Project Report

Dan Jin. Nov. 20, 2012

GWAS

Procedure:

1. **Import genotype and phenotype.** Check if phenotype follows normal distribution (Fig1).
2. **Pre-test:** To get a general idea about the potential hit(s).
   A. **Filter genotype and phenotype data with minimum criteria:**
      a. Only remove individuals with **100%** missing data across all genotypes or without phenotypes using `threshold.geno = (num.alleles - 2)/num.alleles`: Individuals have genotype information of at least one marker (2 alleles)
         i. `137 individuals( 84.5679 %) have missing genotype data <= 99.98979 % with phenotype data.`
      b. Only remove genotypes that have **no data** at all using `threshold.indi=(sample.size-1)/sample.size`: Otherwise, it will cause problem when filtering genotypes with MAF in the next step.
         i. `19598 SNPs ( 100 %) have missing individuals <= 99.27007 %.`
      c. Only remove genotypes with **MAF=0** using `threshold.MAF=1-(sample.size-1)/sample.size`: Otherwise, it will cause problem when calculate p value (system is exactly singular).
         i. `18374 SNPs ( 93.75446 %) have MAF >= 0.729927 %.`
      d. Remove phenotypes with no associated individuals.
      e. After filtering, there were **137 individuals** left with 18374 alleles (**9187 markers**).
   B. **Check phenotype distribution again:** Filtered phenotypes looked normally distributed. So didn't preform any transformation (Fig2).
   C. **Convert genotype to Xa:** There are no heterozygotes, so no need to generate Xd. In addition, having Xd column (will be a column of "-1") will cause problem when calculate p value (system is exactly singular).
   D. **Perform a PCA to detect population structure:** The plot showed that there might have two populations. The histogram of component 1 indicated that the two groups could be separated at Comp.1=0 (Fig3, 4).
   E. **Create covariate matrix Xz** using `threshold.comp1=0`: There were 55 individuals in group 1 and 82 individuals in group 2.

F. **Perform GWAS using linear regression model without covariant**: Calculate p value for each markers.
    a. **Manhattan plot:** Draw Manhattan plot with Bonferroni corrected type I error. Only one marker had p value significant than Bonferroni corrected type I error (Fig5).
    b. **QQ plot:** QQ looked fine. But I will try some other analysis strategies (with covariate, for example.) (Fig6)
G. **Perform GWAS using linear regression model with covariant:**
    a. **Manhattan plot:** Draw Manhattan plot with Bonferroni corrected type I error. Again, only one marker had p value significant than Bonferroni corrected type I error (Fig7).
    b. **QQ plot:** QQ plot looked similar to the one without covariate. It may indicate that population structure is not a big issue in this case (Fig8).

3. **GWAS Test with more restricted condition:**
    A. Set threshold for filter:
        a. Plot percentage of missing data for each person (Fig9), set `threshold.geno = 0.25` (remove individuals with >25% missing genotype)
        b. Plot percentage of missing individulas for each genotype (Fig10), set `threshold.indi = 0.15` (remove genotypes with >15% missing individuals)
        c. Plot MAF for each genotype (Fig11), set `threshold.MAF = 0.05` (remove individuals with <5% MAF)
        d. Remove phenotypes with no associated individuals.
        e. After filtering, there were **130 individuals** left with 4686 alleles (**2343 markers**).
    B. **Check phenotype distribution again** (Fig12)
    C. **Convert genotype to Xa:** There are no heterozygotes, so no need to generate Xd.
    D. **Perform a PCA to detect population structure:** The plot showed that most individuals clustered together. The population structure may not be a big issue in this case. The histogram of component 1 indicated that the two groups could be separated at Comp.1=0.1 (Fig13).
    E. **Create covariate matrix Xz** using `threshold.comp1=0.1`: There were 117 individuals in group 1 and 13 individuals in group 2.
    F. **Perform GWAS using linear regression model without covariant**: Calculate p value for each marker.

a. **Manhattan plot:** Draw Manhattan plot with Bonferroni corrected type I error. Only one marker had p value significant than Bonferroni corrected type I error (Fig14).
b. **QQ plot:** QQ looked fine(?) But I will try some other analysis strategies (with covariate, for example) (Fig15).

G. **Perform GWAS using linear regression model with covariant:**
   a. **Manhattan plot:** Draw Manhattan plot with Bonferroni corrected type I error. Again, only one marker had p value significant than Bonferroni corrected type I error (Fig17).
   b. **QQ plot:** QQ plot looked similar to the one without covariate. It again indicated that population structure is not a big issue in this case (Fig18).

