Analysis of GC-Biased Gene Conversion Using Haplotype Networks

Agustin RR[1], Apostol EC[1], Espigar AM[2], and Lapid C[3]

[1] School of Statistics, University of the Philippines Diliman, Quezon City 1101
[2] Archaeological Studies Program, University of the Philippines Diliman, Quezon City 1101
[3] Philippine Genome Center, University of the Philippines Diliman, Quezon City 1101

ABSTRACT

GC-biased gene conversion (gBGC) is a recombination-associated process where G/C alleles are favored over A/T alleles. The non-symmetric segregation of G/C and A/T is a result of the mismatch repair bias that occurs during meiosis, producing more G/C than A/T-containing gametes. This mimics positive selection by promoting the fixation of G/C over A/T, thereby having implications on studies on functional genomics and evolution. However, despite the existence of evidence indicating the presence of gBGC in mammals, and its effect on evolution, there is still no direct classical statistical test of gBGC on humans. This paper aims to test the hypothesis of GC-biased gene conversion in the highly recombining regions of human DNA. Candidate hotspot regions were determined and CpG islands, gene, and gene-associated regions were eliminated. Haplotype networks were constructed for each of these regions; and recombination and gene conversion events were inferred from these networks. Statistical tests were performed on the inferred gene conversion mutations to test for the gBGC hypothesis. Results show that there is a significant ($p<0.05$) bias towards AT->GC mutations.

Introduction

GC-biased gene conversion (gBGC) is a process where GC alleles are favored over AT alleles. This event occurs during meiotic recombination. When heteroduplex DNA is produced, a mismatch occurs. If the repair system that corrects the mismatch is biased towards GC alleles over AT alleles, the gametes produced will not follow Mendelian proportions (Marais, 2003). Instead, there will be more GC-bearing gametes than AT-bearing gametes, which will accelerate the fixation of GC alleles (Galtier et. al., 2001). The patterns of fixation of guanine or cytosine alleles generated by this process is identical to what is expected from selection; except that gBGC disregards their effects on a population's fitness (Kotska et. al., 2012).

gBGC is still a neutral process even if alleles do not segregate according to the neutral model because despite the bias towards GC alleles, there are no differences in fitness between individuals. This means that gBGC may confound observation of selection and interfere with the selective forces of evolution in populations. (Duret and Galtier, 2009). Hence, there is a need to thoroughly investigate the mechanisms of gBGC.

gBGC also provides the mechanism for human genomic features that were insufficiently explained selectionist and neutralist models, such as GC heterogeneity (eg

isochores, higher segregation frequency of AT → GC alleles at neutral sites) and correlation between high recombination rate and GC-richness (Eyre-Walker, 1999; Fullerton et. al., 2001; Galtier et. al., 2001). Data indicating presence of gBGC has been observed on a wide array of organisms, both eukaryotes, including humans, (Pessia, 2012) and bacteria (Lasalle, 2015).

However, despite the wide body of evidence demonstrating the presence of gBGC in humans, and its effect on evolution, there is still no direct classical statistical test of gBGC on the human genome. This paper aims to verify the hypothesis of GC-biased gene conversion in the highly recombining regions of human DNA. The specific aims are:

1. Acquire and analyze a genetic map for the human genome and identify regions with high recombination rates (recombination hotspots)
2. Filter data to remove genes, CpG islands, and other confounding factors.
3. Download sequence data from the 1000 Genomes Project corresponding to these regions, with (low-recombining) flanking regions as a control.
4. Construct haplotype networks based on the sequence data.
5. Infer gene conversion events, and collect data on associated mutations.
6. Perform statistical testing of stated hypothesis

Review of Literature

gBGC and Genomic GC Content

Holmquist (1992) first proposed that gBGC may be the evolutionary force that propels the heterogeneity of GC distribution across the genome. It was an alternative to two major hypotheses on isochore evolution at that time. The selectionist view proposed that isochores are adaptive structures (Bernardi and Bernardi, 1986) while the neutralist view posited that the differences in GC composition are caused by different mutational bias of DNA polymerases (Filipski, 1987). Later papers reported observations that contradicted these two hypotheses (Goncalves et al. 2000; Eyre-Walker, 1999). With this , Galtier et. al. (2001) argued that gBGC is the most likely major determinant of GC content in mammals. The gBGC hypothesis provides an explanation for observations that were otherwise not demonstrated by the other two models, such as: “(i) the higher GC content of regions undergoing gene conversion; (ii) the correlation between recombination rate and GC content, where recombination seems to be the governing force; and (iii) the nonneutral human polymorphism patterns” (Galtier et. al., 2001).

