulation, we focus on 2 common resources users have to select when running programs with HPC: numbers of CPUs and tasks. All computational tests were run on the University of Nevada, Reno HPC cluster, Pronghorn, which uses Slurm as a workload manager. For all tests, the same gene (chr1-aug-0.14-mRNA-1) and library (*Alcedoecus*: 55 groups of paired-end reads) were used. We examined resource efficiency by measuring the percent of available resources used with an increasing scale of CPUs and tasks.

### CPU<sub>s</sub>

We altered Slurm sbatch options (`-cpus-per-task`) to determine the change in gene assembly rate with an increasing number of CPUs. Efficiency was measured by the number of genes assembled per CPU per hour. All tests were run exclusively on a single node with 1 task and a 2-hour time limit. These tests were run using 1, 2, 4, 8, 16, and 32 CPUs setting the CPU argument in aTRAM to the same value (`-cpus`; details about aTRAM arguments here: <https://github.com/juliema/aTRAM>). We tested these methods with 2 different assemblers, Trinity<sup>45</sup> and ABySS.<sup>46</sup>

### Tasks

The next test focused on parallelization to measure CPU use efficiency and gene assembly rate. All tests were run exclusively on a single node with Slurm argument `-cpus-per-task=1` (see results). By adding tasks, we effectively increase the number of instances of aTRAM running simultaneously. We used the same scaling, (1, 2, 4, 8, 16, 32), for tasks and for the aTRAM `-cpu` argument. With ABySS, we used message passing interface (MPI) mode using `-abyss-mp`. This changes the number of processors used in parallel execution; however, using this mode with ABySS is no longer recommended but does not change the outcome of our research. We used the same scaling for `-abyss-mp`. This argument was not an option to use with Trinity. Values for these 3 parameters (`-ntasks`, aTRAM CPUs, and `-abyss-mp`) matched for all tests. Each test assembled 64 copies of the same gene, and no time limit was given. Due to the high level of I/O (input/output) during assembly, we used a temporary file system to reduce the amount of memory needed to store files during gene assembly. Central processing unit and memory efficiency were obtained with Slurm `seff` command for each job. Tests were run with both Trinity and ABySS.

### Full data set assembly

We assembled a target set of 13364 of the protein-coding genes for 57 feather lice taxa based on the results from the CPU and task tests with aTRAM using ABySS. These genes were all of the annotated genes from the pigeon louse genome. The amino acid sequences from these genes were used as the reference for *tblastn* searches. Exonerate<sup>47</sup> was used to stitch

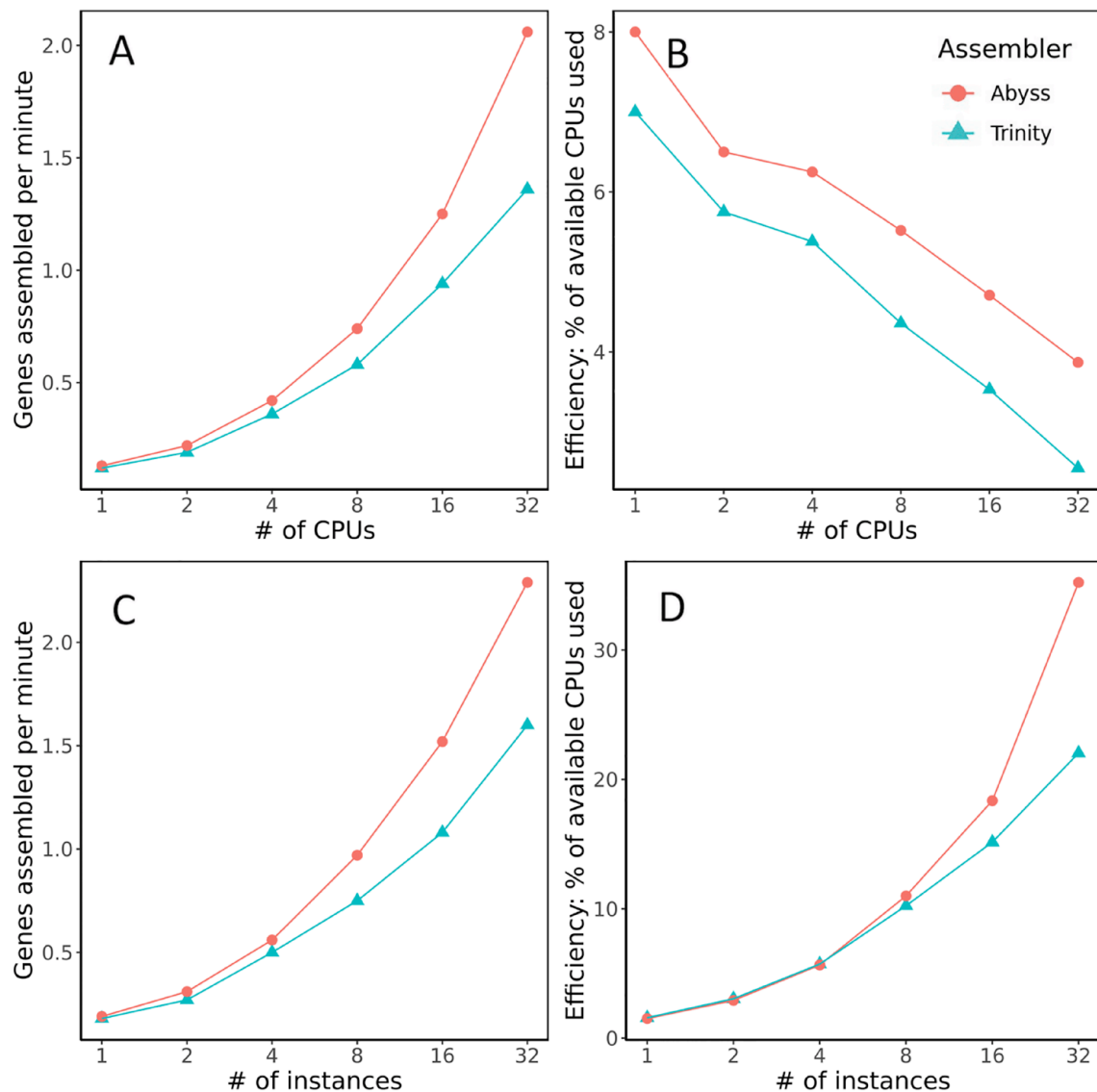
together assembled contigs and concatenate exons using the pigeon louse amino acid reference sequences. Once genes were assembled, genes that were not selected as reciprocal best hit (RBH) for the associated pigeon louse target after a reciprocal best BLAST search were removed. Using the pigeon louse gene set, we ran analysis of variance (ANOVA) tests in R v4.3.1<sup>48</sup> to compare the length of all genes, the length of genes that assembled for at least 1 taxon, and the length of genes that did not assemble for any taxa.