H. Find the position of the hit.
   a. Zoomed in Manhattan plot (Fig16).
   b. Find the its location on chromosome and SNP position.

```
      Chr       SNP      Chr.SNP
2004    1 12477222 2L.12477222
```

   c. Database and literature search for more information about this hit (Please refer to *Discussion*).

Discussion:

I have used linear regression model in this GWAS. Although I did a PCA and GWAS without and with covariate to calculate p value, it seems that the population structure was not a big issue in this case.

Only one marker (**2L.12477222**) has p value significant than Bonferroni corrected type I error in this GWAS, but it looks quite promising. In zoomed in Manhattan plot, the markers adjacent to 2L.12477222 also have higher –log10(p) compared with the markers further away from 2L.12477222, although not higher than Bonferroni corrected type I error. This indicates that the high –log(p) of the potential hit (2L.12477222) is not likely due to genotyping error.

According to Flybase, 2L.12477222 locates within a gene called *bunched (bun).* It has been identified as a positive growth regulator in Drosophila, necessary for cellular growth, proliferation and survival. This gene is required during *Drosophila* development, by regulating peripheral nervous system development [1] and segmental patterning [2]. It also controls photoreceptor patterning during eye development [3]. A couple of mutations have been found in this gene that cause wing development defect. Homozygous mutant clones in the wing disc resulted in

fewer cells than their corresponding wild-type sister clones. And the size of single cell also slightly decreases in mutant clones [4].

Another study demonstrated that *bun* regulates Notch signaling in follicle cells during late oogenesis [5]. It has been reported that Notch-mediated lateral signaling between neighboring cells is crucial to the formation of the boundary between dorsal and ventral fates during late wing development. And a couple of hetero-allelic *bun* mutants did show defects in wing venation [5].

However, the SNP (2L.12477222) found in this GWAS locates in intron rather than in coding region, which means this SNP will not change *bun* protein sequence. It may affect the phenotype by regulating gene expression through ribozyme functionality or mRNA alternative splicing.

In summary, the SNP found in this GWAS may be a good candidate that controls wing development. Further studies may focus on the mechanism of how this SNP affects wing development and whether Notch signaling is involved in this process.

Reference:

[1] Kania, A et al., P-element mutations affecting embryonic peripheral nervous system development in Drosophila melanogaster. Genetics. (1995)

[2] N. Perrimon et al., Zygotic lethal mutations with maternal effect phenotypes in Drosophila melanogaster: II. Loci on the second and third chromosomes identified by P-element-induced mutations. Genetics, (1996)

[3] J.E. Treisman et al, Shortsighted acts in the decapentaplegic pathway in Drosophila eye development and has homology to a mouse TGF-β-responsive gene. Development. (1995)

[4] Gluderer et al., Bunched, the Drosophila homolog of the mammalian tumor suppressor TSC-22, promotes cellular growth. BMC Developmental Biology. (2008)

[5] Leonard Dobens et al., Bunched sets a boundary for Notch signaling to pattern anterior eggshell structures during Drosophila oogenesis. Developmental Biology. (2005)

