The Effect of Mutation Subtypes on the Allele Frequency Spectrum

Kevin Liao[1], Jedidiah Carlson[2], Sebastian Zoellner[1]
[1]Department of Biostatistics, University of Michigan
[2]Department of Bioinformatics, University of Michigan

YOUR PHOTO

Allele Frequency Spectrum

- The allele frequency spectrum (AFS) is a summary of genetic variation used for population genetic inference
- Inferences based on summary statistics using the AFS assume the infinite sites model
- The infinite sites model assumes constant mutation rate across sites
- Mutation rate heterogeneity and biased gene conversion can affect interpretation of the AFS
- Can lead to false population genetic signals such as selection

Mutation Subtype

- Whole genome sequencing data was used from the BRIDGES study (n = 3556)
- Analyses run using genetic data from chromosome 22
- Flanking bases from reference were used to construct 96 different 3-mer mutation subtypes (MST)

Reference: CCA

Sample: CTA

MST: C_T.CCA

- VCFtools was used to fill in ancestral alleles from Ensembl to produce an unfolded AFS for each 3-mer subtype

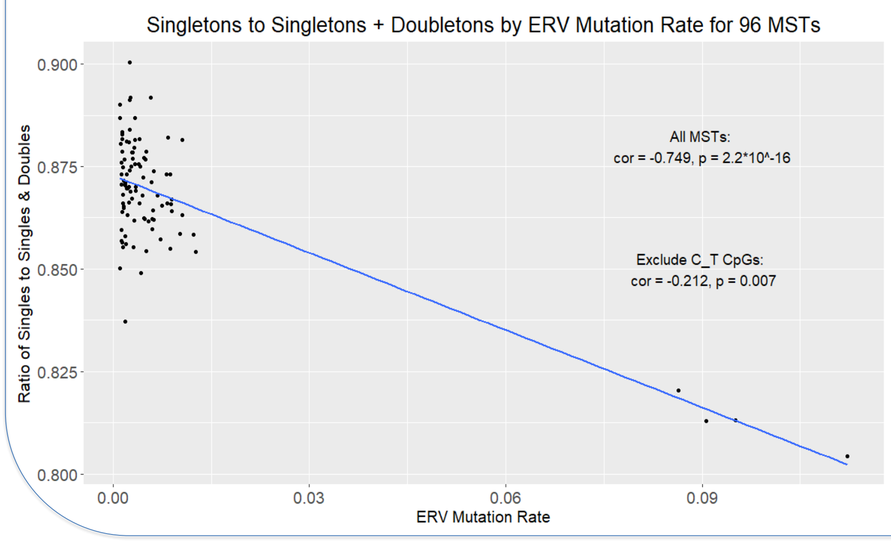
Summary Statistics of AFS

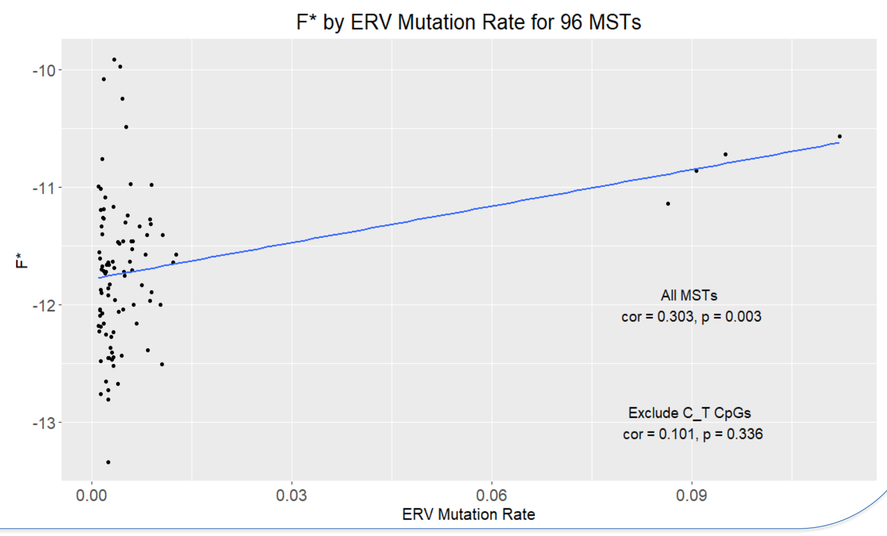
1. Ratio of 1-tons to 1-tons and 2-tons: $p = \frac{\eta_1}{\eta_1+\eta_2}$
2. Tajimas D: $D = \frac{\pi - \frac{S}{a_1}}{\sqrt{var\left(\pi - \frac{S}{a_1}\right)}}$
3. F*: $F^* = \frac{\pi - \frac{n-1}{n}\eta_1}{\sqrt{var\left(\pi - \frac{n-1}{n}\eta_1\right)}}$

η_1: *# of singletons in unfolded AFS*
η_2: *# of doubletons in unfolded AFS*
S: *# of segregating sites*
π: *Mean pairwise differences*