### GC content

GC content was measured ( $(\# \text{ of } Gs + \# \text{ of } Cs) \div \text{length of gene}$ ) for all remaining genes ( $n=499\,426$ ) using an original Python (v3.8.5) script excluding Ns from calculations. Removing Ns from sequences did not impact our results of GC content. Automated target restricted assembly method outputs assembled genes in reading frame so GC content was also calculated at each codon position ( $GC_1$ ,  $GC_2$ , and  $GC_3$ ). From here on, we refer to GC content across all codon positions for an entire gene or taxon as simply GC. Phylogenetic linear regression models (R package *phylolm*<sup>49</sup>) were used to assess the correlation between overall GC content and variation in GC content (standard deviation [SD]). This was done in R for GC and  $GC_3$  per taxon. We focused on  $GC_3$  over  $GC_1$  and  $GC_2$  because it is the least evolutionarily constrained codon position.<sup>50</sup> Because changing the base at the third codon position typically does not change the amino acid, third positions are less constrained and more likely to reflect underlying mutational biases, and this may be linked to GC bias.<sup>51,52</sup> GC,  $GC_1$ ,  $GC_2$ , and  $GC_3$  values were mapped onto phylogenetic trees using the packages *ape* v5.7-1<sup>53</sup> and *ggplot2* v3.3.6<sup>54</sup> in R. Phylogenetic trees were obtained from de Moya et al<sup>28</sup> and pruned to taxa in our data set.

### Phylogenetic signal

We tested for phylogenetic signal of GC and  $GC_3$  to see whether closely related feather lice taxa have similar GC content compared with more distant relatives. Analyses were done using the *phylosignal* v1.3<sup>55</sup> and *phylobase* v0.8.10<sup>56</sup> packages in R. We first measured global phylogenetic signal across the entire phylogeny with both Moran I and Abouheif  $C_{\text{mean}}$  with 100 simulations and 999 repetitions. Moran I and Abouheif  $C_{\text{mean}}$  are measures of spatial autocorrelation used to test the level of similarity of a characteristic between branch tips that are close in proximity.<sup>57,58</sup> Abouheif  $C_{\text{mean}}$  is a slight variation of Moran I by ignoring branch lengths and focusing on the mean of multiple topology possibilities with a weighted matrix of relatedness.<sup>59</sup> This accounts for any inaccuracies in the tree and when paired with Moran I can provide more confidence that any signal found is not reliant on the estimated branch lengths. In addition, Abouheif  $C_{\text{mean}}$  is a robust analysis that provides accurate statistics under many conditions (ie, polytomies or tree size<sup>60</sup>). Both indices calculate a value from  $-1$  to  $1$ ,



**Figure 1.** Gene assembly rate (A and C) and resource use efficiency (B and D) across an increasing number of CPUs (A and B) and aTRAM instances (C and D). An instance in this context represents a task, specifically the number of concurrently running aTRAM instances. Two different assemblers were used: ABySS (pink circles) and Trinity (blue triangles).

meaning the absence (-1) or presence (1) of similarities in a trait at certain phylogenetic distances. Second, because global measures of phylogenetic signal do not specify which lineages show trait similarities, we ran a Local Indicators of Phylogenetic Association (LIPA) analysis to identify local hotspots of autocorrelation.<sup>55</sup>

## Results

### Central processing units

Trinity and ABySS had similar patterns of gene assembly rate and resource efficiency, as measured by genes assembled per CPU per hour, although ABySS assembled genes faster than Trinity overall (Figure 1A and B). We found that aTRAM used computational resources most efficiently (ie, assembled the most genes per CPU) when given 1 CPU (Trinity: 7 genes/CPU/h—ABySS: 8 genes/CPU/h; Table 1). This is a common

outcome for most parallelization problems, as any increase in parallelization introduces overhead in the form of thread communication and synchronization. When addressing parallelization of aTRAM using tasks, we used the Slurm argument `-cpus-per-task=1` for all tests.

### Tasks

For both assemblers, 32 tasks resulted in the highest gene assembly rate and CPU efficiency (Trinity: 1.60 genes/min and 22.04% CPU efficiency; ABySS: 2.29 genes/min and 35.23% CPU efficiency; Table 2). Speedup  $\left(\frac{t(1)}{1(N)}\right)$  was calculated for each stepwise increase in the number of tasks, which measures relative improvement (eg, wall time) when increasing processing elements during execution of a program.<sup>61</sup> We found that ABySS had a faster speedup than Trinity for each test and is,

**Table 1.** Statistics from assembling genes with an increasing number of CPUs on a high-performance computing cluster using 2 different assemblers (Trinity and ABySS).

SCALING CPUS				
TRINITY VS ABYSS				
NUMBER OF CPUS	RUN TIME (MIN)	NUMBER OF GENES ASSEMBLED	GENE ASSEMBLY RATE (GENES/MIN)	RESOURCE EFFICIENCY (GENES/CPU/H)
Trinity				
1	120	14	0.12	7.00
2	120	23	0.19	5.75
4	120	43	0.36	5.38
8	110	64	0.58	4.36
16	68	64	0.94	3.53
32	47	64	1.36	2.55
ABySS				
1	120	16	0.13	8.00
2	120	26	0.22	6.50
4	120	50	0.42	6.25
8	87	64	0.74	5.52
16	51	64	1.25	4.71
32	31	64	2.06	3.87

Each row indicates a single run and how many CPUs were used. Resource efficiency was calculated by the number of genes assembled per available CPU per hour.

therefore, faster when scaling up with larger data sets. Although memory efficiency for both assemblers was quite low, ABySS had a much steeper increase with increasing tasks and Trinity memory efficiency seemed to peak at 16 tasks and began declining with 32 tasks (Figure 1C and D). Because we did not test with more than 32 tasks, our results suggest that we likely have not yet reached the parallelization plateau where synchronization and communication costs outweigh the advantages of parallel computation. Thus, a greater number of tasks would likely further increase assembly rate and efficiency.

#### Full data set assembly

Based on the results from the CPU and task scaling tests, we assembled all of the protein-coding genes from 57 feather lice taxa using the following Slurm arguments: `-nodes=1, -cpus-per-task=1, -ntasks=32, -cpus 32 (aTRAM)`, and `-abyss-np 32` for running ABySS as a parallel MPI job (`-abyss-np`). The average run time was 3.25 days, with a range between 1.38 and 6.25 days. The average number of loci assembled per taxon was 9055 with a range between 6637 and 12139. To assemble all of these loci, we used 1824 CPUs and 146477.04 CPU hours. For the 57 taxa, we assembled 516176 genes total. After removing the genes that did not pass the reciprocal best blast test, our final data set (used for all further analyses) included

499426 protein-coding genes with an average of 8761 genes per taxon.

