ABSTRACT


Gene conversion is the nonreciprocal exchange of genetic material between homologous chromosomes. Multiple lines of evidence from a variety of taxa strongly suggest that gene conversion events are biased towards GC-bearing alleles. However, in Drosophila the data have largely been indirect and unclear, with some studies supporting the predictions of a GC-biased gene conversion model and other data showing contradictory findings. Here we test whether gene conversion events are GC-biased in *Drosophila melanogaster* using whole genome polymorphism and divergence data. Our results provide no support for GC-biased gene conversion and thus suggest that this process is unlikely to significantly contribute to patterns of polymorphism and divergence in this system.
Population Genomic Analysis Reveals No Evidence for GC-Biased Gene Conversion in
_Drosophila melanogaster_

by
Matthew Craig Robinson

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Genetics

Raleigh, North Carolina
2014

APPROVED BY:

Nadia D. Singh
Committee Chair

Eric A. Stone

James Mahaffey

Corbin Jones
DEDICATION

I would like to dedicate this to my family for always supporting me.

“Matt, I don’t understand what the hell you do, but I am proud of you.”

~Last words from my grandfather
BIOGRAPHY

Every journey has a beginning and mine is no exception. I was born in May of 1982 in Durham, North Carolina; while I sometimes can’t recall every specific memory the home movies and pictures do a pretty good job of highlighting the more embarrassing moments. During middle school my dad got a new job and the family moved to Concord, North Carolina. I would finish middle school and high school in Concord, before moving back “home” to attend college at UNC-Chapel Hill. It was during my time at Chapel Hill that I decided to pursue a career in research. After graduation, I worked as a research technician for a few years before going back to school to get a Master of Science at East Carolina University. I came to NC State in August of 2010 to pursue a PhD in Genetics. During my tenure at NC State I have struggled with many different personal issues that have lead me on a journey of self-discovery. While I am graduating from NC State with a Masters and not a PhD, I have discovered that sometimes the journey is more meaningful than the destination. So I will graduate in August of 2014 with a Master of Science in Genetics with my head held high knowing that not only did I gain a wealth knowledge about genetics, but a deeper understanding of myself.
ACKNOWLEDGMENTS

“If I have seen further it is by standing on the shoulders of giants.”

~Sir Isaac Newton

None of what I have accomplished would have been possible without the help of others. I would like to thank my parents, brother, and sister for always supporting me and encouraging me to never settle for anything less than my best. Thanks mom for being what probably was too aggressive with school administration during grade school. Thanks dad for reminding me that contrary to what I think, I have not ruined my life; and to not take things so seriously (“whoo woo, go Dale!”) Thanks Tyler for showing me that you can overcome any challenges in your life and achieve your goals. Thanks Katie, I know you always have my back even if your bark is worse than your bite. A special thanks to Becky for always being able to dig me up from what is covering. While too numerous to mention here a big thanks to the rest of my family and all of my friends for all the support throughout my time at NCSU. I would like to thank my lab mates for all their help and support throughout the years and for making the lab a wonderful environment to be in. I would like to thank my committee members for all of the advice (both professional and personal) that they have offered over the years. Finally, I would thank my advisor Nadia; the knowledge I have gained from you transcends the information contained within this document and is something that I will take with me throughout my life.
# TABLE OF CONTENTS

LIST OF TABLES ................................................................................................. vi
LIST OF FIGURES ............................................................................................... vii

POPULATION GENOMIC ANALYSIS REVEALS NO EVIDENCE FOR GC-BIASED
GENE CONVERSION IN *DROSOPHILA MELANOGASTER* ......................... 1

  Introduction ........................................................................................................ 1
  Results and Discussion ...................................................................................... 2
  Materials and Methods ..................................................................................... 6
  References .......................................................................................................... 7

APPENDICES ....................................................................................................... 10

  Appendix A. Supplementary Figure 1 .............................................................. 11
  Appendix B. Supplementary Figure 2 ............................................................. 13
  Appendix C. Supplementary Figure 3 – Autosomes ................................. 15
  Appendix D. Supplementary Figure 4 – X ................................................. 17
  Appendix E. Supplementary Table 1 .............................................................. 19
  Appendix F. Supplementary Table 2 .............................................................. 21
LIST OF TABLES

APPENDICES

Appendix E. Supplementary Table 1 .................................................................19
Appendix F. Supplementary Table 2 .................................................................21
LIST OF FIGURES

POPULATION GENOMIC ANALYSIS REVEALS NO EVIDENCE FOR GC-BIASED GENE CONVERSION IN *DROSOPHILA MELANOGASTER*

Figure 1. Ratio of AT→GC/GC→AT polymorphisms for each sequence type for the North American and African populations.........................2

Figure 2. Q statistic as a function of sequence types for both the North American and African populations................................................3

Figure 3a. Q statistic as a function of recombination rate for the autosomes of the North American population .............................................3

Figure 3b. Q statistic as a function of recombination rate for the X chromosome of the North American population ........................................3

Figure 4a. $R_N/R_{GC}$ and $R_N/R_{AT}$ by sequence type for the autosomes and the X chromosome of the North American population for all changes4

