Evidence of Purifying Selection in Humans

John Long
Mentor: Angela Yen (Kellis Lab)

Outline

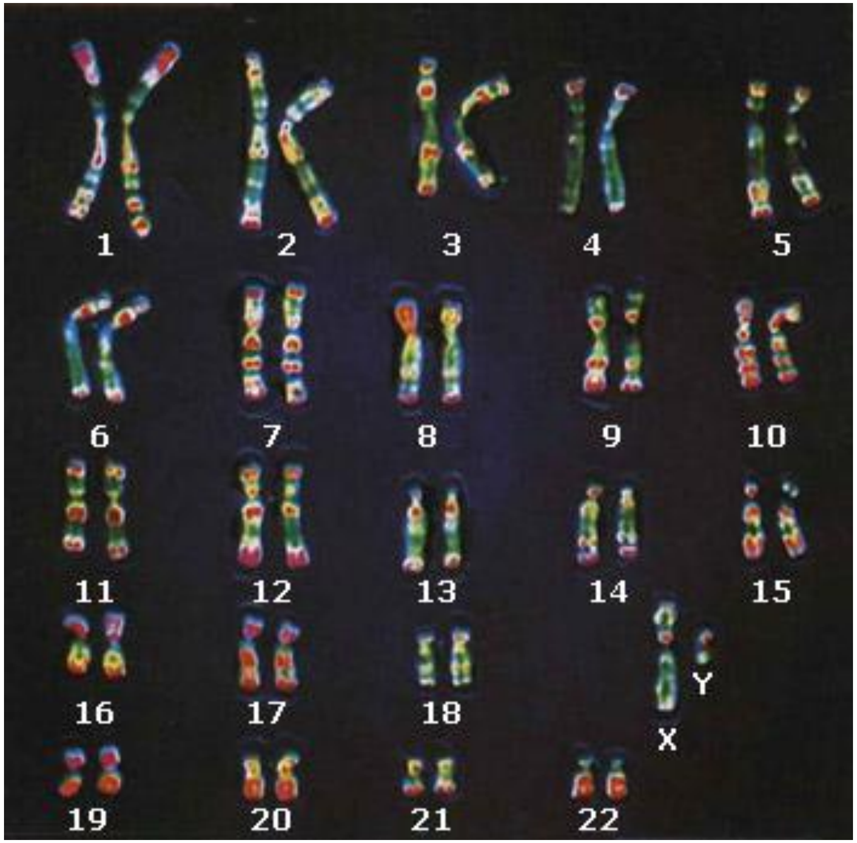
- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- Goal
- Methods
- Progress
- Future Work

Outline

- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- Goal
- Methods
- Progress
- Future Work

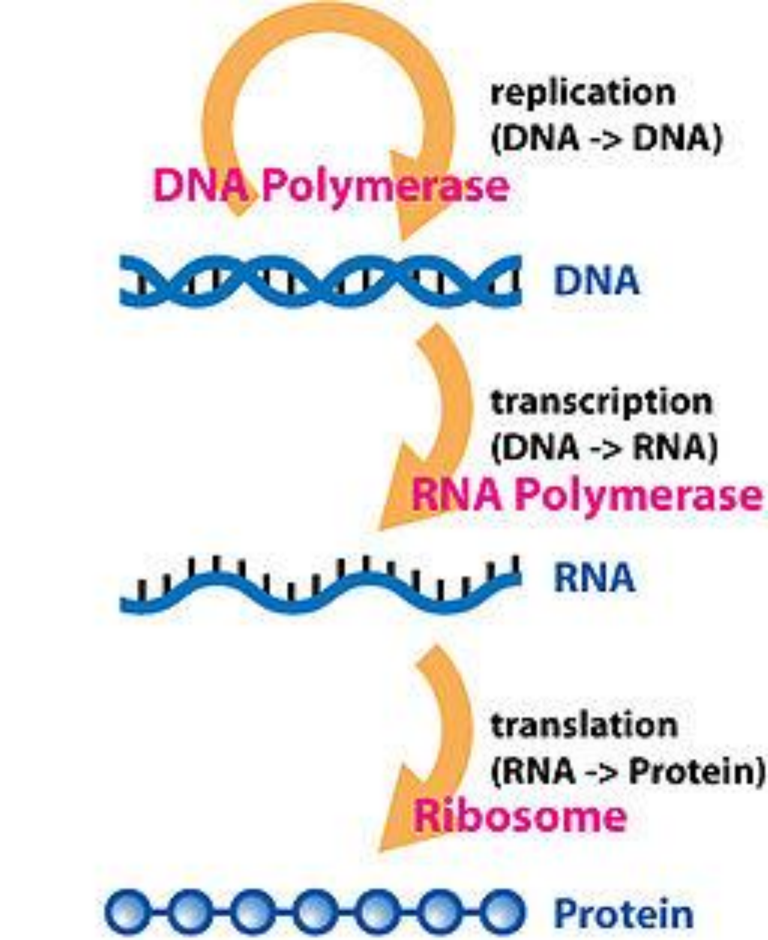
Human Genome

- Genome
  - Set of genetic information
  - Grouped into chromosomes
  - Chromosomes made of nucleotides
- The Human Genome Project (2003)
- Reference Genome
- Function?
  - Genes (2%)
  - Regulatory (10-50%)
  - Junk (50-90%)