Within our data set, 11780 genes assembled for at least 1 taxon and 7182 of those genes were assembled by a minimum of 54 taxa. We found that 1584 genes did not assemble for any taxon. Using the nucleotide sequences from the pigeon louse reference, we compared gene length between all of the reference genes ( $n=13364$ ), genes that we assembled for at least 1 taxon ( $n=11780$ ), and genes that did not assemble for any taxon ( $n=1584$ ). Data were log transformed to fit with assumptions of normality. Gene lengths between these groups were significantly different with a large effect size showing shorter genes were less likely to assemble (ANOVA:  $F_{2, 26726}=2270$ ,  $P<.001$ ,  $\eta^2=0.15$ ). Pairwise  $t$  tests with Bonferroni correction showed a significant difference between all group pairings ( $P\leq.001$ ; Figure 2).

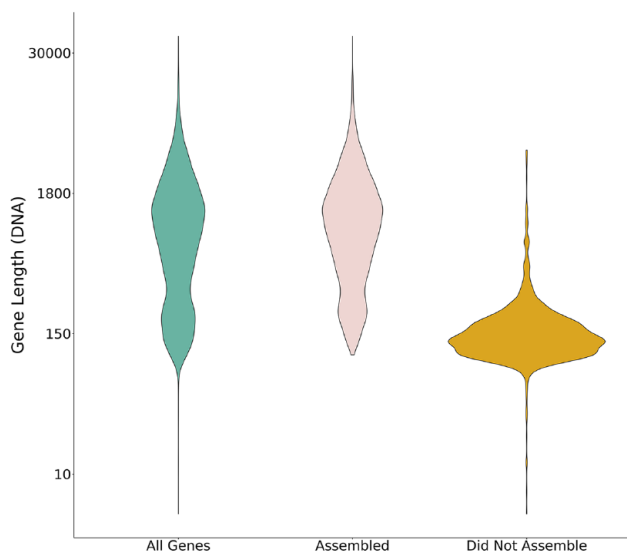
#### GC content

We measured GC content for all genes for each taxon ( $n=499426$ ), as well as GC content for each codon position ( $GC_1$ ,  $GC_2$ , and  $GC_3$ ). Data followed a normal distribution for  $GC$ ,  $GC_1$ , and  $GC_2$ . The distribution for  $GC_3$  was generally normal with a long right tail, indicating a disproportionate group of genes that are GC-rich. The range of GC content for

**Table 2.** Statistics from assembling genes with an increasing number of tasks on a high-performance computing cluster using 2 different assemblers (Trinity and ABySS).

SCALING TASKS: RESOURCE EFFICIENCY				
TRINITY VS ABYSS				
NUMBER OF TASKS	GENES PER MINUTE	SPEEDUP (T(1)/T(N))	CPU EFFICIENCY (%)	MEMORY EFFICIENCY (%)
Trinity				
1	0.18	1.00	1.58	0.42
2	0.27	1.56	3.05	0.37
4	0.50	2.82	5.73	0.48
8	0.75	4.28	10.24	0.55
16	1.08	6.17	15.15	1.42
32	1.60	9.10	22.04	1.1
ABySS				
1	0.19	1.00	1.52	0.35
2	0.31	1.63	2.92	0.34
4	0.56	2.95	5.66	1.27
8	0.97	5.09	11.00	2.51
16	1.52	8.00	18.36	5.00
32	2.29	12.00	35.23	9.98

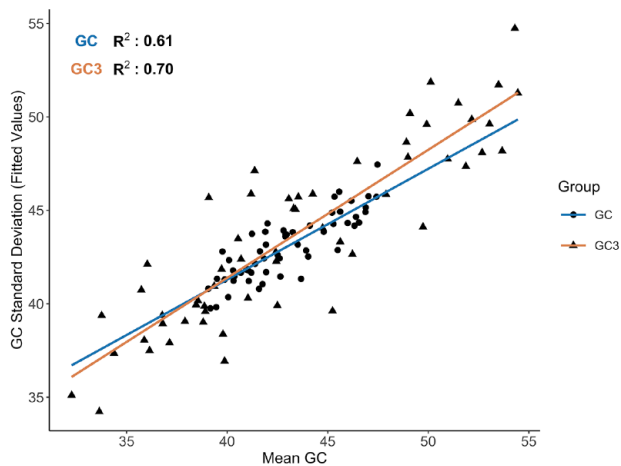
Each row indicates a single run and how many tasks were used. All runs were given 1 CPU per task based on the results from scaling CPUs. Speedup measures the relative improvement of the same run with differing resources. Central processing unit and memory efficiency were obtained from the workload manager (Slurm) output (CPU efficiency =  $\text{cpu\_time} / (\text{run\_time} \times \text{number\_of\_cpus})$ ; memory efficiency is the amount of allocated memory used in a run).



**Figure 2.** Differences in gene length (measured in DNA) from *Columbicola columbae* between 3 groups: all *C. columbae* protein-coding genes ( $n=13364$ ; green), genes that assembled for at least 1 taxon in our data set ( $n=11780$ ; pink), and genes that did not assemble for any taxon ( $n=1584$ ; yellow). The group of genes that did not assemble exhibited significantly shorter gene lengths compared with the other 2 groups. The presented data use logged values, with the y-axis indicating non-logged values.

all genes was 19.57% to 73.33% ( $M=42.96\%$ ;  $SD=6.53$ ; see Supplemental Data for all measures of GC content per taxon). Only 15% of all genes in our data set had a GC content of 50% or greater with most genes being GC-poor. Overall GC content was significantly different between taxa ( $n=499426$ ; ANOVA:  $F_{56, 496925}=1410$ ,  $P \leq .001$ ) with a large effect size ( $\eta^2=0.14$ ). To be sure the significance of the ANOVA was not due to only a few groups with significant differences, we ran a pairwise  $t$  test with Bonferroni correction and found a significant difference in GC content between 91% of groups ( $P \leq .05$ ). The mean GC content across all protein-coding genes for the pigeon louse reference ( $n=13364$ ) was 41.40% ( $SD=5.17$ ), similar to the GC content found in our data set. Both of these findings, however, are higher than GC content found across the entire pigeon louse genome, which is 36%.<sup>42</sup> For each of the 3 codon positions, mean GC content was 47.80% ( $SD=5.32$ ) for GC<sub>1</sub>, 37.70% ( $SD=5.96$ ) for GC<sub>2</sub>, and 43.37% ( $SD=14.15$ ) for GC<sub>3</sub>. As expected, there was much more variation found at GC<sub>3</sub> compared with GC<sub>1</sub> and GC<sub>2</sub>. The distribution of GC content per taxon for GC, GC<sub>1</sub>, GC<sub>2</sub>, and GC<sub>3</sub> is shown in Supplemental Figures 1 to 4.

