The Effect of Mutation Subtypes on the Allele Frequency Spectrum and Population Genetics Inference

Kevin Liao[1], Jedidiah Carlson[2], Sebastian Zöllner[1,3]

[1]Department of Biostatistics, University of Michigan; [2]Department of Computational Medicine and Bioinformatics, University of Michigan;
[3]Department of Psychiatry, University of Michigan

Contact
ksliao@umich.edu
www.kliao12.github.io

Background

Pop Gen Models Treat Sites as Interchangeable

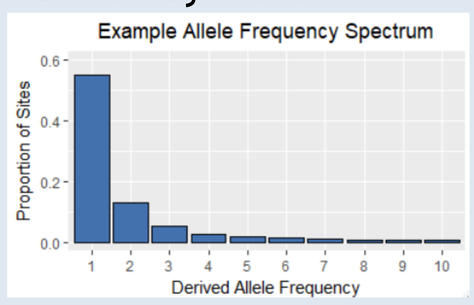
- Current models treat mutations the same:
  A -> G the same as C -> T
- Evolution of sites differ due to unique mutation rates driven primarily by adjacent nucleotides (Carlson, 2018)
- Sites should not be treated the same and can be further differentiated by considering the local nucleotide context

Mutation Subtype:

Base Downstream	Mutation Type	Base Upstream	Mutation Subtype
A	C->T	G	C_T.ACG

The Allele Frequency Spectrum Can Differ by Sites

- AFS is summary of genetic variation in a population that is commonly used for inference
- Current methods use all sites to form a single frequency spectrum
- Shape reflects forces such as natural selection and demographic history



Problem: Signs of selection or demographic history in the overall AFS may be false signals caused simply by its composition of mutation subtypes

Aims

1. What are the factors that affect the allele frequency spectrum at the motif level?
2. Are allele frequency spectrum-based tests of selection and demographic inference biased by failing to account for mutation subtypes?

Results

- Mutation rate heterogeneity and biased gene conversion affect the AFS at the motif level
- Local tests of selection have an inflated rate of false positives due to the local nucleotide composition in a region
- Demographic inference using the distinct AFS for each mutation subtype infers drastically different parameters

Conclusion

AFS-based inference and other population genetics models need to differentiate between sites!

A_G.CAT ≠ C_T.ACG

Dataset

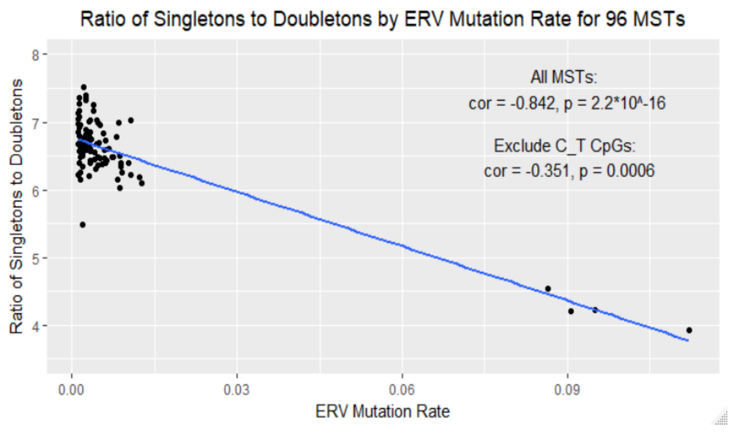
- Bipolar Research in Deep Genome and Epigenome Sequencing Study (BRIDGES)
- 3556 unrelated European individuals
- Average coverage: 9.6x
- Total variants: 59,482,865

Methods

1. Annotate each point mutation with its immediate adjacent nucleotides to form 96 3-mer mutation subtypes (MST)
2. Construct the genome-wide AFS for each MST and compare them by:
   a) Ratio of singletons to doubletons
   b) Tajimas D
3. Compute Tajimas D using the local AFS and the proportion of a single MST in 100kb windows
4. Fit an exponential growth model for each mutation subtypes' AFS using DaDi