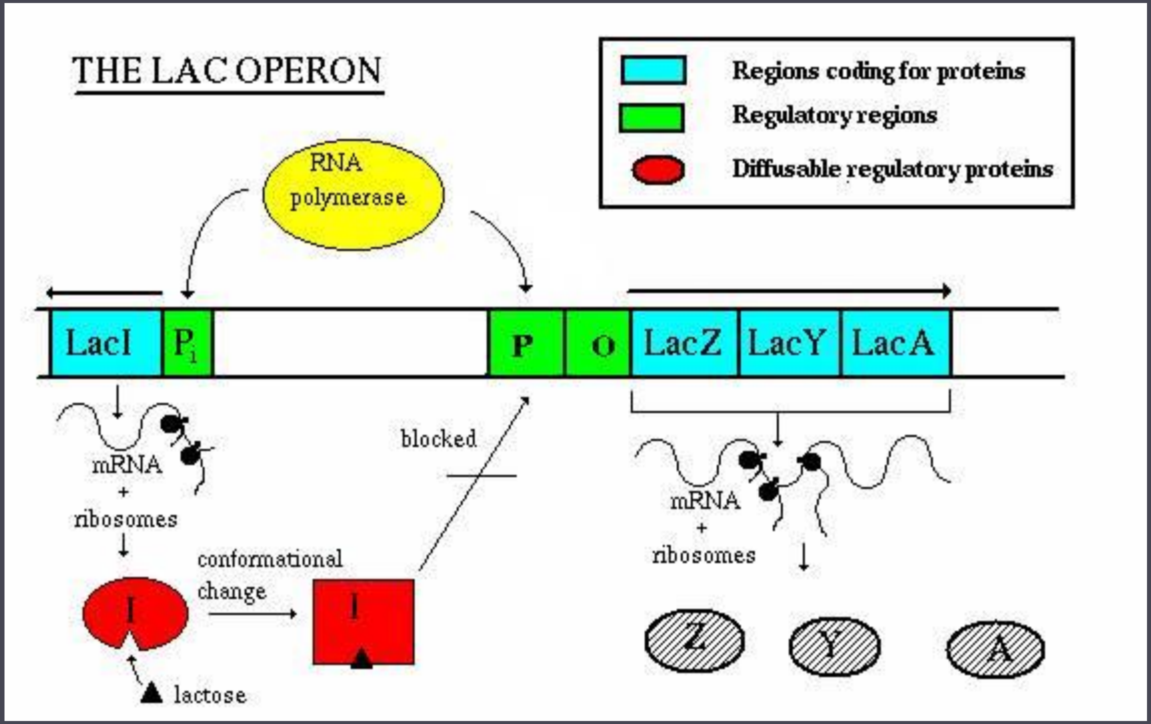
Central Dogma of Biology

Process by which coding DNA regions (genes) get converted to protein



Regulation System

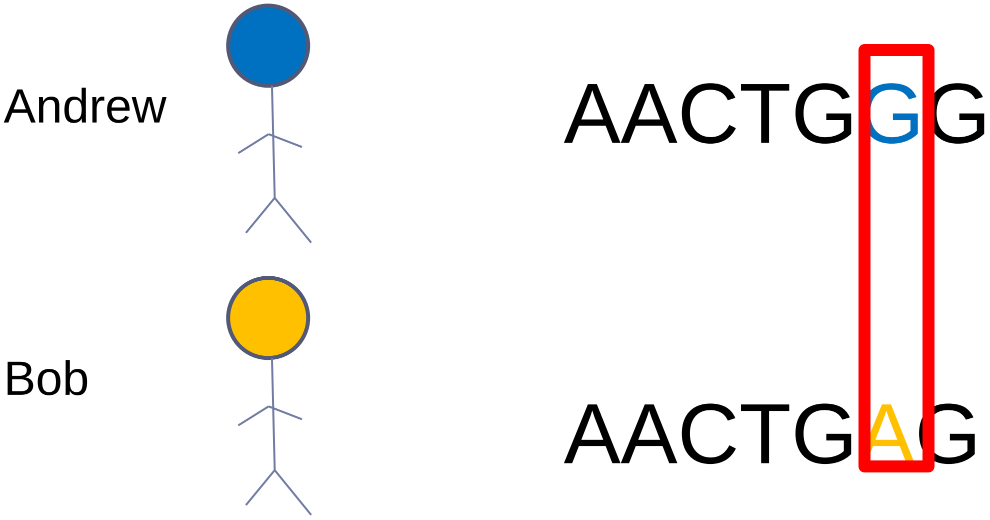
Example of regulation of genes



Natural Selection

- Natural process by which populations evolve
- 2 types
  - Positive selection: increase in frequency of beneficial mutations