Figure 4b. $R_N/R_{GC}$ and $R_N/R_{AT}$ by sequence type for the autosomes and the X chromosome of the North American population for transversions only .................................................................4

Figure 5a. Q as a function of GC content for the autosomes of the North American population.................................................................5

Figure 5b. Q as a function of GC content for the autosomes of the North American population.................................................................5

APPENDICES

Appendix A. Supplementary Figure 1...............................................................11
Appendix B. Supplementary Figure 2...............................................................13
Appendix C. Supplementary Figure 3 – Autosomes.................................15
Appendix D. Supplementary Figure 4 – X..................................................17
Population Genomic Analysis Reveals No Evidence for GC-Biased Gene Conversion in Drosophila melanogaster

Matthew C. Robinson,*1 Eric A. Stone,1 and Nadia D. Singh1
1Department of Biological Sciences, Program in Genetics, North Carolina State University
*Corresponding author: E-mail: microbins@ncsu.edu.
Associate editor: John H. McDonald

Abstract

Gene conversion is the nonreciprocal exchange of genetic material between homologous chromosomes. Multiple lines of evidence from a variety of taxa strongly suggest that gene conversion events are biased toward GC-bearing alleles. However, in Drosophila, the data have largely been indirect and unclear, with some studies supporting the predictions of a GC-biased gene conversion model and other data showing contradictory findings. Here, we test whether gene conversion events are GC-biased in Drosophila melanogaster using whole-genome polymorphism and divergence data. Our results provide no support for GC-biased gene conversion and thus suggest that this process is unlikely to significantly contribute to patterns of polymorphism and divergence in this system.

Key words: GC-biased gene conversion, Drosophila melanogaster, whole genome, polymorphism and divergence data.

Introduction

Gene conversion is the unequal exchange of genetic material between homologous loci and reflects one possible outcome of the recombination process. That is, with gene conversion events, one allele is converted to another due to mismatch repair following the repair of a double-strand break. Meiotic gene conversion events can lead to non-Mendelian segregation of gametes derived from the germ cell in which the gene conversion event occurred. If the repair process is unbiased (i.e., the two possible ways in which the mismatch can be repaired occur with equal probability), the gamete pool as a whole should be unbiased. However, if there is a systematic bias in the repair process toward GC for instance, then GC/AT heterozygotes should produce an excess of GC-bearing gametes, which will increase the probability of transmission of GC-bearing alleles to the next generation. In this way, biased gene conversion (BGC) can increase the frequency of GC alleles in natural populations (Duret and Galtier 2009; Webster and Hurst 2012).

There is mounting evidence that the process of gene conversion is biased toward GC-bearing alleles in a number of taxa. Most of this evidence comes from humans and yeast and much of this evidence is indirect. First, GC content and recombination rate are positively correlated in several taxa, as would be predicted under a model of GC-biased gene conversion (gBGC) (e.g., Fullerton et al. 2001; Marais et al. 2001; Birdsell 2002; Kong et al. 2002; Marais and Piganeau 2002; Meunier and Duret 2004; Webster et al. 2005; Duret and Arndt 2008; Muggle et al. 2011; Pessia et al. 2012). Second, data from several multigene families in humans, birds, yeast, and Arabidopsis are consistent with nonallelic gene conversion increasing GC content for multicopy genes (Galtier et al. 2001; Galtier 2003; Backstrom et al. 2005; Benyovz et al. 2005; although see McGrath et al. 2009; Assis and Kondrashov 2011). Third, polymorphism data from humans are consistent with gBGC, as AT→GC polymorphisms segregate at a higher frequency than GC→AT polymorphisms (Eyre-Walker 1999; Duret et al. 2002; Webster et al. 2003). This pattern is more pronounced in areas of high recombination (Spencer et al. 2006) as would also be predicted by a gBGC model. Fourth, patterns of base substitution in humans are consistent with a gBGC model (e.g., Duret et al. 2002; Webster et al. 2003; Meunier and Duret 2004; Berghound et al. 2009). Further indirect support for gBGC comes from evidence that mismatch repair is generally GC-biased (Birdsell 2002; Mancera et al. 2008) and also base excision repair appears to be GC-biased in many taxa including humans (Brown and Jiricny 1989; Bill et al. 1999; reviewed in Marais 2003). Finally, high-resolution mapping of meiotic products in yeast reveals that meiotic gene conversion events are indeed GC-biased in this system (Mancera et al. 2008).

Data from Drosophila have been less clear, with some studies showing evidence in support of gBGC whereas others show no such evidence. GC content and crossover rate are positively correlated in Drosophila (Marais et al. 2001; Singh et al. 2005; Liu and Li 2008). However, GC content and noncrossover rate are not significantly correlated (Cameron et al. 2012). Polymorphism data from noncoding sequences do indeed suggest that AT→GC polymorphisms segregate at a higher frequency than GC→AT polymorphisms in D. melanogaster (Galtier et al. 2006), and the relative proportions of these mutations are correlated with GC content in sister species D. simulans (Haddrell and Charleworth 2008) as expected under a gBGC model. In contrast, however, no such bias in polymorphism patterns is found in noncoding sequences of D. miranda (Bartolome et al. 2005). Similarly equivocal results are found using divergence data; an early divergence-based study from a variety of Drosophila lineages was inconsistent with gBGC (Ko et al. 2006) although a recent comparative genomic analysis is strongly consistent with
gBGC in the D. melanogaster species group (Capra and Pollard 2011). Finally, analyses of nonallelic gene conversion events have provided no evidence in support of gBGC (Casola et al. 2010; Assis and Kondrashov 2011).