We found a positive correlation between GC content and variation in GC content (measured as SD) for overall GC ( $r=0.61$ ,  $P \leq .001$ ; Figure 3) and for GC<sub>3</sub> ( $r=0.70$ ,  $P \leq .001$ ;



**Figure 3.** A strong positive correlation was found between both mean GC (blue, circles) and GC<sub>3</sub> (GC content at codon position 3; orange, triangles) and their respective SDs. Although accounting for phylogeny,  $R^2$  values were 0.61 for GC and 0.70 for GC<sub>3</sub>. Mean GC, GC<sub>3</sub>, and SD were calculated for all genes assembled per taxon. Each data point represents a distinct genus, and regression lines are depicted in blue (GC) and orange (GC<sub>3</sub>). The y-axis shows the fitted values calculated from the phylogenetic linear regression model.

Figure 3). Species with higher GC content have significantly more variation across their genes, whereas those that are more GC-poor seem to have a more consistent base composition. We found a weak positive relationship between GC content and gene length (ANOVA:  $F_{(1, 496,980)} = 4355$ ,  $P \leq .001$ ;  $\eta^2 = 0.0087$ ) among the genes. There is a slightly stronger positive relationship between GC and gene length for the pigeon louse reference; however, the effect size is still quite small (ANOVA:  $F_{(1, 13,362)} = 158.2$ ,  $P \leq .001$ ;  $\eta^2 = 0.01$ ).

#### GC content on tree and phylogenetic signal

GC content was mapped onto the subsampled phylogenetic tree from de Moya et al,<sup>28</sup> and in some cases, related taxa had similar GC levels (Figure 4). The pattern of GC content seen across the tree is also seen at all codon positions (Supplemental Figures 5 and 6), specifically for GC<sub>3</sub> (Supplemental Figure 7). This pattern suggests that transitions between GC-rich and GC-poor genes may have occurred multiple times across this group, as opposed to a more continuous change from older to more recently diverged taxa (see<sup>19</sup>).

Phylogenetic signal for GC and GC<sub>3</sub>, was not found using Moran I but was when using Abouheif  $C_{\text{mean}}$ . For overall GC, Moran I was insignificant and close to 0 ( $I = -0.0099$ ,  $P = .100$ ) whereas Abouheif  $C_{\text{mean}}$  was significant with a low positive signal ( $C_M = 0.190$ ,  $P = .020$ ). A similar pattern appeared for GC<sub>3</sub> with no global phylogenetic signal detected with Moran I ( $I = -0.010$ ,  $P = .077$ ) but this signal was detected with Abouheif  $C_{\text{mean}}$  ( $C_M = 0.194$ ,  $P = .015$ ). Although no significant phylogenetic signal was found regarding GC using Moran I at the global scale, the LIPA analysis found significant positive autocorrelation for 13 taxa using Moran I, as well as for 15 taxa

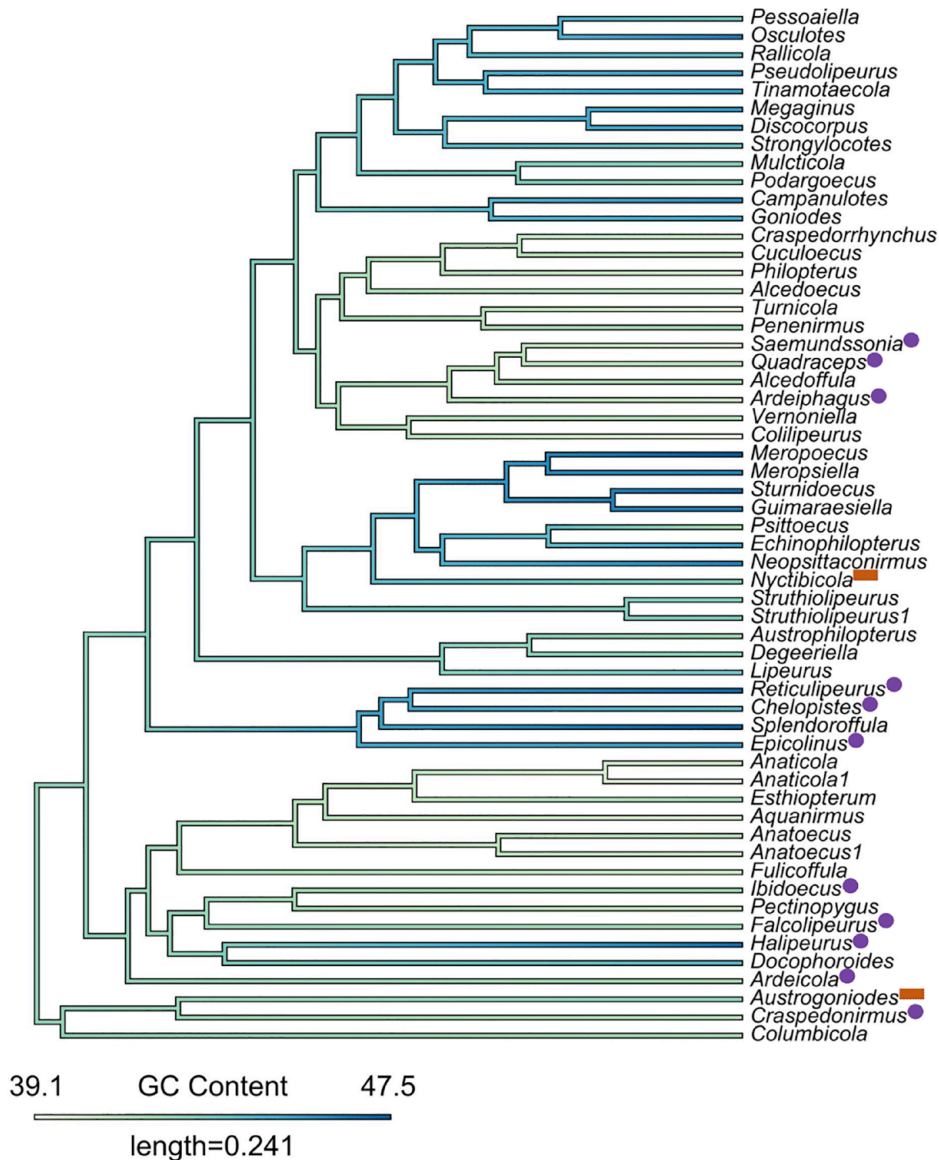
using Abouheif  $C_{\text{mean}}$  (Figure 4). Two of the taxa identified as having significant phylogenetic signal with Abouheif  $C_{\text{mean}}$  were negatively autocorrelated, whereas all of the taxa that exhibited significant phylogenetic signal using Moran I were positively autocorrelated. For GC<sub>3</sub>, the LIPA analysis for both Moran I and Abouheif  $C_{\text{mean}}$  revealed significant autocorrelation for the same 14 taxa; however, 2 of these were negatively autocorrelated with Abouheif  $C_{\text{mean}}$  (Supplemental Figure 7).