Highly-recombining regions have been shown to exhibit gBGC, hotspot drive, indel drive, and mutagenesis (Webster and Hurst, 2012). Recombination rates have been shown to be positively correlated with the GC content in the human genome (Fullerton et. al., 2001), and the role of gBGC in driving this content has been demonstrated (Galtier, 2003).

Implications in Human Functional and Evolutionary Genomics

The similarity of the dynamics of gBGC and selection may confound observation of selection on populations. gBGC must first be eliminated for data that indicates selection before it could be confidently interpreted as such. This means that gBGC extends the null

hypothesis of evolution. In addition, due to the accelerated fixation of G and C alleles, gBGC may impede adaptive selection. Fixation of deleterious alleles, driven by gBGC, has been observed in primates (Galtier et. al., 2009).

Some phenomena specific to human evolution may be explained by gBGC. Human accelerated regions (HARs) are DNA sequences that remained largely unchanged throughout mammalian evolution but underwent accelerated changes after divergence of humans with the chimpanzees (Hubisz and Pollard, 2014). Although the genesis of HARs indicate positive selection, patterns of substitutions of many elements indicate that gBGC is an important evolutionary force in their formation and change (Hubisz and Pollard, 2014; Kaltzmann et. al., 2010).

Methodology

Generation of Candidate Hotspot Regions

Candidate hotspot regions were determined from a build GRCh38 genetic map by selecting regions with a minimum recombination rate of 15 cM/MB and length of at least 3000 bp. Left and right flanking regions were 'selected' that regions with a maximum recombination rate of 3 In cM/MB with maximum distance of 3000bp from the candidate hotspot region.

Generation of Haplotypes

The candidate hotspots, with their flanking regions, were filtered by removing regions that overlapped with areas with CpG islands and gene regions (Appendix 1) using bedtools subtract. Then, variants within the resulting hotspot and their flanking regions were pulled from the 1000 Genomes Project (Appendix 1).

Testing for Recombination and Gene Conversion Events

Using R Studio v.1.2.1335, haplotype networks per region were estimated using maximum parsimony. Subnetworks were approximated based on computed limits of parsimony Templeton et al (1992). Testing of recombination and gene conversion events were based on the physical clustering of inferred homoplasies, following a novel implementation of the CT algorithm (Crandall and Templeton, 1999; Templeton et. al, 2000). A clustering of a set of homoplasies from one inferred parental haplotype (one run) and another clustering from second parental haplotype (one run), resulting in two runs, is indicative of recombination (Templeton et. al, 2000). However, the same can result from gene conversion. Since, the algorithm followed by this study could not differentiate between recombination and gene conversion given two runs, gene conversion events were tested only for observations with three runs, with the assumption that gene conversion took place from one haplotype into the middle of a second haplotype. Hypergeometric test was applied to both putative recombination events (runs = 2) and putative gene conversion events (runs = 3).

Testing for GC-bias

The mutations that resulted from the inferred recombination and gene conversion events were tallied. A Chi-square test for goodness-of-fit was performed to test for the presence of GC bias, where the null hypothesis was 1:1 ratio of AT->GC and GC->AT.

Results

Data Gathered

A total of 340 hotspots, 319 left flanking regions, and 315 right flanking regions were successfully sampled. We also count the conversion event of each of the regions (hotspots = 1168, left flanks = 1475, and right flanks = 1149). The number of mutations was also tallied. The results show that the number of AT -> GC mutations in all regions is significantly larger than the number of GC -> AT mutations, with a total of 19,546 AT -> GC mutations and 17,614 GC -> AT mutations. The table shows the overall tally of the mutations of each region.