It thus remains an open question whether gene conversion is GC-based in Drosophila. Here, we address this question explicitly, using polymorphism data from D. melanogaster and divergence data from D. simulans. We test several predictions of the gBGC model with respect to the numbers of AT → GC versus GC → AT polymorphisms, the frequencies of these mutations, and the relationship between polymorphism frequency and crossover rate. We also test whether contrasting patterns of polymorphism and divergence provide any support for a fixation bias of GC-bearing alleles. Our results suggest that gBGC does not markedly contribute to patterns of polymorphism and divergence in D. melanogaster.

Results and Discussion

Ratio of Polymorphisms

As our first test of whether gene conversion is GC-based in D. melanogaster, we examined counts of AT → GC versus GC → AT polymorphisms. Assuming base composition is at equilibrium, the null expectation is that these two classes of mutation should be equally represented in the genome. In contrast, under a model of gBGC, we expect an excess AT → GC over GC → AT polymorphisms. Following previous examples (e.g., Eyre-Walker 1997; Galtier et al. 2001; Haddrill and Charlesworth 2008), we tested this prediction using polymorphism data from the Drosophila melanogaster Genetic Reference Panel (DGRP) (Mackay et al. 2012). Because the effects of gBGC on patterns of polymorphism and divergence will depend in part on the nature and extent of selective constraint, we restricted our analyses to noncoding sequences. Given the heterogeneity of constraint even among noncoding sequence types (e.g., Haddrill, Charlesworth et al. 2005; Halligan and Keightley 2006; Singh et al. 2009), we partitioned the noncoding sequences by sequence type to control for the degree of selective constraint. In particular, we examined three types of noncoding sequences on the X chromosome and the autosomes: short introns, long introns, and intergenic sequences. For ease of presentation, we report the ratio of AT → GC/GC → AT polymorphisms and note that the null expectation is a ratio of 1 while a ratio > 1 would be consistent with gBGC (fig. 1). In every category, we see a significant deviation from the expected 1:1 ratio of AT → GC versus GC → AT polymorphisms (P < 0.001, all comparisons, \( \chi^2 \) test). However, in all cases, the deviation appears to be due to an excess of GC → AT polymorphisms, which is precisely the opposite of what would be predicted under a gBGC model. This pattern is consistent with recent findings by Poh et al. (2012), who showed a large excess of GC → AT polymorphism for noncoding regions in D. melanogaster.

In addition, the magnitude of the deviation from the null appears similar across sequence types (fig. 1), which would not be predicted under a BGC model. That is, a BGC model might predict that the effects of BGC should be particularly pronounced in neutrally evolving sequences and perhaps less pronounced in sequences under greater functional constraint. Thus, we expect that the least constrained sequences in our data set, that is, short introns, should show the greatest effect of BGC, whereas long introns would show the least (Haddrill, Charlesworth et al. 2005; Halligan and Keightley 2006; Singh et al. 2009). This could manifest as a greater deviation from the null expectation for short introns and the smallest deviation for long introns. This contrasts with our observed results, which indicate similar deviations from expectation in all sequence categories, particularly on the X chromosome (fig. 1).

Moreover, the deviation from expectation under base composition equilibrium may not in fact be due to BGC. Indeed, it has long been suggested that base composition is not at equilibrium in Drosophila (e.g., Kern and Begun 2005; Singh et al. 2005; Ko et al. 2006; Singh et al. 2007; Keightley et al. 2009; Singh et al. 2009; Clemente and Vogl 2012). This deviation from base composition equilibrium may be due to shifting mutational processes or other evolutionary processes that are independent of any potential bias in gene conversion (Singh et al. 2009). Thus, the polymorphism count data in the DGRP are indicative of nonequilibrium base composition, but are inconsistent with gBGC driving this nonequilibrium base composition. Instead, given nonequilibrium base composition, these data may reflect the AT bias of the mutational process in Drosophila (Petrov and Hartl 1999; Haddrill and Charlesworth 2008; Keightley et al. 2009).

Site Frequency Spectrum

As a second test of gBGC, we analyzed the site frequency spectrum (SFS) for each sequence type across both sets of chromosomes. Everything else being equal, under a model of gBGC we expect an excess (relative to neutrality) of high-frequency derived polymorphisms in the case of GC-increasing alleles, and we also expect a dearth of high-frequency derived AT-increasing alleles relative to neutral expectation. To control for selective constraint, we partition...
suggests an excess of high frequency GC-increasing alleles; larger than a randomly sample value from In a comparison of two distributions Whitney frequency spectra, we used a scaled version of the Mann–Whitney U statistic as a function of sequence types for both the North American and African populations. Note that the x axis crosses the y axis at 0.5, which is the null expectation of no difference between the site frequency spectra of GC$\rightarrow$AT versus AT$\rightarrow$GC polymorphisms; values exceeding this value are consistent with gBGC. P values are based off the empirical distribution.