#### Discussion

Considering the disproportionately lower amount of genetic research on insects relative to their abundance and global distribution,<sup>20</sup> we set out to accomplish 2 main goals. First, we aimed to improve gene assembly efficiency with aTRAM by decreasing the amount of time and resources needed using parallelization with HPC, allowing for easier access to genetic data for these understudied groups. Second, we investigated the base composition of the largest set to date of protein-coding genes of feather lice. We found that while 32 CPUs assembled the most genes per minute, resource use was most efficient when using 1 CPU (Trinity: 7 genes/CPU/ hr | ABySS: 8 genes/CPU/h; Figure 1A and B). Using a single CPU, the number of tasks that offered the highest rate of genes assembled per minute, as well as the best CPU and memory efficiency was 32 tasks (Trinity: 1.6 genes/min | ABySS 2.29 genes/min; Figure 1C and D). By manipulating the resources given to aTRAM, we obtained a gene assembly speedup of 9.10 (Trinity) and 12 (ABySS; Table 2) times. This reduced the assembly time of a set of 13 364 genes from 48.72 to 3.8 days. As seen in our analysis, increasing the number of CPUs does not necessarily improve efficiency of a program (Figure 1B). This becomes increasingly complicated when addressing additional components of HPC, such as RAM, permanent disk space, and threading. In general, we recommend running similar tests on other software to gain a general understanding of a program's resource usage with an HPC system.

Our final data set included 499 426 genes and on average had a GC content below 50% (mean GC = 42.96%;  $SD = 6.53$ ). For third codon positions only, we found a similar average GC<sub>3</sub> to that found by Virrueta Herrera et al,<sup>32</sup> however, we found a higher average GC<sub>1</sub> and lower average GC<sub>2</sub>. We also found higher variation in GC content (19.57%–73.33%) across all genes. One explanation for the increased GC content in our data set compared with Baldwin-Brown et al<sup>43</sup> is the bias toward easily aligned genes. We retained only those genes that could be identified by reciprocal best-hit BLAST to the pigeon louse genome. These genes are less likely to contain repetitive sequences, and repetitive genome features are known to be GC-poor. By only retaining easily aligned genes, we have likely removed GC-poor genes from the data set.

Our results suggest feather lice may have higher GC content in coding sequences compared with many other insects that have been investigated. For example, GC content of protein-coding genes from 2 species of parasitoid wasp is around 30%



**Figure 4.** Phylogenetic tree (modified from de Moya et al<sup>28</sup>) with mean GC content mapped onto branches. Local hotspots exhibiting significant phylogenetic signal, as determined by the LIPA analysis, are indicated by asterisks. Purple circles represent a positive relationship, whereas orange rectangles denote a negative relationship. Darker shades of blue indicate higher GC levels, whereas lighter shades of green represent lower GC levels.

(*Aphidius ervi* and *Lysiphlebus fabarum*<sup>36</sup>) and between 33% and 39% for the honeybee (*Apis mellifera*<sup>62</sup>). Some species of insects do have similar GC content to feather lice, such as silkworms (*Bombyx*<sup>63</sup>). Unfortunately, the genetic studies that focus on insects are overrepresented by a small number of model species, such as members of *Drosophila* and *Lepidoptera*,<sup>20,64</sup> ignoring the enormous diversity of insects. To properly understand how GC content of feather lice compares to other insects, more studies are needed on non-model species.

For overall GC and GC<sub>3</sub>, we found a significant correlation between %GC for a species and SD among genes within that species (Figure 3), a pattern also seen in many vertebrates.<sup>11</sup> Species with higher values for GC and GC<sub>3</sub> showed much more variation compared with species with lower GC and GC<sub>3</sub>. This could indicate that GC-rich genes are being selected for in some species, resulting in more GC variation across their

genes. Alternatively, this could be a consequence of GC-biased gene conversion (gBGC), which favors the use of G and C bases.<sup>65</sup> This is particularly prominent in species with high rates of recombination. When an allelic mismatch occurs during the repair of a meiotic double-strand break, gBGC results in a biased frequency of G:C compared with A:T conversions, thus increasing the chance of GC substitution.<sup>10,66</sup> It is thought that gBGC plays a significant role in the variation in GC content within genomes of some taxonomic groups.<sup>67</sup> Many studies have found evidence that suggests gBGC plays a prominent role in the GC-rich base composition of avian<sup>68</sup> and mammalian isochores.<sup>69,70</sup> In addition, Pessia et al<sup>71</sup> examined the genomes of a broader range of organisms (Unikonts, Excavates, Chromalveolates, and Plantae) and found that gBGC is evident in most eukaryotic groups. It is important to note that disentangling gBGC from directional selection requires



further analysis. Although the mechanism is different, both result in a biased increase in 1 allele.<sup>72</sup> Because feather lice have low GC content and high rates of substitution, it may be that gBGC is not a strong influence on GC content in this group and other factors play a larger role. Alternatively, these results could indicate a shift toward higher GC heterogeneity, where different mechanisms result in GC-rich and GC-poor regions of the genome (eg,<sup>73</sup>).

Another prominent hypothesis that could also explain GC variation and evolution focuses on selection acting on the molecular machinery that results in nucleotide biases. Specifically, deamination of methyl-cytosine produces thymine causing a T:G mismatch requiring repair.<sup>74</sup> Methyl-cytosine deamination and GC content create a positive feedback loop that can either decrease or increase GC content in an organism.<sup>75</sup> In addition, some studies have found a positive correlation between GC content of transposable element (TE) and genome-wide GC content (eg,<sup>76</sup>). GC-rich TEs can increase GC content within the region of insertion if that region has a lower %GC than the TE, and the same rule would follow for GC-poor TEs. However, parasitic lice have a relatively low percentage of TEs in their genome (<5%)<sup>42,77,78</sup>. Interestingly, this fraction of genome-wide TEs is lower than that of silkworms (~40% of genome made up of TEs<sup>79</sup>), with whom feather lice share a similar average GC content.<sup>73</sup> Continuing to investigate the mechanisms driving GC content and the evolutionary consequences of GC-rich or poor genomes is critical for deepening our understanding of molecular evolution and, ultimately, species diversification.<sup>80,81</sup> To determine the driving force behind changes in base composition of feather lice among other organisms, future research needs to focus on a combination of synonymous vs non-synonymous substitutions ( $d_N/d_S$  ratios),<sup>82</sup> effective population size ( $N_e$ ),<sup>83</sup> and codon usage bias.<sup>84</sup>

Phylogenetic signal indicates groups of related organisms that may exhibit similar ecological or genetic traits<sup>85</sup> because of phylogenetic relatedness alone. We estimated the phylogenetic signal of GC content in feather lice, which revealed that close relatives had more similar GC content than would be expected by chance. Specifically, local phylogenetic signal of GC content was detected for overall GC (Figure 4) and GC<sub>3</sub> (Supplemental Figure 7). In addition, using Abouheif  $C_{mean}$ , we identified positive global phylogenetic signal of mean GC and GC<sub>3</sub> (GC:  $C_M=0.190$ ,  $P=.020$ ; GC<sub>3</sub>:  $C_M=0.194$ ,  $P=.015$ ). The LIPA analysis found 15 (GC) and 14 (GC<sub>3</sub>) species with significant autocorrelation, most being positive. Thus, GC content is likely a feature of phylogenetically closely related lineages. Because of this, we can have more confidence that base composition alone will not bias the construction of phylogenetic trees, a common concern in systematics.<sup>86</sup>

