# Final Exam (Biometrical genetics and plant breeding)

Niranjan Pokhrel

2025-05-06

## Selection scan and genome wide association analysis for NUE related traits in maize.

**For selection scan studies 10 maize lines from tropical and temperate lines each were downloaded in the folder**

**Tropical lines is in directory: /work/agro932/niranjan27/Final_exam/sralite**

**Temperate lines is in directory: /work/agro932/niranjan27/Final_exam/Tropical_vcf**

**Since temperate lines were in fastq format downloaded from NCBI it is extracted and converted to vcf file**

**However, the tropical lines were already as vcf file.**

### Pipeline for extracting the vcf file from fastq file for temperate maize.

```
# Aligning the FASTQ files with reference genome and indexing the resulting BAM files

#!/usr/bin/env bash
#SBATCH -D /work/agro932/niranjan27/Final_exam          # working dir
#SBATCH -o /work/agro932/niranjan27/logs/align-%A_%a.out
#SBATCH -e /work/agro932/niranjan27/logs/align-%A_%a.err
#SBATCH -J align_wg
#SBATCH -p schnablelab
#SBATCH --time=96:00:00
#SBATCH --array=0-9                       # 10 samples indexes 0-9
#SBATCH -c 16
#SBATCH --mem=64G
#SBATCH --mail-user=npokhrel3@huskers.unl.edu
#SBATCH --mail-type=BEGIN,END,FAIL

set -euo pipefail

module load bwa samtools

# paths
REF=/work/agro932/niranjan27/Reference/Zea_mays.Zm-B73-REFERENCE-NAM-5.0.dna.toplevel.fa
```

```
FASTQ_DIR=/work/agro932/niranjan27/Final_exam/fastq
OUT_DIR=/work/agro932/niranjan27/Final_exam/aligned
mkdir -p "$OUT_DIR"

# sample list (ten SRR IDs)
SAMPLES=(
  SRR5725631 SRR5725632 SRR5725633 SRR5725634 SRR5725635
  SRR5725636 SRR5725637 SRR5725641 SRR5725642 SRR5725643
)
SAMPLE="${SAMPLES[$SLURM_ARRAY_TASK_ID]}"

# input FASTQs
R1=${FASTQ_DIR}/${SAMPLE}.sralite.1_1.fastq
R2=${FASTQ_DIR}/${SAMPLE}.sralite.1_2.fastq
[[ -f $R1 && -f $R2 ]] || { echo "missing FASTQ for $SAMPLE"; exit 1; }

# reference index (only task 0 checks/builds)
if [[ $SLURM_ARRAY_TASK_ID -eq 0 && ! -f ${REF}.bwt ]]; then
  echo "$(date)  Indexing reference"
  bwa index "$REF"
fi
wait   # guarantee index exists before others start

# align, sort, index BAM
RG="@RG\tID:${SAMPLE}\tSM:${SAMPLE}\tPL:ILLUMINA"
SAMTOOLS_THREADS=$(( SLURM_CPUS_PER_TASK - 1 ))

bwa mem -t "$SLURM_CPUS_PER_TASK" -R "$RG" "$REF" "$R1" "$R2" \
  | samtools sort -@ "$SAMTOOLS_THREADS" -o "${OUT_DIR}/${SAMPLE}.sorted.bam" -

samtools index "${OUT_DIR}/${SAMPLE}.sorted.bam"

echo "$(date) finished $SAMPLE"
```

## SNP calling from the bam file created before and creating the bcf files.

```
#!/usr/bin/env bash
#SBATCH -D /work/agro932/niranjan27/Final_exam
#SBATCH -o /work/agro932/niranjan27/Final_exam/logs/snp-%A_%a.out
#SBATCH -e /work/agro932/niranjan27/Final_exam/logs/snp-%A_%a.err
#SBATCH -J snp_call_array                  #job name changed
#SBATCH -p schnablelab
#SBATCH --time=72:00:00
#SBATCH --array=0-9                        # 10 BAMs indexes 0-9
#SBATCH -c 16
#SBATCH --mem=64G
#SBATCH --mail-user=npokhrel3@huskers.unl.edu
#SBATCH --mail-type=BEGIN,END,FAIL

set -euo pipefail
```

```
module load samtools bcftools

# Directories
REF=/work/agro932/niranjan27/Reference/Zea_mays.Zm-B73-REFERENCE-NAM-5.0.dna.toplevel.fa
BAMS_DIR=/work/agro932/niranjan27/Final_exam/aligned
OUTDIR=/work/agro932/niranjan27/snp_calls_whole_genome
LOGDIR=/work/agro932/niranjan27/Final_exam/logs

mkdir -p "$OUTDIR" "$LOGDIR"

# BAM list & sample name
BAMS=(${BAMS_DIR}/*.sorted.bam)
BAM=${BAMS[$SLURM_ARRAY_TASK_ID]}
SAMPLE=$(basename "$BAM" .sorted.bam)

# index BAM if missing
[[ -f ${BAM}.bai ]] || samtools index "$BAM"

# SNP calling
bcftools mpileup -Ou -f "$REF" -a AD,DP "$BAM" \
| bcftools call -mv -Ob -o "$OUTDIR/${SAMPLE}.bcf"

bcftools index "$OUTDIR/${SAMPLE}.bcf"

echo "$(date) Finished SNP calling for ${SAMPLE}"
```