To quantify the difference between these two site frequency spectra, we used a scaled version of the Mann–Whitney U statistic we call Q (see Materials and Methods). In a comparison of two distributions X and Y, Q estimates the probability that a randomly sampled value from Y will be larger than a randomly sample value from X. Here, Q > 0.5 suggests an excess of high frequency GC-increasing alleles whereas a Q of 0.5 indicates that neither GC- nor AT-increasing alleles tend to be at higher frequency than the other. Estimates of Q for each sequence type are presented in figure 2. In all cases, the Q statistic is not significantly greater than 0.5 as would be predicted if gene conversion in D. melanogaster were GC-biased. To the contrary, Q does not significantly deviate from the empirical distribution in all cases for all sequence types, for both the X and the autosomes (P > 0.07, all comparisons). Moreover, the estimated Q statistics appear similar in magnitude across sequence types (fig. 2).

As suggested above, this similarity among sequence types is not predicted by a BGC model, as the effects of BGC should arguably be more pronounced in less constrained sequences. Thus, echoing the results based on polymorphism counts in the DGRP, the site frequency spectra of GC- and AT-increasing alleles in the DGRP appear inconsistent with gBGC.

**Recombination Rate**

One additional prediction of the gBGC model is that, everything else being equal, the magnitude of the bias in gene conversion should be greatest where rates of gene conversion are highest. Indeed, previous work in humans has convincingly demonstrated that the effects of gBGC are most pronounced in regions of high recombination (Dreszer et al. 2007; Katzman et al. 2011). We sought to test this prediction of the gBGC model in D. melanogaster by partitioning SNPs by recombination rate and comparing the SFS of GC-enriching versus AT-enriching mutations within each recombination rate category using the Q statistic. The gBGC model predicts 1) an excess of GC- versus AT-enriching mutations which would manifest as Q > 0.5 for a given recombination rate and 2) that the magnitude of this excess should increase monotonically as a function of recombination rate. These predictions are contingent on consistency in selective pressures; as with previous analyses, we attempt to control for varying selective constraint across the genome by partitioning the data by sequence type. In Drosophila, crossover rate has generally been used as a proxy for overall recombination rate and thus we use crossover rates as a proxy for recombination rate in this analysis.

Our results are presented in figure 3 and clearly show that these predictions are not met. In 16/18 categories, Q did not significantly deviate from the empirical distribution (P > 0.08, all comparisons). The two exceptions are autosomal long and short introns in low recombination (P = 0.01 and P = 0.003, respectively). In both cases, Q > 0.5, which is consistent with gBGC. However, estimates of Q overall show no consistent relationship with recombination rate, which is a prediction of the gBGC model.

It is important to note that both crossovers and noncrossovers are associated with gene conversion tracts, and both
types of recombination events have the potential to be associated with gBGC (see Duret and Galtier 2009). However, recent evidence from yeast suggests that gBGC is only associated with crossovers (Lesecque et al. 2013). If this is also the case in Drosophila, then as stated earlier the gBGC model predicts a positive correlation between the degree of GC bias and recombination rate. If, however, noncrossover events predominantly associate with gBGC in Drosophila, then the gBGC model would predict an inverse relationship between recombination rate and the degree of bias. This is because recent evidence indicates that rates of noncrossovers and rates of crossing over are negatively correlated in D. melanogaster (Comeron et al. 2012). Our results are largely inconsistent with a negative correlation between estimates of Q and recombination rate (fig. 3). Autosomal intronic sequences show a qualitative hint of a negative correlation between recombination rate and the degree of bias (fig. 3), but this is should be interpreted with caution because the estimates of the Q statistic for both medium and high recombination rates are 1) not significantly different from expectation and 2) in the direction opposite to the gBGC prediction in three out of four cases (fig. 3). If both crossovers and noncrossovers contribute to gBGC in Drosophila, it would be difficult to predict the expected relationship between the estimated degree of bias and recombination rate in this system, given the negative correlation between the rates of these types of events. However, the observation that the Q statistic itself is largely not significantly different from the null expectation in any recombination rate category argues against the action of gBGC in D. melanogaster.

Fixation Bias
As a final test for gBGC in D. melanogaster, we used polymorphism and divergence data to test for a bias in fixation probabilities specifically for GC-increasing alleles. Borrowing on the intuition from the McDonald–Kreitman test (McDonald and Kreitman 1991), we compared the counts of polymorphic and divergent GC-increasing alleles to such counts of mutations unaffected by potentially BGC, GC→CG and AT→TA. If gene conversion is indeed GC-biased and GC-bearing alleles have increased fixation probabilities, then we expect an excess of divergent GC-increasing mutations. Thus, the ratio of polymorphism to divergence for GC-increasing mutations (RGC) should be lower than that ratio for "neutral" (GC→CG and AT→TA) mutations (R0). In contrast, AT-increasing alleles should show a dearth of fixation relative to neutrality resulting in the ratio of polymorphisms to divergence for mutations of this class (RA) exceeding R0. Thus, under a gBGC model, we expect RGC/RA > 1 and R0/RGC < 1.