- **Figures (Only show key plots here due to limitation of space.)**

Key plots from Pre-test	
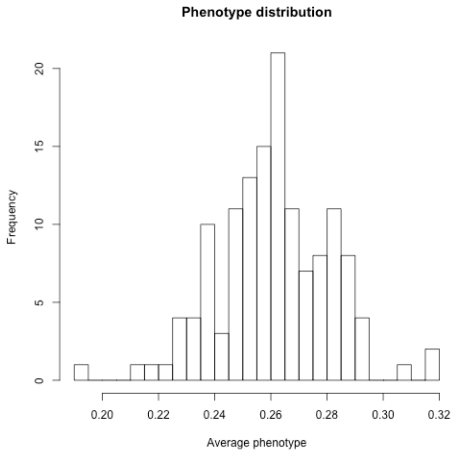	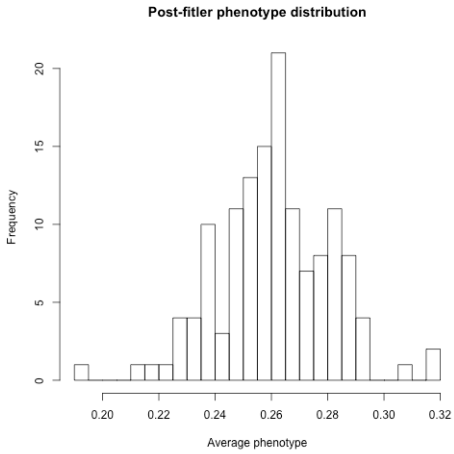
Fig1: Phenotype distribution after import	Fig2: Phenotype distribution after data filtering
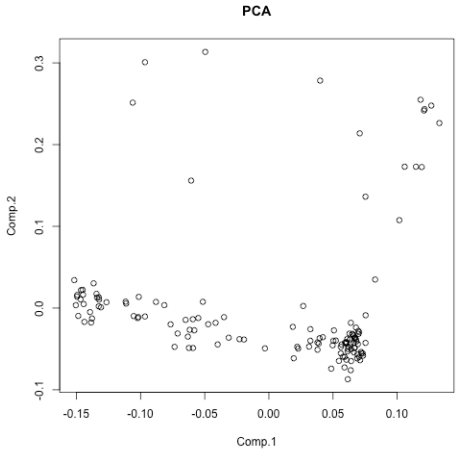	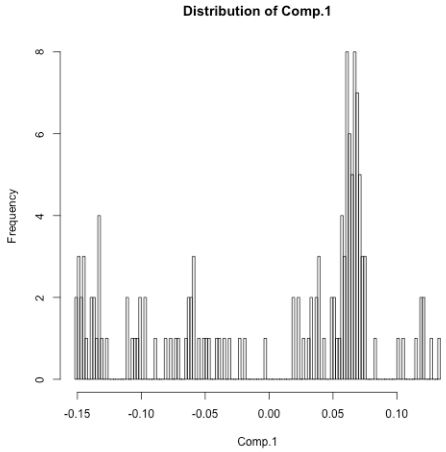
Fig3: PCA	Fig4: Distribution of PCA Component1
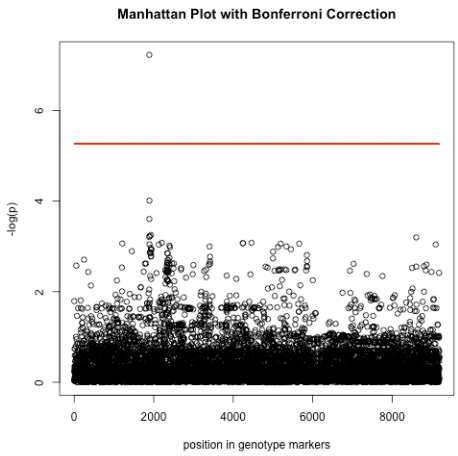	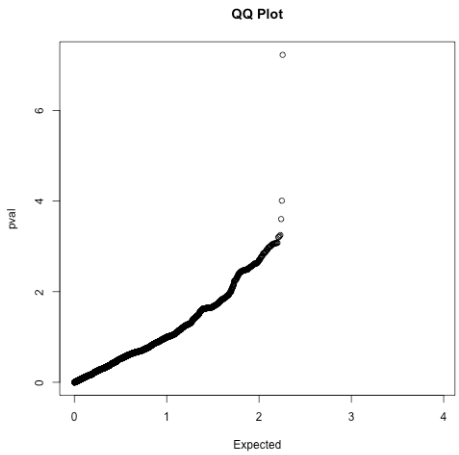
Fig5: Manhattan plot	Fig6: QQ plot

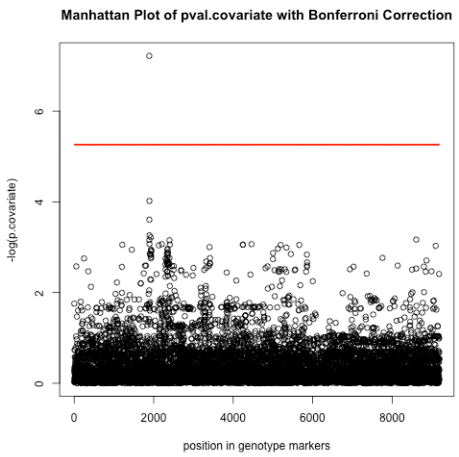

Fig7: Manhattan plot (using p values calculated with covariate)

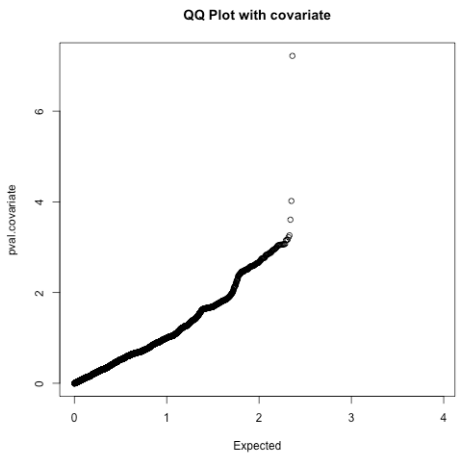

Fig8: QQ plot (using p values calculated with covariate)