Feather lice genera have been grouped into different ecomorphs based on similar morphological characteristics. These specialized phenotypes allow them to live in different areas of

their avian host's body to escape host defense mechanisms<sup>24,87-89</sup> and found evidence of convergent evolution in these ecomorphs but it is unknown if each ecomorph type experiences the same selective pressures on the same genes or expresses similar genetic pathways. Signatures of convergent evolution have been found in other species that exhibit ecomorphs, which can tell us more about the genetic architecture of closely and distantly related organisms and the role selective pressures play in diversification.<sup>90,91</sup> Furthermore, these data provide new opportunities for future studies to explore insect protein evolution and adaptive evolution between feather lice genera and species.<sup>92</sup>

Scientists without a background in bioinformatics or computer science, however, often struggle with using the programs necessary for these evolutionary studies. Because this software often needs to be executed with HPC, researchers frequently hire specialists if they have available funds. This burden is much heavier on minorities, as they are less likely to receive funding<sup>93,94</sup>. These programs need to be built in a way that is more accessible and resource-efficient when used by scientists at large. Future development should increase focus on improving the program's ability to identify and allot available computational resources on various HPC architectures without the need for complex manipulation by the user. We have increased accessibility to aTRAM by developing a Singularity container that provides a more simple manner to install and use the program ([https://github.com/averygrant/atram\\_singularity](https://github.com/averygrant/atram_singularity)). Our aTRAM container consists of all of the files and dependencies needed for running aTRAM leading to fewer installation steps.

Genomic base composition is a fundamental feature of the genome of all organisms. Our results show that most feather louse protein-coding genes are GC-poor with the greatest variation found in GC<sub>3</sub>. However, GC content varies considerably between species. On average, lice have a higher GC content than other insects, although there is considerable variation in GC content among insect species. For example, feather lice GC content is similar to silkworms,<sup>63</sup> while being very different from the pea aphid (*Acyrtosiphon pisum*).<sup>33</sup> More detailed comparisons of GC content among insects will require examination of this feature at a genomic scale, ideally comparing orthologous genes. It is expected that GC content may vary between different gene regions,<sup>31,95</sup> such as nuclear vs mtDNA<sup>96,97</sup> and introns vs exons,<sup>98</sup> and this variation needs to be taken into account. As insect genomes become more available for a higher diversity of species, more research can focus on untangling the impact of molecular mechanisms and selective pressures on GC-rich or GC-poor organisms and, ultimately, shed light on protein evolution in insects.

## Acknowledgements

The authors thank Sebastian Smith and John Anderson for their help with the computational components of this publication.

## Author Contributions

ARG and JMA conceived and designed the study; ARG performed the formal analyses, created all visualizations, and wrote the original draft; all authors contributed to data collection, and reviewing and editing; ARG, JMA, and KPJ analyzed results; JMA supervised and funded the research; all authors approved the final manuscript.