This contrasts dramatically with our results (fig. 4a), which indicate that the ratio of polymorphisms to divergence for GC-increasing mutations (RGC) significantly exceed 1 for all sequence classes (P < 0.001, all comparisons, G-test) except short X-linked introns (P = 0.13) and RGC > 1 and R0/RGC < 1.

This is consistent with previous work (Poh et al. 2012). Because our "neutral" changes (GC→CG and AT→TA) are all transversions and our GC-increasing (AT→GC) and AT-increasing (GC→AT) changes are a mixture of transitions and transversions, we repeated this analysis only considering transversions to ensure that our results were not driven by the difference in the types of mutations comprising the "neutral" versus the "nonneutral" classes. Similar to what was found when considering all changes, our results indicate that the ratio of polymorphism to divergence for GC-increasing mutations (RGC) significantly exceeds 1 for all sequence classes (P < 0.001, all comparisons, G-test) except long X-linked introns where RGC/RA > 1 and R0/RGC < 1.

Thus, both when considering transitions and transversions together and when considering transversions alone, these polymorphism and divergence data provide no evidence in support of a fixation bias unique to GC-enriching mutations as would be expected under a gBGC model.

Robustness
Our analysis of polymorphism within D. melanogaster and divergence between D. melanogaster and D. simulans provides compelling evidence that gene conversion in D. melanogaster...
is not GC-biased. However, there are several confounding factors intrinsic to our analysis; here we systematically explore them to assess the robustness of our results. First, because the DGRP is derived from a North American population, it is possible that demographic effects have compromised our ability to detect the effects of gBGC. Indeed, population bottlenecks which have clearly played a role in the demographic history of North American populations of D. melanogaster [Begun and Aquadro 1993; Baudry et al. 2004; Haddrill, Thornton et al. 2005; Duchen et al. 2013] have marked effects on levels of polymorphism and divergence as well as the site frequency spectra of segregating mutations (e.g., Wall et al. 2002; Haddrill, Thornton et al. 2005; Duchen et al. 2013). This could in principle mask the influence of gBGC.

To investigate this possibility, we repeated a suite of our analyses using polymorphism data from an African population. Africa is the presumed origin of this species (Lachaise et al. 1988; reviewed in Stephan and Li 2006), and thus, patterns of polymorphism in this ancestral range are likely be less compromised by demographic effects. Note that these African sequence data are restricted to a 2.1 Mb X chromosome fragment (see Materials and Methods). We first examined polymorphism counts for AT→GC versus GC→AT mutations. As was the case with the North American data, we see a significant departure from expectation under base composition equilibrium (P < 0.001, all comparisons, χ² test) (fig. 1). Also consistent with what was observed with the North American polymorphism data, the direction of the deviation from expectation is contrary to the predictions of the gBGC model.

We also examined the site frequency spectra of GC-versus AT-enriching mutations in this African sample (fig. 2). As was the case with the North American sample, the predictions of the gBGC model are largely not supported by these data. We see no evidence in support of a general, marked excess of high-frequency derived GC-enriching relative to AT-enriching mutations in the African population data set. Estimates of Q do not significantly deviate from null expectation for any sequence type (P > 0.19, all comparisons), which is contrary to the predictions of the gBGC model.

When coupled with the polymorphism count data, these SFS data from Africa largely recapitulate the patterns observed in our North American sample and fail to provide compelling evidence of gBGC in D. melanogaster. Thus, the lack of evidence in support of gBGC in the North American data does not appear to be driven by a lack of power given potentially confounding demographic effects.

Another possible confounding factor in our analysis is that of CC content. Counts of polymorphisms are particularly susceptible to such a bias given nonequilibrium base composition in D. melanogaster (e.g., Kern and Begun 2005; Singh et al. 2005; Ko et al. 2006; Singh et al. 2007; Knightley et al. 2009; Singh et al. 2009; Clemente and Vogl 2012), one might expect that CC-rich regions might have an excess of AT-enriching mutations over GC-enriching ones simply due to mutational opportunity. Thus, failing to account for mutational opportunity might obscure the signal of gBGC.

To control for the potential effects of mutational opportunity, we partitioned our data by CC content and repeated the SFS analysis. Our results are presented in figure 5. In 16/18 cases, the estimated Q values are not significantly different from the null expectation (P > 0.06, all comparisons). Two autosomal categories did deviate significantly from the empirical distribution: short introns in low and high GC content (P < 0.003, both comparisons). However, in the case of low GC content short autosomal introns, the deviation from expectation is again in the direction that is contrary to what is predicted by a gBGC model.

We thus suggest that our analysis of polymorphism in the context of genomic variation in CC content fails to provide evidence that gBGC contributes substantially to patterns of polymorphism and divergence in D. melanogaster. For the vast majority of CC content categories and sequence types, the predictions of the model are not met. Although there are a few exceptions to this pattern (fig 5), they are sporadic and we believe it is unlikely that gBGC would generate such an unpredictable pattern. Together, these two suites of robustness analysis convincingly demonstrate that the lack of evidence in support of gBGC in D. melanogaster is not due to confounding factors such as demography or nonequilibrium base composition.

