# Introduction

* Mitochondrial DNA in evolutionary and disease biology
* Variation
  + Haplogroups are evolutionary proxies
  + Microarrays don't include sufficient mtSNVs for accurate imputation
* Importance of high-quality alignments (1 paragraph; extra 100-200 words)
  + Reference alignment v reference sequence
  + Consistency of gaps
  + Curated, globally representative reference panel
* MitoImpute
  + Impute missing variants
  + In silico microarrays
  + ADNI data

# Methods

## Reference Alignment

* PhyloTree
* Aligned to rCRS
  + Why to still include the rCRS? (1-2 sentences; ~50 words)
* Aligned in batches using MAFFT and combined
* Resolved gap placement manually
* More detail about creation of reference alignment (1 paragraph; extra 100-200 words)
  + Difference between reference alignment and reference panel
  + Consistency of gap placement
  + Decision process

## Reference Panel

* Accessing/downloading sequences
* Exclusion of sequences
* Aligning against reference
* Quality control filtering
* Final sequences in panel

## Validation Panel

* Creation of 101 in silico microarrays
* Haplogroup assignment

## Imputation Protocol

* Describe X-chromosome imputation protocol adapted from Goncalves
* No recombination rates of MCMC
* Varying khap and MAF parameters
* Haplogroup assignment
* Assessment of imputation accuracy
  + Haplogroup and genotype concordance
  + Linear Mixed Model ANOVAs

# Results

## Reference Alignment and Reference Panel

* Paragraph (200-300 words)
  + Where the sequences are from
  + Geographic bias of samples
    - Might be hard to obtain, sort through?
  + Representation of lineages
  + Assessment of alignment quality?

## In silico Microarrays

### **Parameter Tuning**

* Fill in existing parts

### **Overall Microarray Performance**

* Fill in existing parts
* Performance drops with more reference haplotypes (khap) and with rarer alleles (MAF)

### **Overall Haplogroup Concordance**

* Fill in existing parts
* Results from specific regions
  + Representative of majority and minority populations

## Alzheimer’s Disease Neuroimaging Initiative Data

* Fill in existing parts

# Discussion

* Describe Usage of SnakeMake
* Pipeline improves haplogroup assignment in datasets with missing data
  + Utilisation by long-term studies with limited mtSNVs
  + Utility in non-European populations
    - GenBank contains sequences not found in projects such as the 1,000 Genomes Project (ie Indigenous Australians, Middle Easterners, Pacific Islanders) (~50 words)
* Discuss how the different parameters affected outcome (1 paragraph; 200-300 words)
  + Is the decreasing accuracy at higher khap counterintuitive?
  + Same with decreasing imputation accuracy with including rarer sequences
  + Does not including sites < MAF 1% mean rare haplotypes cannot be accurately imputed?
    - Bias against sequences underrepresented in reference panel/GenBank
* Utility of the alignment outside imputation (1 paragraph; 100-200 words)
  + Consistent gap placement
  + Well-curated alignment with many sequences
    - Would take a large computation time for others to replicate
* Utility of reference panel outside imputation (1 paragraph; 100-200 words)
  + Reference panel is still a curated alignment
  + Ability to subsample from reference panel
    - Phylogenetics
    - Population genetics
    - Evolutionary analyses