  - Negative (purifying) selection: decrease in frequency of deleterious mutations
- Selection occurs in populations (not individuals)
- Over long periods of time
  - 10,000 – millions of years

What is an allele?



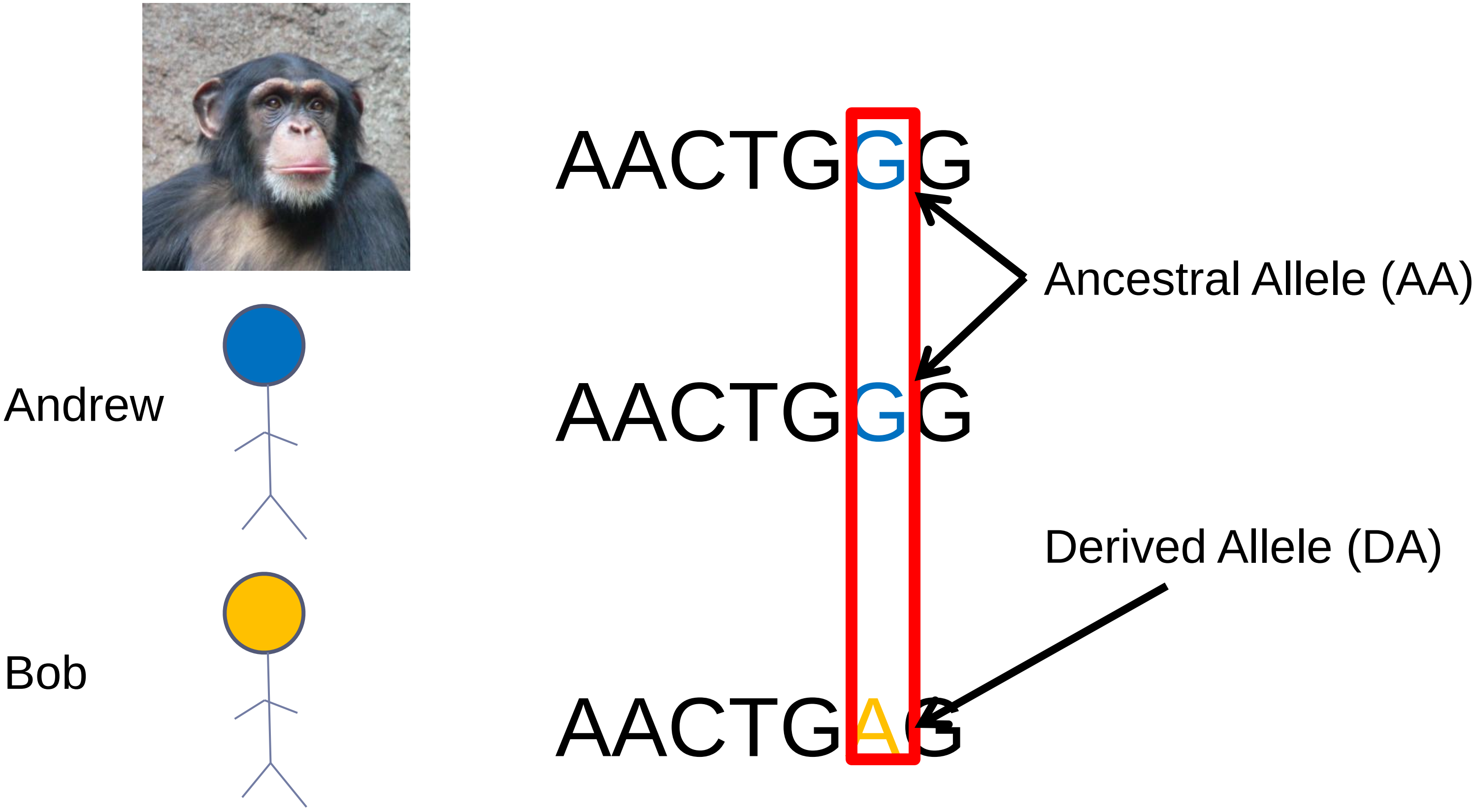
Andrew

AACTGGG

Bob

AACTGAG

Ancestral Allele (AA) and Derived Allele (DA)