It should be noted that we cannot exclude the possibility that the process of gene conversion is mechanically GC-biased in Drosophila, but that its effects on genome evolution are either too minor to be detected or are overwhelmed by
other evolutionary forces. However, given the preponderance of evidence contrary to the predictions of the gBGC model and the apparent robustness of our results, our analysis clearly indicates that even if gene conversion is GC-biased in D. melanogaster, this bias is unlikely to contribute to patterns of polymorphism and divergence in this system.

Further Considerations

Although the goal of this study was to test the hypothesis that gene conversion is GC-biased in D. melanogaster, several additional aspects of our data merit discussion. First, it is very important to note that the signatures of gBGC tested here are indistinguishable from the signatures of selection for increased GC content. For instance, both gBGC and selection favoring GC-bearing alleles would result in an excess of high-frequency derived GC-increasing alleles as well as increased fixation probabilities of GC-increasing alleles. As a consequence, our data are similarly inconsistent with natural selection systematically favoring increased GC content in D. melanogaster.

However, it remains possible that there is selection on GC content that is sequence-type specific. It is clear that patterns of single nucleotide substitution in Drosophila vary as a function of sequence type (e.g., Singh et al. 2009), and this may be driven in part by selection on sequence-type-specific GC content. This could manifest in part as selection on local GC content, which has been observed in Drosophila (Kenigsberg and Tanay 2013). Selection on sequence-specific GC content would yield differential fixation of GC-increasing versus AT-increasing alleles among sequence types and might thus contribute to the variation in the R_{GC}/R_{AT} and R_{GC}/R_{AT} observed among sequence classes (fig. 4), for instance, although we note that this is speculative.

With respect to the ratio of AT→GC versus GC→AT polymorphisms, it is perhaps unsurprising that on the X and autosomes, intergenic and long introns sequences show similar values (fig. 1) given that these sequences appear subject to similar levels of constraint in Drosophila (e.g., Bergman and Kreitman 2001). What is surprising, however, is the difference in the short intron AT→GC/GC→AT ratio between the X and the autosomes. On the X chromosome (in both North America and Africa), the AT→GC/GC→AT ratio for short introns is quite similar to the ratio observed for other X-linked sequences (fig. 1). On the autosomes, the AT→GC/GC→AT ratio is quite different and is in fact much closer to the expected equilibrium value. This is consistent with previous work based on divergence data indicating that among intronic and coding sequences, short introns are closest to base composition equilibrium (Singh et al. 2009). This in turn suggests that the pattern observed here on the X is contrary to expectation and merits explanation. One possible explanation is a recent shift in mutational patterns specific to the X-chromosome. This shift would be most apparent in short introns, which largely evolve in a manner that is consistent with neutrality (Haddrill, Charlesworth et al. 2005; Halligan and Keightley 2006; Singh et al. 2009). This could thus potentially explain why AT→GC/GC→AT differs between the X and the autosomes especially for short introns and why X-linked short introns are far from equilibrium, contrary to expectation.

Finally, previous work has shown a recombination-associated substitutional bias toward GC (Singh et al. 2005). That is, the rate of GC-enriching substitutions increases with increasing recombination. In principle, such a bias could result from a bias in the fixation process, perhaps mediated by selection or gBGC, or from a bias in the mutational process. The results presented here are inconsistent with a role of gBGC or natural selection in generating this pattern, as we see no evidence of a fixation bias toward GC-bearing alleles that increases with increased recombination. This previously reported pattern may therefore reflect a bias in the mutational process; this will be a topic of future investigation.

Materials and Methods

Polymorphism Data

The polymorphism data used in this study come in part from the Drosophila Genetic Reference Panel (DGRP) (Mackay et al. 2012). This sample corresponds to 162 inbred lines derived from a Raleigh, NC population sample. Freeze 1 SNPs identified using the joint Genotype for Inbred Lines (JGIL) (Stone 2012) were used for our analyses.

In addition to the DGRP, we used sequence data from 20 lines of Drosophila melanogaster from an African (Ugandan) population (for line details see Pool and Aquadro [2006]). These sequence data are restricted to a 2.1 Mb region along the X chromosome between the genes garnet and scalloped (coordinates 13,621,237–15,721,714 on chromosome X [release 5.37]) and were generated using a combination of targeted-enrichment and Illumina sequencing (Singh et al. 2013). SNP calls for this data set were also done with JGIL (Stone 2012) and have been reported previously (Singh et al. 2013).

We retained only bi-allelic SNPs for our analyses. To polarize these SNPs, we reconstructed the ancestral state of each polymorphism using all species whole-genome alignments of D. melanogaster and three additional species: D. simulans, D. erecta, and D. yakuba. These alignments were graciously provided by R. Kulathinal. For ancestral state reconstruction, we required bases to be called in at least D. simulans in addition to D. erecta and/or D. yakuba (data from both D. erecta and D. yakuba were used if bases were called in both of these species). Only those SNPs where the ancestral state could be unambiguously inferred using parsimony were retained for analysis. Positions in which bases called in an outgroup did not match either of the allelic states in D. melanogaster were excluded. We required at least two outgroups for ancestral reconstruction given that doing so nearly completely eliminates the misinference of ancestral state due to mutation rate variation that can accompany parsimony-based inferences of ancestral state (Hernandez et al. 2007). Ancestral state reconstruction was performed using custom python scripts. All python scripts are available upon request.