Key plots from GWAS

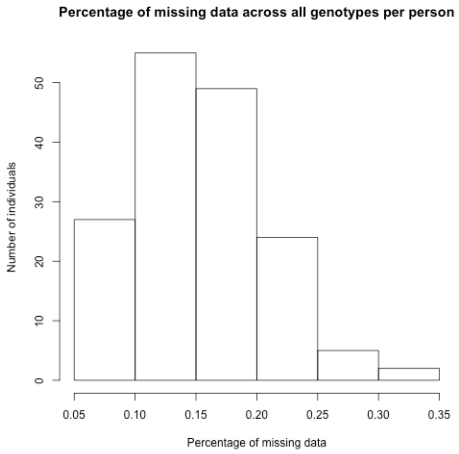

Fig9: Distribution of Percentage of missing data/person before filter

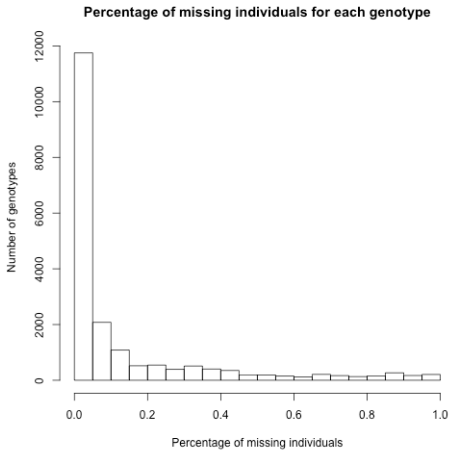

Fig10: Distribution of Percentage of missing data/genotype before filter

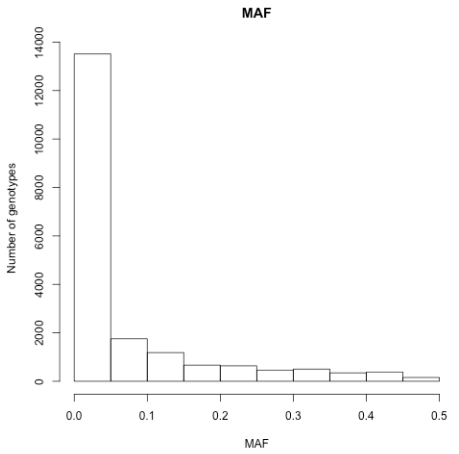

Fig11: Distribution of MAF for each genotype before fitler

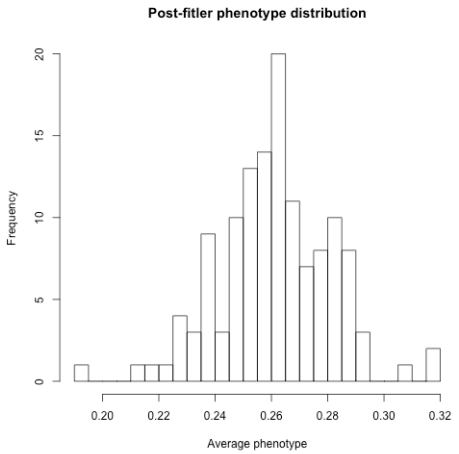

Fig12. Phenotype distribution after data filtering

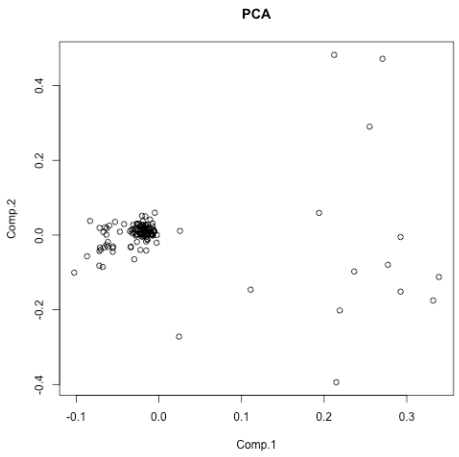

Fig13: PCA

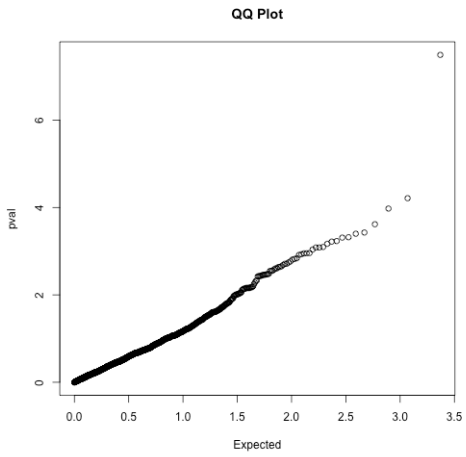

Fig14: QQ plot

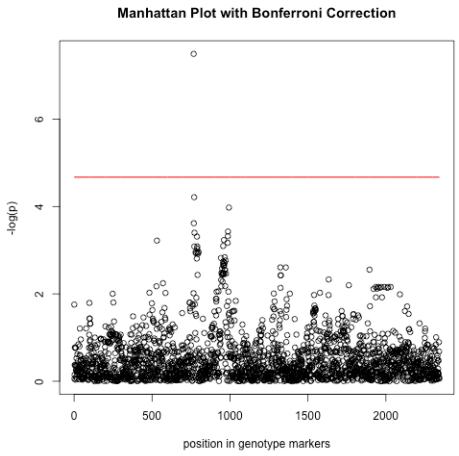

Fig15: Manhattan plot

Fig16: Zoomed in Manhatton plot

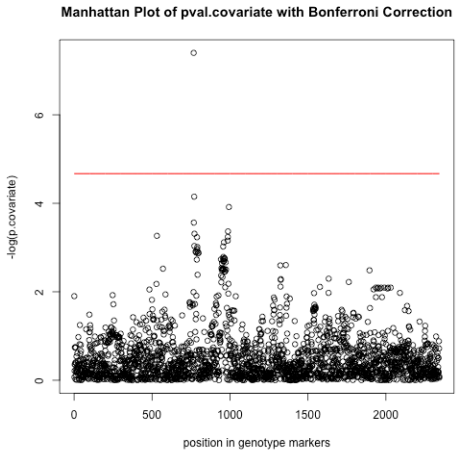

Fig17: Manhattan plot (using p values calculated with covariate)

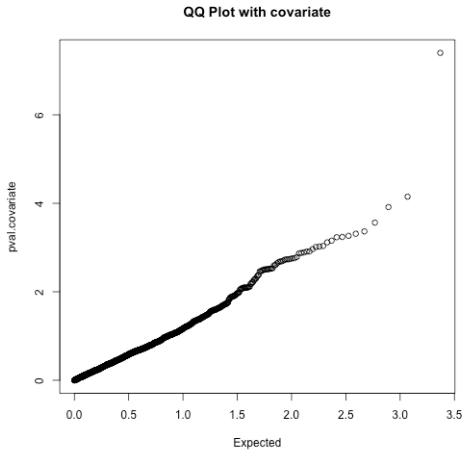

Fig18: QQ plot (using p values calculated with covariate)

A brief instruction of using ProjectScript.r

To run this script, please copy ProjectScript.r to the same fold where containing the input data files (project_genotype.txt, project_phenotype.txt, project_information.txt), then type the following lines in R.

```
source("ProjectScript.r")
result.ls=list()
```

For Pre-test (Take really very long time to code Xa):

```
result.ls=control.fun(-1,-1,-1,0)
```

For GWAS:

```
result.ls=control.fun(0.25,0.15,0.05,0.1)
```