Mutation Rate / Parallel Mutations

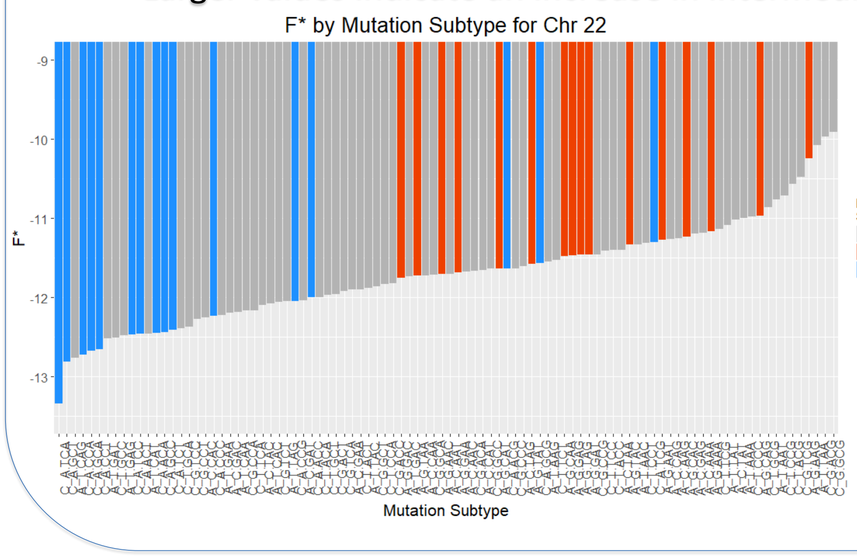
- Parallel singletons are falsely counted as a doubleton
- Lower proportion of singletons for subtypes with higher mutation rates such as CpG mutations (Harpak 2016, PLoS Genet)
- Mutation rates were estimated from extremely rare variants (Carlson 2018, bioRxiv)

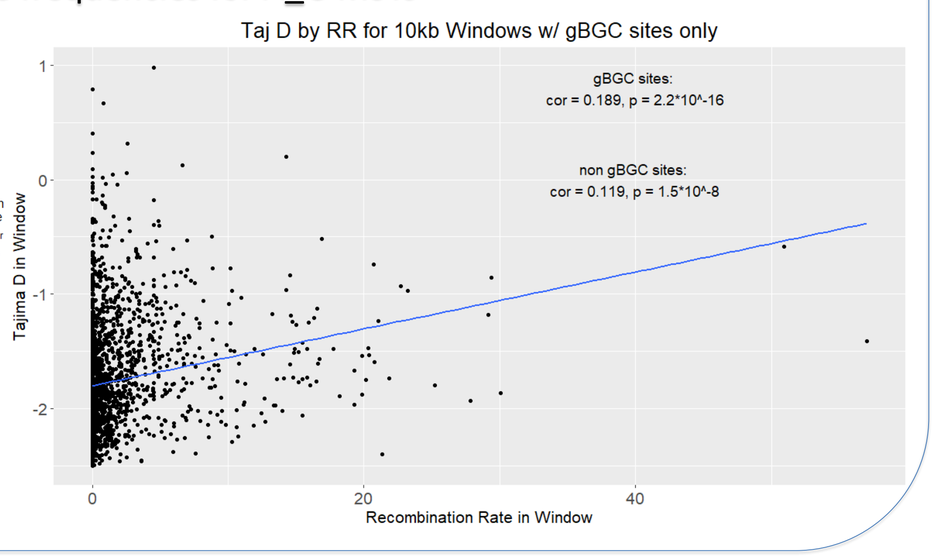




Biased Gene Conversion

- Base mismatch repair during recombination where C/G repairs are more likely for mismatches.
- gBGC causes more A_G MSTs and leads to increases in intermediate frequencies (Glemin 2015, Genome Research)
- Wilcoxon rank sum test shows F* for A_G MSTs have systematically larger values (p = 0.0088)
- Larger values indicate an increase in intermediate frequencies for A_G MSTs





References

Harpak A, Bhaskar A, Pritchard JK (2016) Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. PLoS Genet 12(12): e1006489.

Jedidiah Carlson, Adam E. Locke, Matthew Flickinger, Matthew Zawistowski, Shawn Levy, The BRIDGES Consortium, Richard M. Myers, Michael Boehnke, Hyun Min Kang, Laura J. Scott, Jun Z. Li, Sebastian Zöllner. bioRxiv 108290; doi:

Glémin, S., P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier et al., 2015 Quantification of GC-biased gene conversion in the human genome. Genome Res. 25(8): 1215–1228

Conclusions

- Mutation rate and wrongly classified parallel mutations lower proportion of true singletons
- Biased gene conversion leads to increase in intermediate frequencies for A_G mutation subtypes
- Can mutation rate heterogeneity be corrected for during inference using the AFS?
- Can we quantify the effect of gBGC on the AFS?
- What is causing C_A MSTs to have an excess of rare variants?