## ORCID iD

Avery R Grant  <https://orcid.org/0000-0002-7623-4060>

## SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

## REFERENCES

- Wernegreen JJ. Ancient bacterial endosymbionts of insects: genomes as sources of insight and springboards for inquiry. *Exp Cell Res*. 2017;358:427-432. doi:10.1016/j.yexcr.2017.04.028
- Du MZ, Zhang C, Wang H, Liu S, Wei W, Guo FB. The GC content as a main factor shaping the amino acid usage during bacterial evolution process. *Front Microbiol*. 2018;9:2948. doi:10.3389/fmicb.2018.02948
- Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. Genome size diversity and its impact on the evolution of land plants. *Genes*. 2018;9:88. doi:10.3390/GENES9020088
- du Toit Z, du Plessis M, Dalton DL, Jansen R, Grobler JP, Kotzé A. Mitochondrial genomes of African pangolins and insights into evolutionary patterns and phylogeny of the family Manidae. *BMC Genomics*. 2017;18:746. doi:10.1186/s12864-017-4140-5
- Luo H, Thompson LR, Stingl U, Hughes AL. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol*. 2015;32:2738-2748. doi:10.1093/molbev/msv149
- Castillo AI, Nelson ADL, Lyons E. Tail wags the dog? Functional gene classes driving genome-wide GC content in *Plasmodium* spp. *Genome Biol Evol*. 2019;11:497-507. doi:10.1093/gbe/evz015
- Supple MA, Shapiro B. Conservation of biodiversity in the genomics era. *Genome Biol*. 2018;19:1-12. doi:10.1186/s13059-018-1520-3/FIGURES/3
- Cao J, Wu X, Jin Y. Lower GC-content in editing exons: implications for regulation by molecular characteristics maintained by selection. *Gene*. 2008;421:14-19. doi:10.1016/j.gene.2008.05.012
- Li X, Du D. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS ONE*. 2014;9:88339. doi:10.1371/journal.pone.0088339
- Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 2008;4:e1000071. doi:10.1371/journal.pgen.1000071
- Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*. 2010;20:1001-1009. doi:10.1101/GR.104372.109
- Matsubara K, Kuraku S, Tarui H, et al. Intra-genomic GC heterogeneity in sauropsids: evolutionary insights from cDNA mapping and GC 3 profiling in snake. *BMC Genomics*. 2012;13:604. doi:10.1186/1471-2164-13-604
- Mugal CF, Arndt PF, Holm L, Ellegren H. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3*. 2015;5:441-447. doi:10.1534/G3.114.015545/-/DC1
- Rao YS, Chai XW, Wang ZF, Nie QH, Zhang XQ. Impact of GC content on gene expression pattern in chicken. *Genet Sel Evol*. 2013;45:9. doi:10.1186/1297-9686-45-9
- Sémon M, Mouchiroud D, Duret L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum Mol Genet*. 2005;14:421-427. doi:10.1093/hmg/ddi038
- Niu Z, Xue Q, Wang H, et al. Mutational biases and GC-biased gene conversion affect GC content in the plasmodes of *Dendrobium* genus. *Int J Mol Sci*. 2017;18:2307. doi:10.3390/ijms18112307
- Eyre-Walker A, Hurst LD. The evolution of isochores. *Nat Rev Genet*. 2001;2:549-555. doi:10.1038/35080577
- Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet*. 2010;6:e1001107. doi:10.1371/JOURNAL.PGEN.1001107
- Šmarda P, Bureš P, Horová L, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci USA*. 2014;111:E4096-E4102. doi:10.1073/pnas.1321152111
- Hotelling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: where are we now? *Proc Natl Acad Sci USA*. 2021;118:e2109019118. doi:10.1073/pnas.2109019118
- Matoulek D, Ježek B, Vohnoutová M, Symonová R. Advances in vertebrate (cyto)genomics shed new light on fish compositional genome evolution. *Genes (Basel)*. 2023;14:244. doi:10.3390/genes14020244
- Grimaldi D, Engel MS. *Evolution of the Insects*. 1st ed. Cambridge University Press; 2005.
- Kunin WE. Robust evidence of declines in insect abundance and biodiversity. *Nature*. 2019;574:641-642. doi:10.1038/d41586-019-03241-9
- Price RD, Hellenthal RA, Palma RL, Johnson KP, Clayton DH. *Chewing Lice: World Checklist and Biological Overview*. Illinois Natural History Survey; 2003.
- Johnson KP, Shreve SM, Smith VS. Repeated adaptive divergence of microhabitat specialization in avian feather lice. *BMC Biol*. 2012;10:52. doi:10.1186/1741-7007-10-52
- Johnson KP, Allen JM, Olds BP, et al. Rates of genomic divergence in humans, chimpanzees and their lice. *Proc R Soc B*. 2014;281:20132174. doi.org/10.1098/rspb.2013.2174
- Clayton DH, Bush SE, Johnson KP. *Coevolution of Life on Hosts: Integrating Ecology and History*. University of Chicago Press; 2015.
- de Moya RS, Allen JM, Sweet AD, et al. Extensive host-switching of avian feather lice following the cretaceous-paleogene mass extinction event. *Commun Biol*. 2019;2:445. doi:10.1038/s42003-019-0689-7
- Huttener R, Thorrez L, In't Veld T, et al. GC content of vertebrate exome landscapes reveal areas of accelerated protein evolution. *BMC Evol Biol*. 2019;19:144. doi:10.1186/s12862-019-1469-1
- Johnston JS, Yoon KS, Strycharz JP, Pittendrigh BR, Clark JM. Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. *J Med Entomol*. 2007;44:1009-1012. doi:10.1603/0022-2585
- Yoshizawa K, Johnson KP. Changes in base composition bias of nuclear and mitochondrial genes in lice (Insecta: Psocodea). *Genetica*. 2013;141:491-499. doi:10.1007/s10709-013-9748-z
- Virrueta Herrera S, Sweet AD, Allen JM, Walden KKO, Weckstein JD, Johnson KP. Extensive in situ radiation of feather lice on tinamous. *Proc R Soc B Biol Sci*. 2020;287:20193005. doi:10.1098/rspb.2019.3005
- International Aphid Genomics Consortium. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol*. 2010;8:e1000313. doi:10.1371/journal.pbio.1000313
- Xue J, Zhou X, Zhang CX, et al. Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation. *Genome Biol*. 2014;15:521. doi:10.1186/s13059-014-0521-0
- Wang L, Tang N, Gao X, et al. Genome sequence of a rice pest, the white-backed planthopper (*Sogatella furcifera*). *Gigascience*. 2017;6:1-9. doi:10.1093/GIGASCIENCE/GIW004
- Dennis AB, Ballesteros GI, Robin S, et al. Functional insights from the GC-poor genomes of two aphid parasitoids, *Aphidius ervi* and *Lysiphlebus fabarum*. *BMC Genomics*. 2020;21:376. doi:10.1186/s12864-020-6764-0
- Sweet AD, Johnson KP, Cao Y, et al. Structure, gene order, and nucleotide composition of mitochondrial genomes in parasitic lice from Amblycera. *Gene*. 2021;5:145312. doi:10.1016/j.gene.2020.145312
- Dominguez Del Angel V, Hjerde E, Sterck L, et al. Ten steps to get started in genome assembly and annotation. *F1000Res*. 2018;7:149. doi:10.12688/f1000research.13598.1
- Tavares WFC, Roberto Miranda Assis M, Borin E. Quantifying detecting HPC resource wastage in cloud environments. Paper presented at: 2021 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW); October 26-29, 2021:41-46; Belo Horizonte. doi:10.1109/SBAC-PADW53941.2021.00017
- Allen JM, Boyd B, Nguyen N-P, et al. Phylogenomics from whole genome sequences using aTRAM. *Syst Biol*. 2017;66:786-798. doi:10.1093/sysbio/syw105
- Allen JM, Huang DI, Cronk QC, Johnson KP. ATRAM—automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*. 2015;16:98. doi:10.1186/s12859-015-0515-2
- Allen JM, LaFrance R, Folk RA, Johnson KP, Guralnick RP. aTRAM 2.0: an improved, flexible locus assembler for NGS data. *Evol Bioinform Online*. 2018;14:1176934318774546. doi:10.1177/1176934318774546
- Baldwin-Brown JG, Villa SM, Vickrey AI, et al. The assembled and annotated genome of the pigeon louse *Columbicola columbae*, a model ectoparasite. *G3*. 2021;11:jkab009. doi:10.1093/g3journal/jkab009