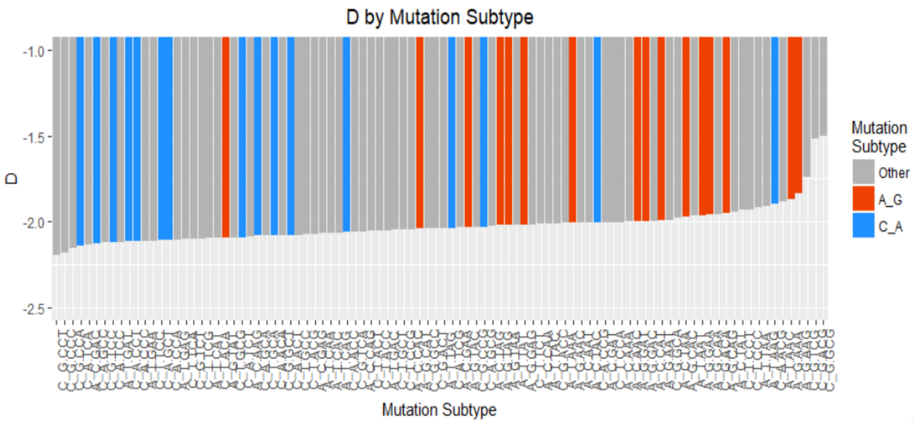
1. Factors affecting the AFS

Mutation Rate Heterogeneity



- Mutation rates differ across subtypes due primarily to adjacent nucleotides
- Sites with higher mutation rates (CpG -> TpG) have lower proportion of singletons
- Parallel singletons falsely counted as doubletons

Biased Gene Conversion (gBGC)

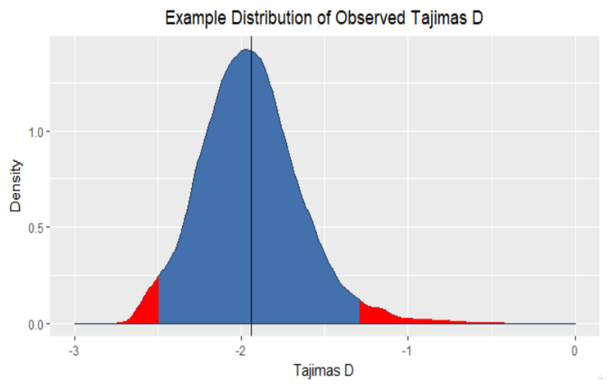


- Base mismatch repair process during recombination where C/G repairs more likely
- gBGC mimics selection on AFS and causes increase in intermediate frequency alleles or rare alleles
- Systematic higher ranks of D for A->G and lower ranks for C->A is consistent with gBGC

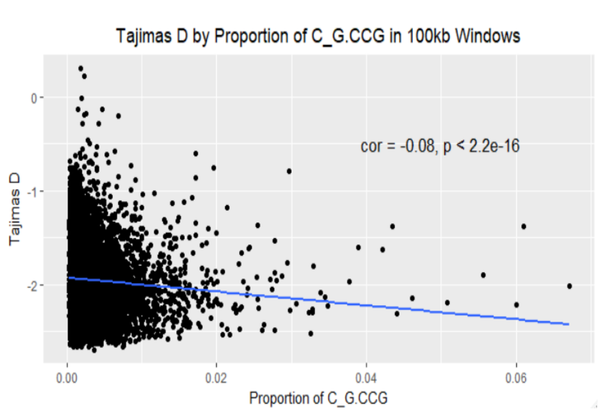
2. Effect of mutation subtypes on AFS-based inference