Andrew

Bob

AACTGGG

AACTGGG

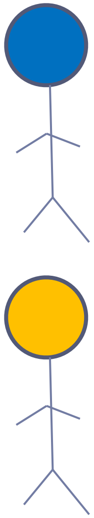
AACTGAG

Ancestral Allele (AA)

Derived Allele (DA)

Single Nucleotide Polymorphism (SNP)

AACTGGG



Andrew

AACTGGG

Bob

AACTGAG

Mutation:
Single Nucleotide Polymorphism

Allele Frequency

- Remember that A is derived allele

AACTGGG

AACTGAG

AACTGAG

AACTGAG

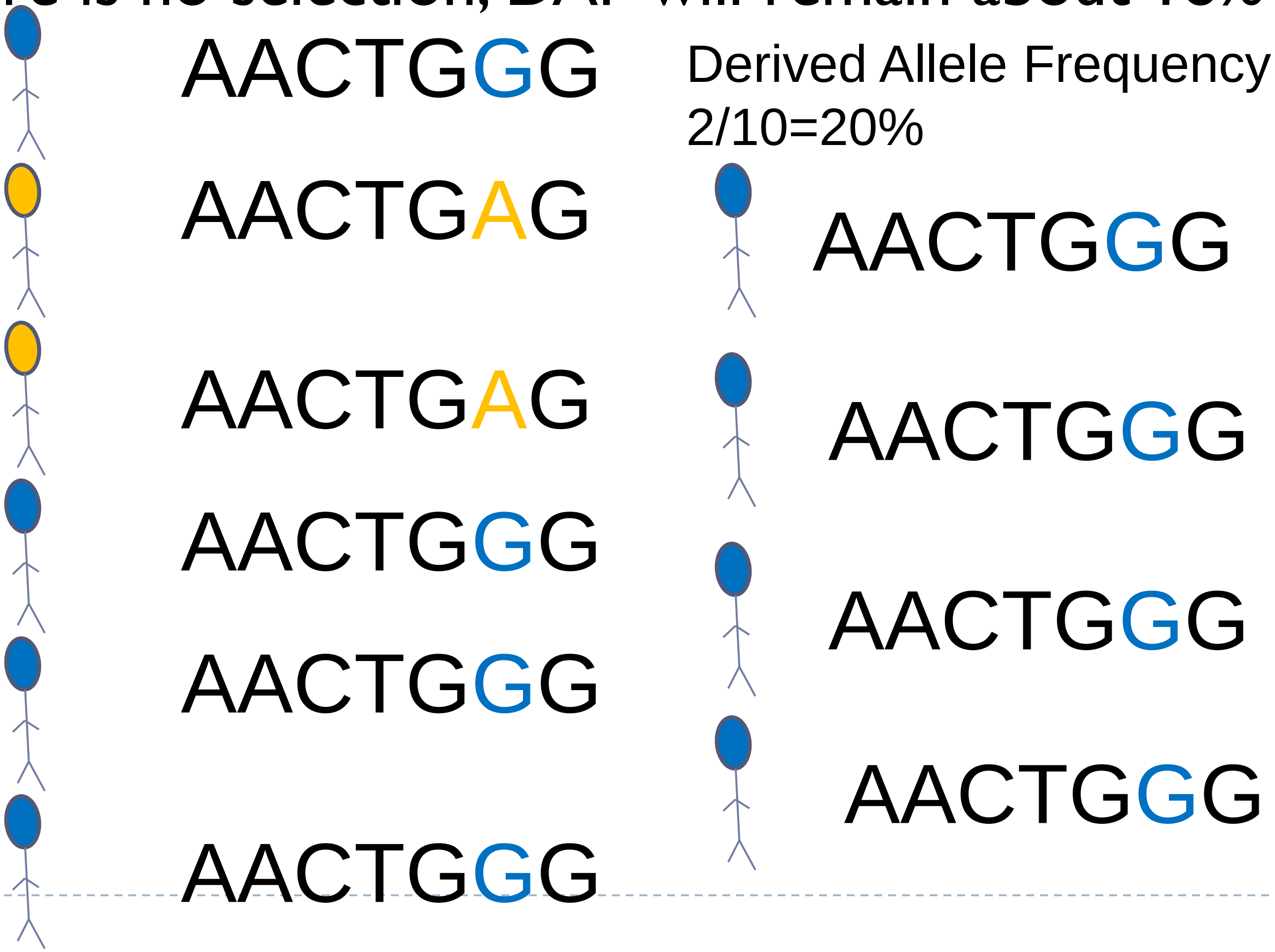
AACTGGG

AACTGGG

Derived Allele Frequency (DAF):
1/2=50%

No Selection

- If the result of a mutation is neutral, there is no selection
- If there is no selection, DAF will remain about 18%