Note: Please move Pre-test outputs to a new folder before preforming GWAS. Otherwise, Pre-test outputs will be overwritten by GWAS outputs.

Result of this script:

1. Create two .txt files containing filtered genotype and phenotype data.
2. Return one list, result.ls. See below for more details about the contents in the list.
3. Create 11 plots in .png format. See below for more details about each .png file.

Summary of the content in result.ls:

1. $pval: p values calculated without covariate
2. $pval.covariate: p values calculated with covariate
3. $hits.ls: all genotype marker (marker number and their p value) significant than Bonferroni correction
4. $hits.info: chromosome and SNP position information for each hit in hits.ls

Structure of result.ls:

```
result.ls
|-$pval
|-$pval.covariate
|-$hits.ls
| |-$position
| |-$pval
|-$hits.info
  |-$Chr
  |-$SNP
  |-$Chr.SNP
```

List of .png image files:

1. 01_Phenotype distribution.png
2. 02_Percentage of missing data.png
3. 03_Percentage of missing individuals.png
4. 04_MAF.png
5. 05_Post-filter phenotype distribution.png
6. 06_PCA.png
7. 07.1_QQ Plot.png
8. 07.2_QQ Plot with covariate.png
9. 08.1_Manhattan Plot with Bonferroni correction.png
10. 08.2_Manhattan Plot with covariate with Bonferroni correction.png
11. 09_Zoomed in manhattan Plot with Bonferroni correction.png