MUTATIONS	REGIONS			TOTAL
	HOTSPOTS	FLANK 1	FLANK 2	
AT -> GC	6398	7563	5585	19546
GC -> AT	5689	6979	4946	17614
TOTAL	12087	14542	10531	37160

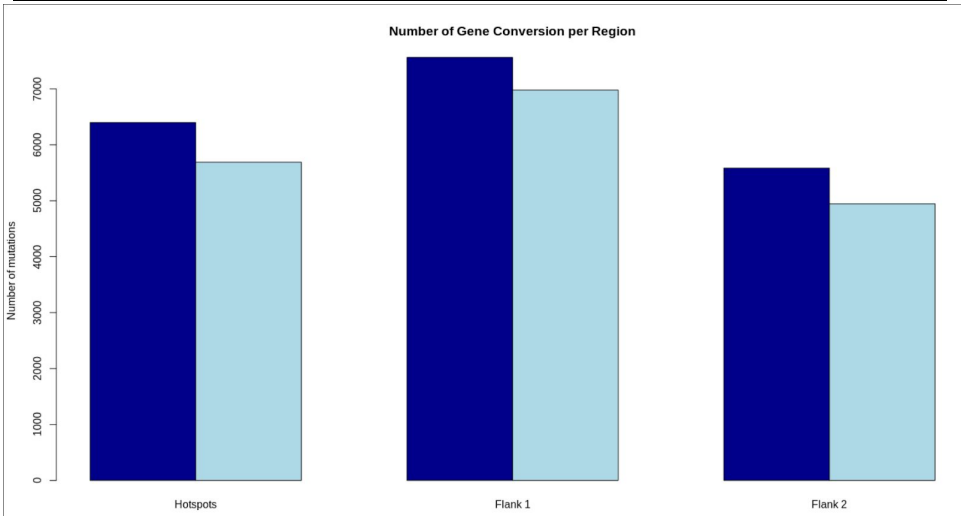

Figure 1. Number of gene conversion per region. The plot shows the number of mutations happened in hotspots, left and right flanks regions. The dark blue represents the number of AT to GC gene conversions while the light blue represents the number of GC to AT gene conversions.

Statistical test

Given the data, we performed a Chi-square goodness-of-fit test to test our hypothesis. The result shows that there is sufficient evidence to say, at 0.05 level of significance, that there is bias in gene conversion in favor of AT -> GC than GC -> AT in across all regions (χ^2 = 100.45, degrees of freedom = 1, p-value = $2.2e^{-16}$). We also

performed a Chi-square goodness-of-fit test for each of the region type - hotspots, left flank and right flank. All tests reject the null hypothesis.

Discussion

Previous studies (Fullerton et. al., 2001; Arbeithuber et. al., 2015) suggest that gene conversion events are more likely to occur in highly recombining regions. However, our results show that the location of gene conversion does not discriminate relative to value of recombination rate. GC-biased gene conversion was also observed in areas with relatively low recombination rates. It must be noted, however, that several factors may have brought about this observation. For example, correlation between sample length and number of loci must be checked. Given that the number of subnetworks calculated is dependent on the number of variants observed, it may be that the number of inferred events.

Despite the limitations in this study, it provides a preliminary data that shows a strong and significant bias towards GC alleles over AT alleles.

Recommendations

Although the results reflect a significant difference between the direct tally of GC-> AT and AT->GC mutations, the methods that lead to this observation still requires further refinement. But prior to rectifying the methods, preliminary analyses of variables must first be performed. Statistical tests on variables such as input lengths, number of variant loci, number of subnetworks, and other factors must be performed to infer possible correlations, which will inform on how to improve the methods.

Inference procedures must also be revisited. The approach used in this study required two candidate parental haplotypes within the network. This means that the estimates derived are conservative, hence, a number of possible recombination events may have been unaccounted.

Another factor that may contribute to the conservative counts is the inability of the algorithm used to differentiate recombination vs gene conversion for events with two runs. Methods that will enable distinction of the two may be explored.