## Converting the bcf file of temperate maize lines to vcf files

```
module load bcftools vcftools
REF=/work/agro932/niranjan27/Reference/Zea_mays.Zm-B73-REFERENCE-NAM-5.0.dna.toplevel.fa

mkdir -p /work/agro932/niranjan27/Final_exam/merged_fst
cd       /work/agro932/niranjan27/Final_exam/merged_fst

bcftools merge /work/agro932/niranjan27/Final_exam/bcf_files/*.bcf \
-Oz -o temperate.raw.vcf.gz
bcftools index temperate.raw.vcf.gz

# Normalize temperate to same referene genome as tropical
bcftools norm -f "$REF" -m -both \
temperate.raw.vcf.gz -Oz -o temperate.norm.vcf.gz
bcftools index -f temperate.norm.vcf.gz

# Normalizing the tropical data with same reference genome to remmove any uncertainty

# bgzip + index the raw file once
bcftools view -Oz -o BGEM_trop.vcf.gz \
/work/agro932/niranjan27/Final_exam/Tropical_vcf/BGEM_15founders.recode.vcf
bcftools index -f BGEM_trop.vcf.gz

# normalise
```

```
bcftools norm -f "$REF" -m -both \
BGEM_trop.vcf.gz -Oz -o tropical.norm.vcf.gz
bcftools index -f tropical.norm.vcf.gz


# Since we had 15 tropical files, keeping only 10 vcf files of tropical one
# list the 15 IDs
bcftools query -l /work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical.norm.vcf.gz \
> tropical_all.txt

nano tropical_all.txt            # delete any 5 IDs, save as tropical10.txt

bcftools view -S tropical10.txt -Oz \
-o /work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical10.norm.vcf.gz \
/work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical.norm.vcf.gz
bcftools index -f /work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical10.norm.vcf.gz


# Merging the tropical and temperate individuals to create single vcf file

mkdir -p /work/agro932/niranjan27/Final_exam/merged_fst

bcftools merge -Oz \
-o /work/agro932/niranjan27/Final_exam/merged_fst/maize_merged.vcf.gz \
/work/agro932/niranjan27/Final_exam/temperate_vcf/temperate.norm.vcf.gz \
/work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical10.norm.vcf.gz

bcftools index /work/agro932/niranjan27/merged_fst/maize_merged.vcf.gz

# Here the single vcf file created to do population genomics related analysis
```

### Calculating genome wide Fst values

```
# Calculate per SNP Fst

vcftools --gzvcf maize_merged.vcf.gz \
--weir-fst-pop temperate.txt \
--weir-fst-pop tropical.txt \
--out temp_vs_trop_10vs10

# After filtering, kept 20 out of 20 Individuals
# Outputting Weir and Cockerham Fst estimates.
# Weir and Cockerham mean Fst estimate: 0.39787
# Weir and Cockerham weighted Fst estimate: 0.52008
# After filtering, kept 50278096 out of a possible 50278096 Sites
```

### Plotting the fst value computed

```r
library(data.table)
library(ggplot2)


# Load data
fst_data <- fread("/work/agro932/niranjan27/Final_exam/merged_fst/temp_vs_trop_win50k.windowed.weir.fst

# Plot histogram of fst value
hist(fst_data$MEAN_FST,
     breaks = 50,
     col = "steelblue",
     main = "Histogram of FST values",
     xlab = "Mean FST",
     ylab = "Number of windows")

# Plotting the manhatten plot of Fst value

# Clean: replace NA and negative FST values
fst_data <- fst_data %>%
  filter(!is.na(MEAN_FST)) %>%
  mutate(MEAN_FST = ifelse(MEAN_FST < 0, 0, MEAN_FST))

# Ensure CHROM is treated as an ordered factor
fst_data$CHROM <- factor(fst_data$CHROM, levels = mixedsort(unique(as.character(fst_data$CHROM))))

# Calculate chromosome lengths and offsets
chr_lengths <- fst_data %>%
  group_by(CHROM) %>%
  summarize(chr_len = max(BIN_START)) %>%
  mutate(chr_start = lag(cumsum(chr_len), default = 0))

# Merge back to data and compute cumulative position
fst_data <- fst_data %>%
  left_join(chr_lengths, by = "CHROM") %>%
  mutate(cum_pos = BIN_START + chr_start)

# Midpoint for chromosome labels
axis_df <- chr_lengths %>%
  mutate(mid = chr_start + chr_len / 2)


ggplot(fst_data, aes(x = cum_pos, y = MEAN_FST)) +
  geom_point(size = 0.3, alpha = 0.6, color = "blue") +
  scale_x_continuous(
    label = axis_df$CHROM,
    breaks = axis_df$mid,
    expand = expansion(mult = c(0.01, 0.01))
  ) +
  labs(
    title = "Sliding Window FST (Tropical vs Temperate Maize)",
    x = "Chromosome",
    y = "Mean FST"
  ) +
```

```
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, size = 8),
    plot.title = element_text(hjust = 0.5)
  )
```