44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–2120. doi:10.1093/bioinformatics/btu170
45. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29:644–652. doi:10.1038/nbt.1883
46. Jackman SD, Vandervalk BP, Mohamadi H, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res*. 2017;27:768–777. doi:10.1101/gr.214346.116
47. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31. doi:10.1186/1471-2105-6-31
48. R Core Team. R: a language and environment for statistical computing. *R J*. <https://www.R-project.org/>
49. Ho LST, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol*. 2014;63:397–408. doi:10.1093/sysbio/syu005
50. Bofkin L, Goldman N. Variation in evolutionary processes at different codon positions. *Mol Biol Evol*. 2007;24:513–521. doi:10.1093/molbev/msl178
51. Kliman RM, Bernal CA. Unusual usage of AGG and TTG codons in humans and their viruses. *Gene*. 2005;352:92–99. doi:10.1016/j.gene.2005.04.001
52. Palidwor GA, Perkins TJ, Xia X. A general model of codon bias due to GC mutational bias. *PLoS ONE*. 2010;5:e13431. doi:10.1371/journal.pone.0013431
53. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35:526–528. doi:10.1093/bioinformatics/bty633
54. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2009.
55. Keck F, Rimet F, Bouchez A, Franc A. PhyloSignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol*. 2016;6:2774–2780. doi:10.1002/ece3.2051
56. Hackathorn R, Bolker B, Butler M, et al. PhyloBase: base package for phylogenetic structures and comparative data. R package version 0.8.10. Published 2010. <https://github.com/fmichonneau/phylobase>
57. Gittleman JL, Kot M. Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst Biol*. 1990;39:227–241. doi:10.2307/2992183
58. Abouheif E. A method for testing the assumption of phylogenetic independence in comparative data. *Evol Ecol Res*. 1999;1:895–909.
59. Pavoine S, Ollier S, Pontier D, Chessel D. Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. *Theor Popul Biol*. 2008;73:79–91. doi:10.1016/j.tpb.2007.10.001
60. Münkemüller T, Lavergne S, Bzeznik B, et al. How to measure and test phylogenetic signal. *Methods Ecol Evol*. 2012;3:743–756. doi:10.1111/J.2041-210X.2012.00196.X
61. Ristov S, Prodan R, Gusev M, Skala K. Superlinear speedup in HPC systems: why and when? *FedCSIS*. 2016;8:889–898. doi:10.15439/2016F498
62. Qi W, Yan C, Li W, et al. Distinct patterns of simple sequence repeats and GC distribution in intragenic and intergenic regions of primate genomes. *Aging*. 2016;8:2635–2654. doi:10.18632/aging.101025
63. Zhou QZ, Fang SM, Zhang Q, Yu QY, Zhang Z. Identification and comparison of long non-coding RNAs in the silk gland between domestic and wild silkworms. *Insect Sci*. 2018;25:604–616. doi:10.1111/1744-7917.12443
64. Kyriacou RG, Mulhair PO, Holland PWH. GC content across insect genomes: phylogenetic patterns, causes and consequences. *J Mol Evol*. 2024;92:138–152.
65. Eyre-Walker A. Recombination and mammalian genome evolution. *Proc R Soc B Biol Sci*. 1993;252:237–243. doi:10.1098/RSPB.1993.0071
66. Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet*. 2012;28:101–109. doi:10.1016/j.tig.2011.11.00
67. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 2009;10:285–311. doi:10.1146/annurev-genom-082908-150001
68. Bolívar P, Mugal CF, Nater A, Ellegren H. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system. *Mol Biol Evol*. 2015;33:216–227. doi:10.1093/MOLBEV/MSV214
69. Galtier N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet*. 2003;19:65–68. doi:10.1016/S0168-9525(02)00002-1
70. Kudla G, Helwak A, Lipinski L. Conversion and GC-content evolution in mammalian Hsp70. *Mol Biol Evol*. 2004;21:1438–1444. doi:10.1093/molbev/msh146
71. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol*. 2012;4:675–682. doi:10.1093/gbe/evs052
72. Borges R, Szöllösi GJ, Kosiol C. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*. 2019;212:1321–1336. doi:10.1534/genetics.119.302074
73. Jørgensen FG, Schierup MH, Clark AG. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. *Mol Biol Evol*. 2007;24:611–619. doi:10.1093/MOLBEV/MSL190
74. Shen JC, Rideout WM 3rd, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res*. 1994;22:972–976. doi:10.1093/nar/22.6.972
75. Fryxell KJ, Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*. 2000;17:1371–1383. doi:10.1093/oxfordjournals.molbev.a026420
76. Symonová R, Suh A. Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mob DNA*. 2019;10:49. doi:10.1186/s13100-019
77. Kirkness EF, Haas BJ, Sun W, et al. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci USA*. 2010;107:12168–12173. doi:10.1073/pnas.1003379107
78. Xu Y, Ma L, Liu S, et al. Chromosome-level genome of the poultry shaft louse *Menopon gallinae* provides insight into the host-switching and adaptive evolution of parasitic lice. *Gigascience*. 2024;13:1–15. doi:10.1093/gigascience/giae004
79. Xu HE, Zhang HH, Xia T, Han MJ, Shen YH, Zhang Z. BmTEdb: a collective database of transposable elements in the silkworm genome. *Database (Oxford)*. 2013;2013:bat055. doi:10.1093/database/bat055
80. Figueat E, Ballenghien M, Romiguier J, Galtier N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol*. 2015;7:240–250. doi:10.1093/gbe/evu277
81. Trávníček P, Čertner M, Ponert J, Chumová Z, Jersáková J, Suda J. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol*. 2019;224:1642–1656. doi:10.1111/nph.15996
82. Bolívar P, Guéguen L, Duret L, Ellegren H, Mugal CF. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol*. 2019;20:5. doi:10.1186/s13059-018-1613-z
83. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol*. 2012;4:658–667. doi:10.1093/gbe/evs027
84. Clément Y, Sarah G, Holtz Y, et al. Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet*. 2017;13:e1006799. doi:10.1371/journal.pgen.1006799
85. Jombart T, Pavoine S, Devillard S, Pontier D. Putting phylogeny into the analysis of biological traits: a methodological approach. *J Theor Biol*. 2010;264:693–701. doi:10.1016/j.jtbi.2010.03.038
86. Van Den Bussche RA, Baker RJ, Huelsenbeck JP, Hillis DM. Base compositional bias and phylogenetic analyses: a test of the “flying DNA” hypothesis. *Mol Phylogenet Evol*. 1998;10:408–416. doi:10.1006/mpev.1998.0531
87. Clay T. Some problems in the evolution of a group of ectoparasites. *Evolution*. 1949;3:279–299. doi:10.2307/2405715
88. Bush SE, Kim D, Reed M, Clayton DH. Evolution of cryptic coloration in ectoparasites. *Am Nat*. 2010;176:529–535. doi:10.1086/656269
89. Kolencik S, Stanley EL, Punnath A, et al. Parasite escape mechanisms drive morphological diversification in avian lice. *Proc Biol Sci*. 2024;29:20232665. doi:10.1098/rspb.2023.2665
90. Brewer MS, Carter RA, Croucher PJ, Gillespie RG. Shifting habitats, morphology, and selective pressures: developmental polyphenism in an adaptive radiation of Hawaiian spiders. *Evolution*. 2015;69:162–178. doi:10.1111/evo
91. McGlothlin JW, Kobiela ME, Wright HV, Kolbe JJ, Losos JB, Brodie ED 3rd. Conservation and convergence of genetic architecture in the adaptive radiation of anolis lizards. *Am Nat*. 2022;200:E207–E220. doi:10.1086/721091
92. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–338. doi:10.1146/annurev.genet.39.073003.114725
93. Ginther DK, Schaffer WT, Schnell J, et al. Race, ethnicity, and NIH research awards. *Science*. 2011;333:1015–1019. doi:10.1126/science.1196783
94. Carnethon MR, Kershaw KN, Kandula NR. Disparities research, disparities researchers, and health equity. *JAMA*. 2020;323:211–212. doi:10.1001/jama.2019.19329
95. Sparks ME, Hebert FO, Johnston JS, et al. Sequencing, assembly and annotation of the whole-insect genome of *Lymantia dispar* dispar, the European gypsy moth. *G3*. 2021;11:jkab150. doi:10.1093/g3journal/jkab150
96. Sadílek D, Urfus T, Vilimová J, Hadrava J, Suda J. Nuclear genome size in contrast to sex chromosome number variability in the human bed bug, *Cimex lectularius* (Heteroptera: Cimicidae). *Cytometry A*. 2019;95:746–756. doi:10.1002/cyto.a.23729
97. Li X, Yan L, Pape T, Gao Y, Zhang D. Evolutionary insights into bot flies (Insecta: Diptera: Oestridae) from comparative analysis of the mitochondrial genomes. *Int J Biol Macromol*. 2020;149:371–380. doi:10.1016/j.ijbiomac.2020.01.249
98. Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep*. 2012;1:543–556. doi:10.1016/j.celrep.2012.03.013