Polarized SNPs were further filtered to exclude positions with missing data. Thus, every SNP position considered in this study had a base called in every line in the population sample...
from which it was drawn as well as D. simulans and at least one additional outgroup. SNP filtering was performed using custom python scripts. All python scripts are available upon request.

Genomic Context
To explore gBGC across the genome, we partitioned the genome into six different sequence types: short introns (<100 bp), long introns (>100 bp), and intergenic regions on the autosomes and the X chromosome based on annotations from Flybase (release 5.37). SNPs with more than one annotation were discarded from our analysis.

We also partitioned SNPs by two other features of genomic context: GC content and recombination rate. The GC content of SNPs was calculated as the GC content of the intergenic or intronic region in the reference genome in which SNP was found. Recombination rates were estimated using the Drosophila recombination rate calculator (Singh et al. 2005; Fiston-Lavié et al. 2010) using the midpoint of a given (intronic or intergenic) sequence.

SNPs for each sequence type were then sorted into bins of high, medium, or low recombination rate or GC content. For both recombination rate and GC content, we used specific threshold values between bins to ensure that each bin contained approximately the same number of SNPs. The sorting of sequences by sequence type, chromosome, GC content, and recombination rate was performed using custom python scripts. All python scripts are available upon request.

Comparing Site Frequency Spectra
Comparisons between site frequency spectra were conducted using a statistic we denote as “Q.” The Q statistic is simply a scaled version of the Mann–Whitney U statistic whose values are constrained between 0 and 1. We employed statistical tests using Q in analogy to the U-tests performed in Kozak et al. (2011).

Consider independent observations from a distribution X and independent observations from a distribution Y. Then, we compute Q as

\[ Q = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[ \mathbb{I}(X_i < Y_j) + \frac{1}{2} \mathbb{I}(X_i = Y_j) \right] = \frac{U}{mn}. \]

In words, Q is the tie-adjusted fraction all of \((X_i, Y_j)\) pairs for which \(X_i < Y_j\). Thus, if there is no tendency for samples from Y to be larger for samples from X, the expectation is \(Q = 0.5\). We tested against this null hypothesis using a permutation-based approach in which 1,000 samples from the empirical null distribution were generated. For each, we sampled without replacement observations from the values Q was computed by comparing these observations to the remaining.

Calculation and significance testing of the Q statistic were performed using a custom python script. All python scripts are available upon request.

Fixation Bias
We tested for a fixation bias for GC-increasing alleles using polymorphism and divergence data. Following the McDonald–Kreitman (McDonald and Kreitman 1991) framework, we compared the ratios of AT→GC polymorphism/divergence versus “neutral” polymorphism/divergence. In the absence of a fixation bias, the polymorphism/divergence ratios should be equal between AT→GC and “neutral” sites. If there is fixation bias toward GC-increasing alleles, then there will be an excess of divergent AT→GC changes relative to neutral expectations; this would manifest as a polymorphism/divergence ratio of the AT→GC less than the polymorphism/divergence ratio of the “neutral” class. We can test whether these ratios are significantly different using a \(\chi^2\) test. For our “neutral” class, use pooled AT→T/A and GC→G changes. Analysis for fixation bias was performed using a custom python script. All python scripts are available upon request.

Supplementary Material
Supplementary figures S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors gratefully acknowledge Rob Kulathinal who kindly provided the multi-species whole-genome alignments of D. melanogaster and three additional species: D. simulans, D. erecta, and D. yakuba. They also thank C. Jones, J. Mahaffey, and D. Lawrie for helpful discussions of this project. Finally, they thank the two anonymous reviewers for their helpful comments and suggestions on this manuscript.

References


APPENDICES
Appendix A

Supplementary Figure 1

A.  

B.  

Derived allele frequency

Derived allele count (bin size = 7)

Short intron AT->GC
Short intron GC->AT
Neutrality

Long intron AT->GC
Long intron GC->AT
Neutrality
Supplementary Figure 1. Site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the autosomes (a) Short intron, b) Long intron, c) Intergenic). The x-axis is the number of derived alleles at a given SNP grouped into bins (bin size = 7). The y-axis is the frequency of the total number of derived alleles within each bin.
Appendix B

Supplementary Figure 2

A.

B.
Supplementary Figure 2. Site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the X chromosome (a) Short intron, b) Long intron, c) Intergenic). The x-axis is the number of derived alleles at a given SNP grouped into bins (bin size = 7). The y-axis is the frequency of the total number of derived alleles within each bin.
Appendix C

Supplementary Figure 3 – Autosomes

A. Short intron

B. Long intron
Supplementary Figure 3. Site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the autosomes (a) Short intron, b) Long intron, c) Intergenic) sorted by recombination rate. The x-axis is the number of derived alleles at a given SNP grouped into bins (bin size = 7). The y-axis is the frequency of the total number of derived alleles within each bin.
Appendix D