AACTGGG

AACTGAG

AACTGAG

AACTGGG

AACTGGG

AACTGGG

Derived Allele Frequency (DAF): 2/10=20%

AACTGGG

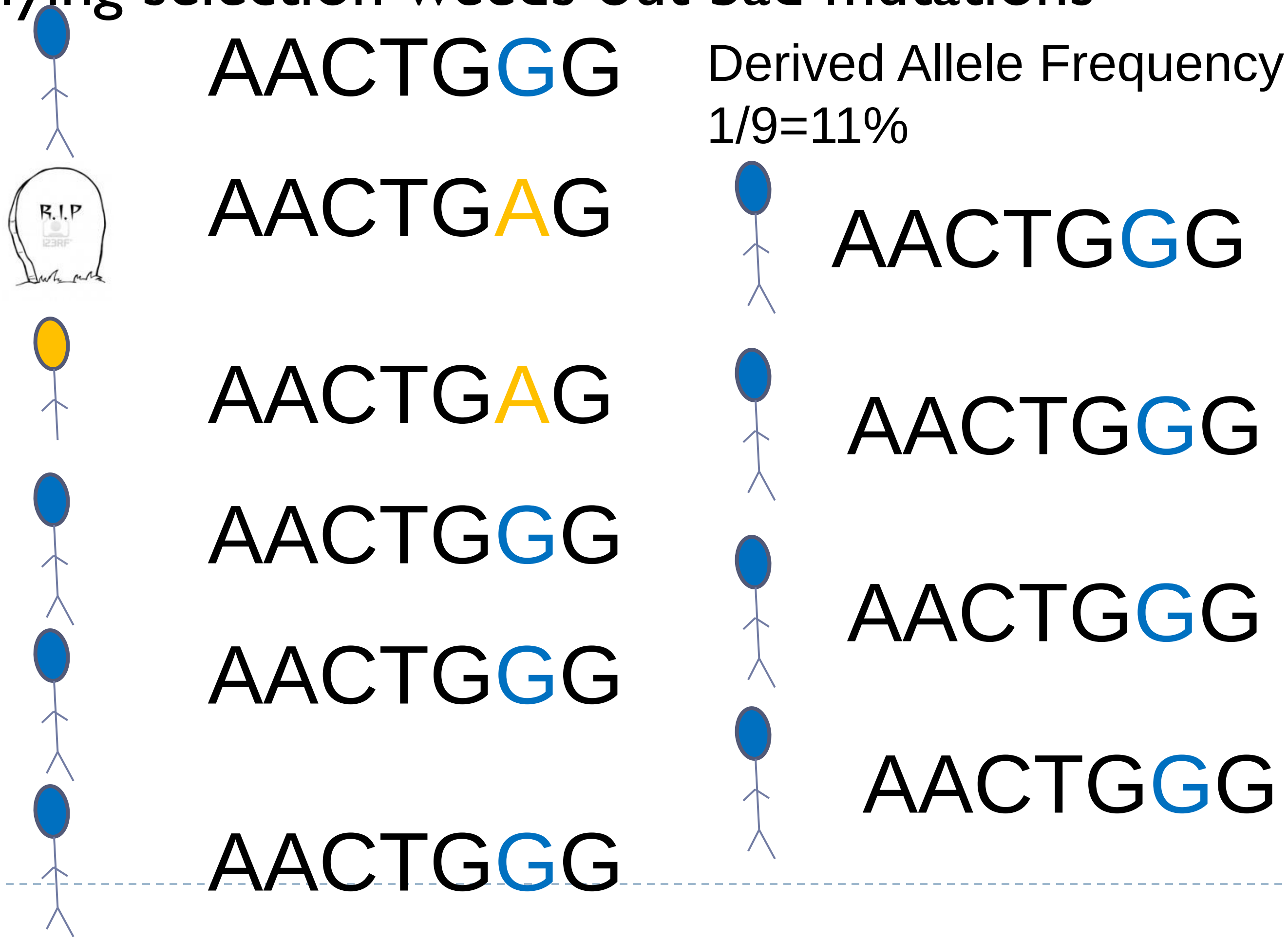
AACTGGG

AACTGGG

AACTGGG

Purifying Selection

- A random mutation is more likely a bad mutation
- Purifying selection weeds out bad mutations