Conclusion

There is a significant bias towards AT->GC mutations over GC->AT mutations. This is a preliminary finding and requires further re-analysis for confirmation.

References

Arbeithuber, B. Betancourt, A.J., Ebner,T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences, 112* (7): 2109-2114; doi: 10.1073/pnas.1416622112

Bernardi, G. and Bernardi, G.. (1986). Compositional constraints and genome evolution. *Journal of Molecular Evolution 24*(1-2):1–11

Crandall K.A. and Templeton A.R.(1999) Statistical approaches to detecting recombination. In: Crandall KA, editor. The Evolution of HIV. Baltimore: *John Hopkins University Press*: 153-76.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics 10*:285-311 https://doi.org/10.1146/annurev-genom-082908-150001

Eyre-Walker, A. (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics 152*: 675–683

Filipski, J. (1987). Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett. 217*(2):184–186. https://doi.org/10.1016/0014-5793(87)80660-9

Fullerton, S.M., Carvalho, A.B., and Clark, A.G (2001). Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Molecular Biology and Evolution 18*(6): 1139-1142. https://doi.org/10.1093/oxfordjournals.molbev.a003886

Galtier, N. (2003). Gene conversion drives GC content evolution in mammalian histones. *Trends in Genetics, 19*(2):65-68. https://doi.org/10.1016/S0168-9525(02)00002-1

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics 159*(2):907-911

Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates *Trends in Genetics, 25*(1):1-5. https://doi.org/10.1016/j.tig.2008.10.011

Goncalves I., Duret L., Mouchiroud D. (2000). Nature and structure of human genes that generate retropseudogenes. *Genome Research. 10*: 672–678

Holmquist, G. P.. (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet. 51*:17–37

Hubisz, M.J., and Pollard, K.S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development (29)*: 15-21 https://doi.org/10.1016/j.gde.2014.07.005

Katzman, S., Kern, A.D., Pollard, K.S., Salama, S.R., and Haussler, H. (2010). GC-Biased Evolution Near Human Accelerated Regions. *PLOS genetics*. https://doi.org/10.1371/journal.pgen.1000960

Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S. (2012). The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Molecular Biology and Evolution, 29*(3): 1047–1057. doi:10.1093/molbev/msr279

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. (2015). GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLOS genetics*. https://doi.org/10.1371/journal.pgen.1004941

Marais, G.A.B. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics. (19)*6:330-338, https://doi.org/10.1016/S0168-9525(03)00116-1

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G.A.B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biology and Evolution, 4*(7): 675–682, https://doi.org/10.1093/gbe/evs052

Templeton A.R., Crandall, K.A., and Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. cladogram estimation. *Genetics. 132*(2):619-33.

Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. (2000). Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics 156*(3):1259-75.

Webster M.T. and Hurst, L.D. (2012). Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in Genetics, 28*(3): 101-109; doi:10.1016/j.tig.2011.11.002

Appendix 1: Detailed Methodology

Generation of filtered hotspots

The build GRCh38 genetic map was obtained. Recombination rates were calculated based on the ratio of genetic and physical distances provided by the map. The CpG islands and gene regions were downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables), both from the Dec. 2013 (GRCh38/hg38). CpG island data was obtained from group: All Tracks, track: CpG Islands. The CpG Islands dataset had 31,144 items. Gene region data was obtained from group: Genes and Gene Predictions, track: GENCODE v29. The gene region data had 226,811 items and included not only protein-coding genes but also the following: long non-coding RNA genes, small non-coding RNA genes, pseudogenes, immunoglobulin/T-cell receptor gene segments, protein-coding transcripts (full and partial length), nonsense mediated decay transcripts, and long non-coding RNA loci transcripts. After the filtering step, 351 regions were identified.

Generation of vcf files

Using Tabix, variants from the hotspot and flanking regions were pulled from the 1000 Genomes Project phase 3 data (haplotypes = 5008), based on dbSNP release 149 and underwent dbSNP liftover of GRCh37 coordinates to GRCh38 coordinates (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/).