Supplementary Figure 4 – X
A. Short intron

Supplementary Figure 4 – X
B. Long intron
Supplementary Figure 4. Site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the X (a) Short intron, b) Long intron, c) Intergenic) sorted by recombination rate. The x-axis is the number of derived alleles at a given SNP grouped into bins (bin size = 7). The y-axis is the frequency of the total number of derived alleles within each bin.
## Supplementary Table 1

| Derived allele | AT->GC | GC->AT | AT->GC | GC->AT | AT->GC | GC->AT | AT->GC | GC->AT | AT->GC | GC->AT | AT->GC | GC->AT | AT->GC | GC->AT |
|---------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Short intron   |        |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
| Low           | 1252   | 933    | 195    | 1515   | 1207   | 1359   | 2948   | 55046  | 36133  | 98205  | 27617  | 54416  | 17698  | 34874  | 16540  | 34138  | 96860  | 34729  |
| Medium        | 219    | 2436   | 216    | 274    | 269    | 3347   | 6486   | 3059   | 6699   | 3471   | 6995   | 2039   | 4031   | 1903   | 411    | 2064   | 497    |
| Low           | 108    | 149    | 322    | 101    | 161    | 1304   | 2674   | 1431   | 2994   | 1471   | 3096   | 729    | 1525   | 813    | 1877   | 836    | 1956   |
| Medium        | 68     | 88     | 71     | 113    | 63     | 728    | 1187   | 836    | 743    | 765    | 1719   | 419    | 856    | 453    | 1036   | 570    | 1338   |
| Low           | 12      | 64     | 52     | 73     | 53     | 368     | 913    | 532    | 1167   | 523    | 1189   | 279    | 489    | 323    | 679    | 336    | 727    |
| Medium        | 16      | 20     | 33     | 46     | 29     | 294     | 553    | 340    | 766    | 395    | 797    | 177    | 315    | 209    | 454    | 234    | 509    |
| Low           | 26     | 32     | 49     | 33     | 43     | 201     | 404    | 319    | 607    | 292    | 564    | 114    | 272    | 160    | 353    | 160    | 358    |
| Medium        | 50     | 21     | 22     | 34     | 39     | 205     | 349    | 219    | 476    | 231    | 452    | 110    | 173    | 119    | 281    | 133    | 256    |
| Low           | 17     | 24     | 22     | 37     | 36     | 141     | 365    | 216    | 587    | 181    | 349    | 71     | 136    | 102    | 207    | 111    | 213    |
| Medium        | 64     | 21     | 24     | 16     | 14     | 106     | 251    | 197    | 333    | 153    | 316    | 64     | 126    | 84     | 196    | 68     | 196    |
| Low           | 71     | 12     | 21     | 33     | 17     | 114     | 388    | 140    | 290    | 130    | 263    | 70     | 103    | 64     | 194    | 71     | 179    |
| Medium        | 78     | 19     | 14     | 21     | 23     | 89     | 368    | 123    | 256    | 129    | 251    | 51     | 103    | 60     | 194    | 96     | 196    |
| Low           | 65     | 18     | 20     | 19     | 18     | 97      | 553    | 113    | 240    | 125    | 251    | 51     | 125    | 77     | 129    | 93     | 147    |
| Medium        | 92     | 11     | 23     | 29     | 12     | 77      | 465    | 140    | 245    | 107    | 269    | 67     | 80     | 75     | 136    | 79     | 142    |
| Low           | 94     | 14     | 21     | 16     | 16     | 87      | 441    | 103    | 215    | 108    | 219    | 48     | 85     | 60     | 124    | 68     | 143    |
| Medium        | 124    | 13     | 23     | 21     | 23     | 70      | 351    | 130    | 198    | 119    | 261    | 50     | 93     | 80     | 127    | 67     | 141    |
| Low           | 115    | 16     | 19     | 19     | 33     | 120     | 320    | 146    | 325    | 167    | 318    | 91     | 136    | 94     | 192    | 89     | 181    |
| Medium        | 134    | 26     | 24     | 29     | 27     | 178     | 365    | 187    | 366    | 175    | 353    | 90     | 171    | 103    | 202    | 109    | 203    |
| Low           | 141    | 24     | 25     | 48     | 39     | 213     | 312    | 222    | 458    | 261    | 443    | 140    | 201    | 144    | 253    | 136    | 258    |
| Medium        | 146    | 41     | 51     | 66     | 55     | 318     | 466    | 367    | 668    | 395    | 666    | 239    | 314    | 240    | 385    | 244    | 394    |
| Low           | 155    | 12     | 99     | 104    | 115    | 125     | 202    | 116    | 974    | 1629    | 944    | 1430    | 315    | 742    | 576    | 862    | 578    | 855    |

Appendix E
Supplementary Table 1. Counts for each of the site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the autosomes (Short intron, Long intron, Intergenic), sorted by recombination rate (bin size equals 7).
Supplementary Table 2

<table>
<thead>
<tr>
<th>Derived allele</th>
<th>Medium</th>
<th>Low</th>
<th>Intergenic</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Intergenic</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Intergenic</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT→GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC→AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT→GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC→AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT→GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC→AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT→GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC→AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT→GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix F
Supplementary Table 2. Counts for each of the site frequency spectra of GC- and AT- increasing alleles for the three sequence types on the X (Short intron, Long intron, Intergenic), sorted by recombination rate (bin size equals 7).