### Calculating the nucleotide diversity

```
# Run VCF tool for nucleotide diversity

# Tropical
vcftools --gzvcf /work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical10.norm.vcf.gz \
--keep tropical.txt \
--window-pi 50000 \
--window-pi-step 25000 \
--out tropical_pi

# Temperate
vcftools --gzvcf /work/agro932/niranjan27/Final_exam/temperate_vcf/temperate.norm.vcf.gz \
--keep temperate.txt \
--window-pi 50000 \
--window-pi-step 25000 \
--out temperate_pi


#Calculate nucleotide diversity.

trop <- read.table("/work/agro932/niranjan27/Final_exam/merged_fst/tropical_pi.windowed.pi", header=TRU
temp <- read.table("/work/agro932/niranjan27/Final_exam/merged_fst/temperate_pi.windowed.pi", header=TRU

mean(trop$PI, na.rm=TRUE)  # Tropical
mean(temp$PI, na.rm=TRUE)  # Temperate
```

### Plotting nucleotide diversity

```
library(data.table)
library(ggplot2)
library(tidyverse)

# Load data
trop <- fread("/work/agro932/niranjan27/Final_exam/merged_fst/tropical_pi.windowed.pi")
temp <- fread("/work/agro932/niranjan27/Final_exam/merged_fst/temperate_pi.windowed.pi")


# Add population info
trop$pop <- "Tropical"
temp$pop <- "Temperate"

# Combine
```

```r
pi_all <- rbind(trop, temp)

# Convert CHROM to numeric
pi_all$CHROM <- as.numeric(as.character(pi_all$CHROM))

# Sort by chromosome and position
pi_all <- pi_all %>% arrange(CHROM, BIN_START)

# Calculate cumulative position for plotting
chr_lengths <- pi_all %>%
  group_by(CHROM) %>%
  summarise(chr_len = max(BIN_END)) %>%
  mutate(chr_start = lag(cumsum(chr_len), default = 0))

pi_all <- left_join(pi_all, chr_lengths, by = "CHROM")
pi_all$position <- pi_all$BIN_START + pi_all$chr_start

# Plot as a continuous line graph
ggplot(pi_all, aes(x = position, y = PI, color = pop)) +
  geom_line(alpha = 0.6, size = 0.5) +
  scale_color_manual(values = c("Tropical" = "red", "Temperate" = "blue")) +
  labs(x = "Genomic Position", y = "Nucleotide Diversity ( )",
       title = "Nucleotide Diversity Across Genome",
       color = "Population") +
  theme_minimal(base_size = 14) +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())
```

## Calculating Tajima's D value

```
# Temperate
vcftools --gzvcf /work/agro932/niranjan27/Final_exam/temperate_vcf/temperate.norm.vcf.gz \
--TajimaD 10000 \
--out temperate_tajima

# Tropical
vcftools --gzvcf /work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical10.norm.vcf.gz \
--TajimaD 10000 \
--out tropical_tajima
```