AACTGGG

AACTGAG

AACTGAG

AACTGGG

AACTGGG

AACTGGG

Derived Allele Frequency (DAF): 1/9=11%

AACTGGG

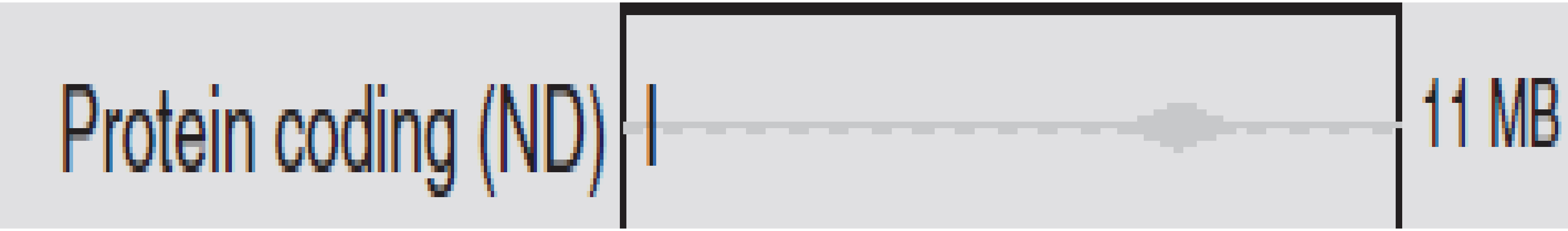
AACTGGG

AACTGGG

AACTGGG

Significance of Selection

- Selection suggests that a set of regions is important
- Purifying selection is more common than positive selection because random mutations are likely bad
- DAF value at a position indicates level of selection
- A lower **mean** DAF across sets of regions indicates purifying selection
- Previous research on genes by Dr. Ward



Bar indicates mean DAF in gene regions

Outline

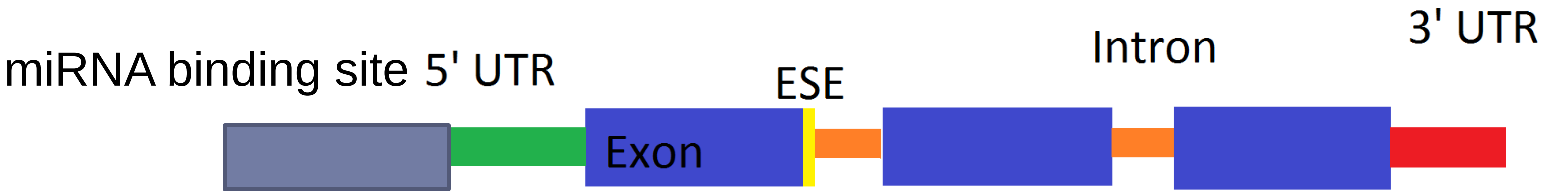
- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- **Goal**
- Methods
- Progress
- Future Work

Goal

- Find evidence of purifying selection in the following regions:
  - 5’ Untranslated Regions
  - Exonic Splicing Enhancers
  - miRNA binding sites
- DAF used to measure selection
- How much of the regions are functional

Sets of Regions

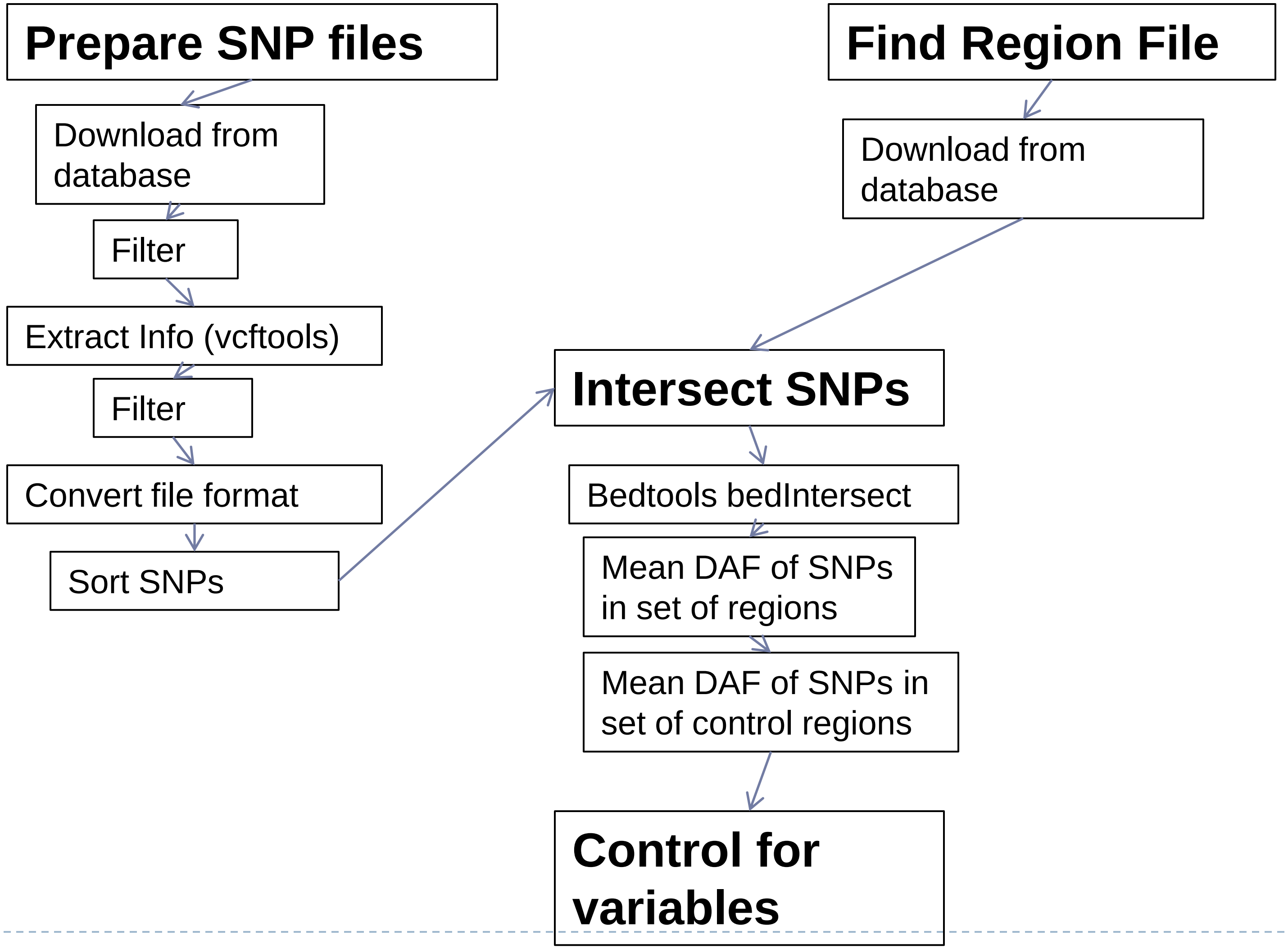
- 5’ Untranslated Region
  - Regions that occur right before a coding region
- Exonic Splicing Enhancers
  - Regions where exonic splicers tend to bind
- Micro RNA binding Sites
  - Regions where Micro RNA tends to bind
  - miRNA: a regulatory molecule