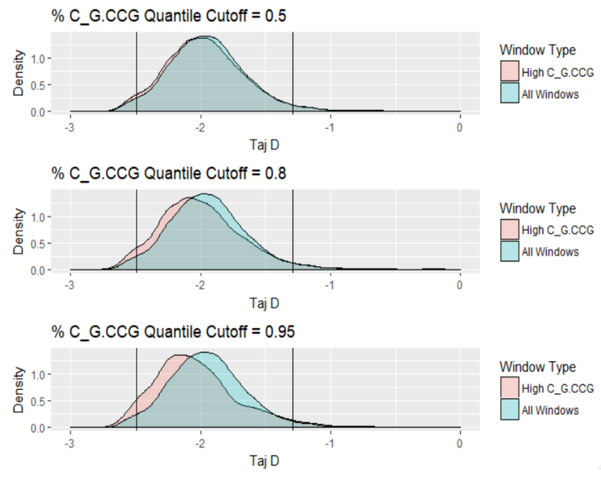
Tests of Selection

- Tajimas D uses AFS to test for local regions of selection
  - Compute D from AFS in local regions of genome
  - Regions with D falling in tails of empirical distribution are significant of selection



Problem: Could windows falsely fall in tails of Tajimas D empirical distribution simply by having more of a particular mutation subtype?

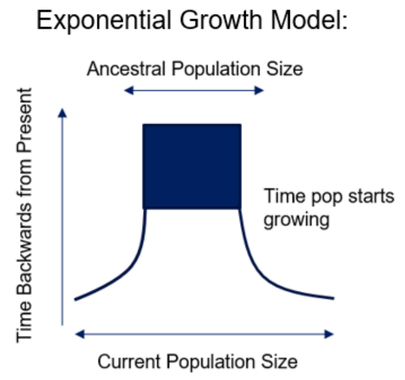




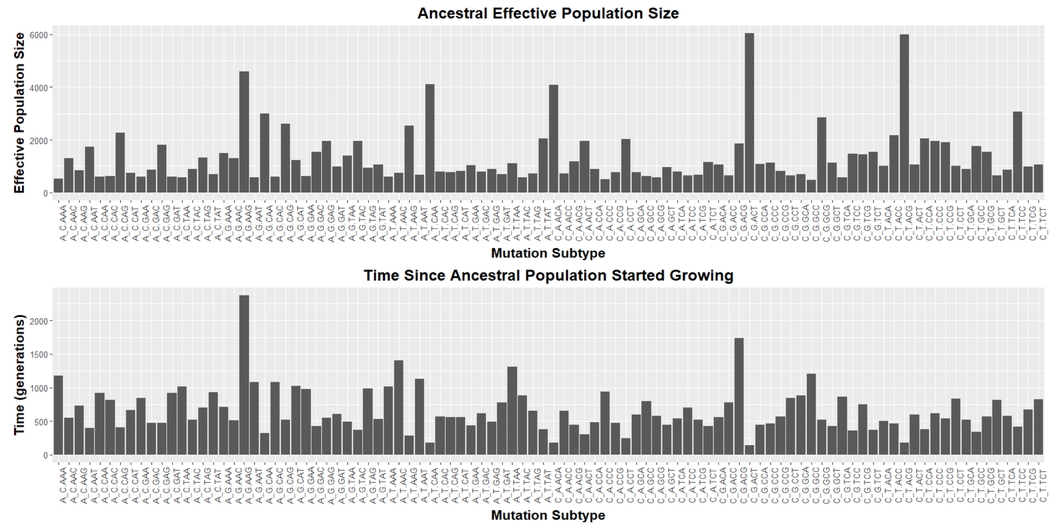
- Artificial negative correlation between proportion of a C_G subtype and D
- Windows with "high" proportion of C_G subtype have heavier left tail
- False signals of selection caused by having "high" proportion of C_G.CCG

Demographic Inference

- $\delta a \delta i$ uses AFS to infer demographic history
- Assuming an exponential growth model:
  - Uses diffusion approximation to estimate expected AFS
  - Infers ancestral effective population size and time since it started growing



Problem: Does demographic inference using the distinct AFS for each mutation subtype give varying results?



Inferred time since ancestral population started growing and effective population size, showing drastic differences by subtype