## Plotting tajima's D values

```r
library(data.table)
library(ggplot2)
library(dplyr)
library(gtools)

# TROPICAL LINES
# Load Tajima's D data
```

```r
tajima <- fread("/work/agro932/niranjan27/Final_exam/Tropical_vcf/tropical_tajima.Tajima.D")
colnames(tajima) <- c("CHROM", "BIN_START", "N_SNPS", "TajimaD")

# Remove extreme values for better visualization (optional)
tajima <- tajima %>% filter(TajimaD > -5, TajimaD < 5)

# Ensure CHROM is treated as a factor and sorted naturally
tajima$CHROM <- factor(tajima$CHROM, levels = mixedsort(unique(tajima$CHROM)))

# Calculate chromosome lengths and offsets
chr_lengths <- tajima %>%
  group_by(CHROM) %>%
  summarize(chr_len = max(BIN_START)) %>%
  mutate(chr_start = lag(cumsum(chr_len), default = 0))

# Merge back to get cumulative positions
tajima <- tajima %>%
  left_join(chr_lengths, by = "CHROM") %>%
  mutate(cum_pos = BIN_START + chr_start)

# Midpoint of each chromosome for axis labeling
axis_df <- chr_lengths %>%
  mutate(mid = chr_start + chr_len / 2)

# Plot
ggplot(tajima, aes(x = cum_pos, y = TajimaD)) +
  geom_point(size = 0.4, color = "darkred", alpha = 0.6) +
  scale_x_continuous(label = axis_df$CHROM, breaks = axis_df$mid) +
  labs(
    x = "Chromosome",
    y = "Tajima's D",
    title = "Tajima's D Across Genome (Tropical Maize)"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 0, size = 8),
    plot.title = element_text(hjust = 0.5)
  )

# TEMPERATE LINES

# Load Tajima's D data for temperate maize
tajima <- fread("/work/agro932/niranjan27/Final_exam/temperate_vcf/temperate_tajima.Tajima.D")
colnames(tajima) <- c("CHROM", "BIN_START", "N_SNPS", "TajimaD")

# Remove extreme values (optional)
tajima <- tajima %>% filter(TajimaD > -5, TajimaD < 5,N_SNPS >= 10)

# Ensure chromosomes are treated as ordered factors
tajima$CHROM <- factor(tajima$CHROM, levels = mixedsort(unique(tajima$CHROM)))

# Calculate cumulative genome positions
chr_lengths <- tajima %>%
```

```
  group_by(CHROM) %>%
  summarize(chr_len = max(BIN_START)) %>%
  mutate(chr_start = lag(cumsum(chr_len), default = 0))

# Merge and compute cumulative positions
tajima <- tajima %>%
  left_join(chr_lengths, by = "CHROM") %>%
  mutate(cum_pos = BIN_START + chr_start)

# Axis labels
axis_df <- chr_lengths %>%
  mutate(mid = chr_start + chr_len / 2)

# Plot
ggplot(tajima, aes(x = cum_pos, y = TajimaD)) +
  geom_point(size = 0.4, color = "blue", alpha = 0.6) +
  scale_x_continuous(label = axis_df$CHROM, breaks = axis_df$mid) +
  labs(
    x = "Chromosome",
    y = "Tajima's D",
    title = "Tajima's D Across Genome (Temperate Maize)"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 0, size = 8),
    plot.title = element_text(hjust = 0.5)
  )
```

### Genome Wide Association Analysis

```
# Loading the rMVP package for gwas analysis
install.packages("rMVP")
library(rMVP)


# converting genotype and phenotype data into rmvp readable file

MVP.Data(fileHMP="Data/BGEM_MMC_imputed_181entries_62077markers_perse.hmp.txt",
         filePhe="Data/Phenotypic_data.csv",
         sep.phe=";",
         fileKin=FALSE,
         filePC=FALSE,
         #maxLine=10000,
         out="Niranjan.mvp.hmp"
)

pheno<- read.csv("Assignment_3_midterm/Data/Phenotypic_data.csv")

# loading the genotype and phenotype
genotype <- attach.big.matrix("/home/agro932/niranjan27/Genetics-Assignment/Assignment_3_midterm/Niranj
View(genotype)
phenotype <- read.table("/home/agro932/niranjan27/Genetics-Assignment/Assignment_3_midterm/Niranjan.mvp
```

```r
View(phenotype)
map <- read.table("/home/agro932/niranjan27/Genetics-Assignment/Assignment_3_midterm/Niranjan.mvp.hmp.g


#Running GWAS

# Set output directory
setwd("/home/agro932/niranjan27/Genetics-Assignment/Final_term/Plots")

trait <- phenotype[c("Taxa", "SIL_LN")] #can change the trait here
imMVP <- MVP(
  phe=trait,          #NA is acceptable in phenotype
  geno=genotype,
  map=map,             #if you have pre-computed GRM, please keep there open, otherwise rMVP will compu
  #CV.GLM=Covariates,     #if you have environmental covariates, please keep all 'CV.*' open
  #CV.MLM=Covariates,
  #CV.FarmCPU=Covariates,
  nPC.GLM=5,              #if you have added PCs into covariates, please keep there closed
  nPC.MLM=3,              #if you don't want to add PCs as covariates, please comment out the parameter
  nPC.FarmCPU=3,
  maxLine=10000,          #smaller value would reduce the memory cost
  #ncpus=10,
  vc.method="BRENT",      #only works for MLM
  method.bin="static",    # "FaST-LMM", "static" (#only works for FarmCPU)
  threshold=0.05,
  method=c("FarmCPU"),   #can adjust GLM and MLM here
  file.output=c("pmap", "pmap.signal", "plot", "log")
)

## Note: we had six phenotypic trait to run GWAS so in this line of code: "trait <- phenotype[c("Taxa",
```