Outline

- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- Goal
- **Methods**
- Progress
- Future Work

Methods



Prepare SNP files

Download from database

Filter

Extract Info (vcftools)

Filter

Convert file format

Sort SNPs

Find Region File

Download from database

Intersect SNPs

Bedtools bedIntersect

Mean DAF of SNPs in set of regions

Mean DAF of SNPs in set of control regions

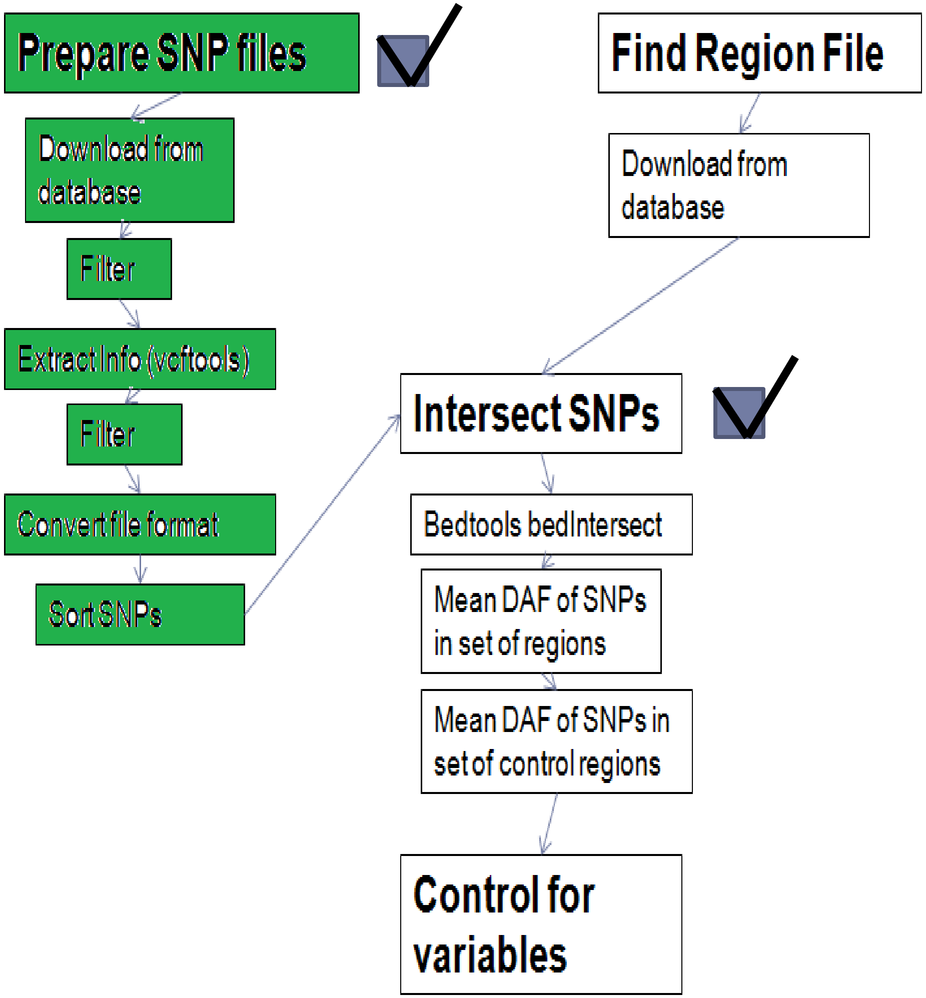
Control for variables

Outline

- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- Goal
- Methods
- Progress
- Future Work

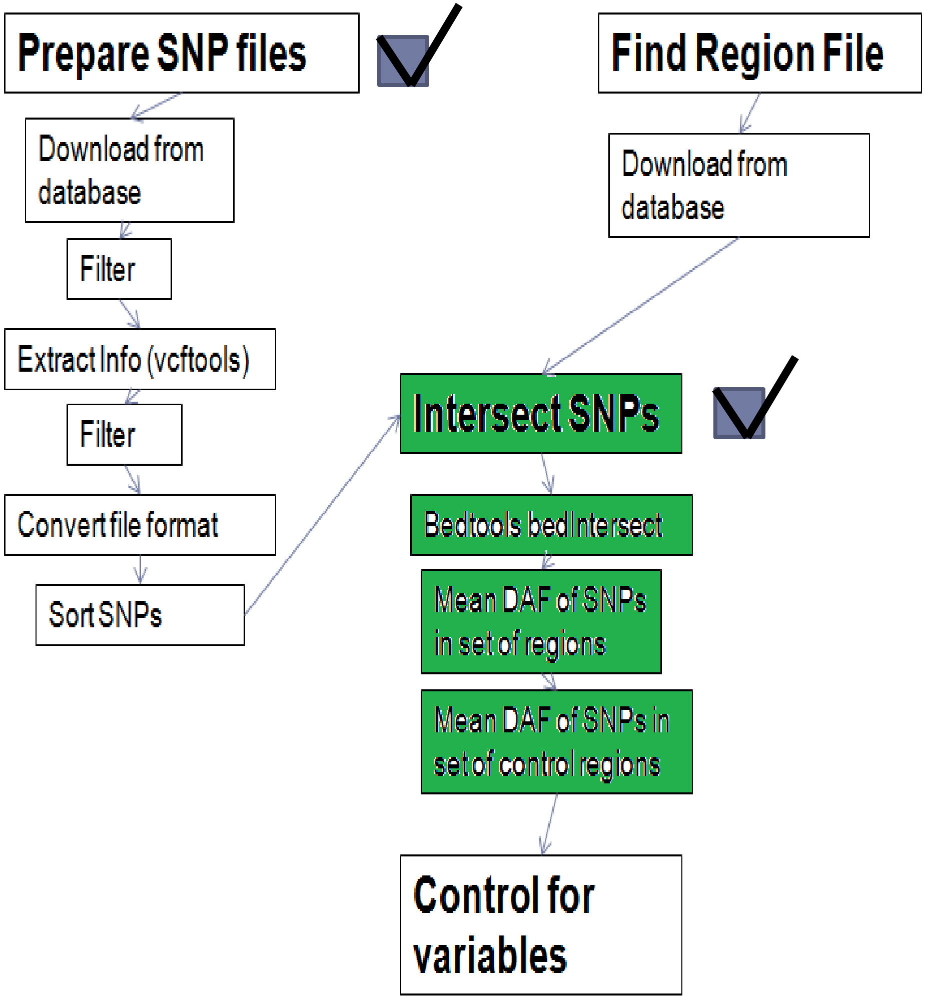
Progress

- Prepared SNP files
  - VCF format (1000 genomes project)
  - Extracted useful information (vcftools)
  - Deleted SNPs
  - Converted to bed file format
  - Sorted to match bed format sorted order
  - Unix and awk commands



Progress

- Intersect SNP and Bed files
  - intersectBed command bedtools
  - Mean DAF of SNPs was calculated using an awk script submitted as job
  - Similarly calculated for control regions



Outline

- Background
  - Genomes
  - Expression
  - Regulation
  - Selection
- Goal
- Methods
- Progress
- Future Work

Future

- Run the program on region files
- Script is adaptable
- Hope to find lower DAF value
- Confirm these regions are important
- Move to other annotated regions of genome

Acknowledgements

- PRIMES program
- Dr. Manolis Kellis, Luke Ward, and Angela Yen
- Angela Yen
- Parents