Cjelogenomske asocijacijske studije (GWAS) u PLINKu Sveučilišni računski centar Srce

Nikolina Pleić

mag.math.
Sveučilište u Splitu, Medicinski fakultet
Zavod za biologiju i humanu genetiku

23. listopada 2023.

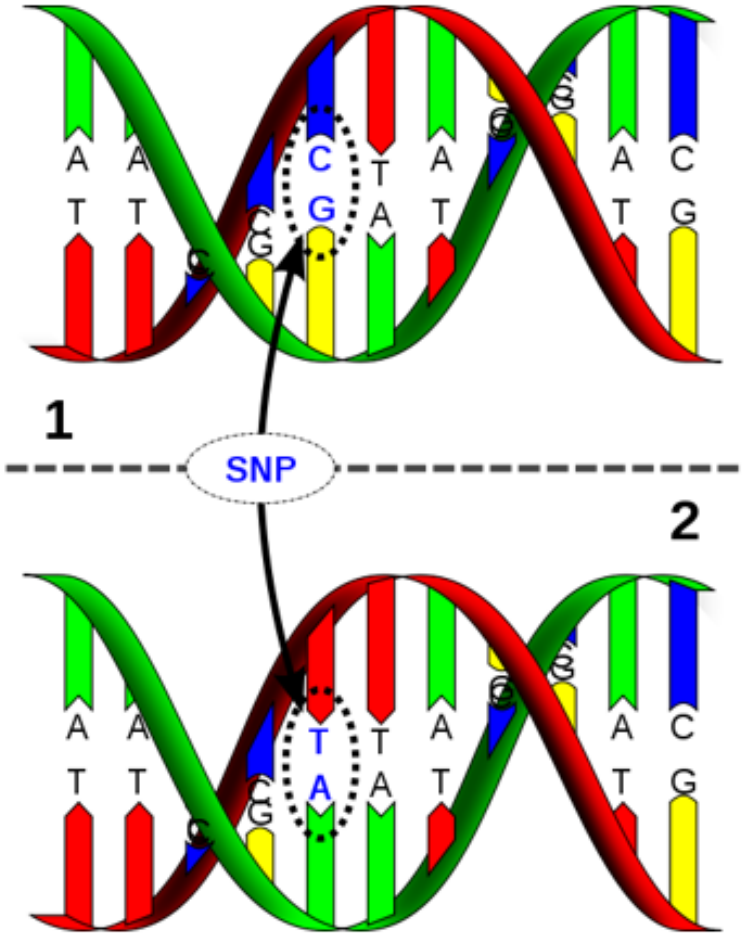
GWAS

- Sekvenca ljudskog genoma se sastoji od 3.2 milijarde nukleotida (baznih parova tj. DNA 'slova' A,C,G,T)
- Ostaje konstantna kroz cijeli život pa je idealna polazna točka za znanstvena istraživanja
- Cjelogenomske asocijacijske studije (GWAS) istražuju statističku povezanost između genetičke varijacije i fenotipova u populaciji

Izvor: National Human Genome Research Institute (NHGRI) LinkedIn

GWAS

- Na bilo kojoj poziciju u genomu (lokus) mogu postojati razlike između genoma u populaciji
- Kromosom nasljeđen od majke: baza G
- Kromosom nasljeđen od oca: baza A
- Ove dvije verzije se nazivaju aleli
- Ovakva varijacija se naziva polimorfizam jednog nukleotida (eng. single nucleotide polymorphism, SNP)
- Postoji preko 10 milijuna SNP-ova u ljudskom genomu



GWAS

Pod **fenotipom** podrazumijevamo vanjski izgled nekog organizma, kakav se razvio pod utjecajem vanjskih čimbenika (vlaga, svjetlo, temperatura, hrana) zajedno s nasljednim osobinama (**genotip**).

- **Monogenski fenotip:** Određen jednim genom.
- **Oligogenski fenotip:** Pod utjecajem nekoliko gena.
- **Poligenski fenotip:** Pod utjecajem mnogih genetičkih varijanti.
- **Složeni fenotip (Complex trait):** Fenotip koji nije monogenski fenotip. Obično se radi o poligenskom fenotipu koji je također pod utjecajem mnogih okolišnih čimbenika.
- **Uobičajena bolest (Common disease):** Bolest/stanje koje je uobičajeno u populaciji (recimo, prevalencija od 0,1% ili više). Primjeri: MS (prevalencija reda veličine 0,1%), shizofrenija ($\sim$1%) ili dijabetes tipa 2 ($\sim$10%).

GWAS

- Obično su SNP-ovi bialelični, odnosno u populaciji su prisutna samo dva alela i to pretpostavljamo u GWASu
- U principu, međutim, sva četiri moguća alela SNP-a mogu biti prisutna u populaciji
- **Učestala varijanta (Common variant):** Ima frekvenciju barem 1% u populaciji (također se koristi i 5% kao granica).
- **Nisko-frekventna varijanta:** Ima frekvenciju barem 0.1% u populaciji i ta frekvencija je niža od frekvencije učestale varijante.
- **Rijetka varijanta:** Ima frekvenciju nižu od nisko-frekventne varijante.

GWAS

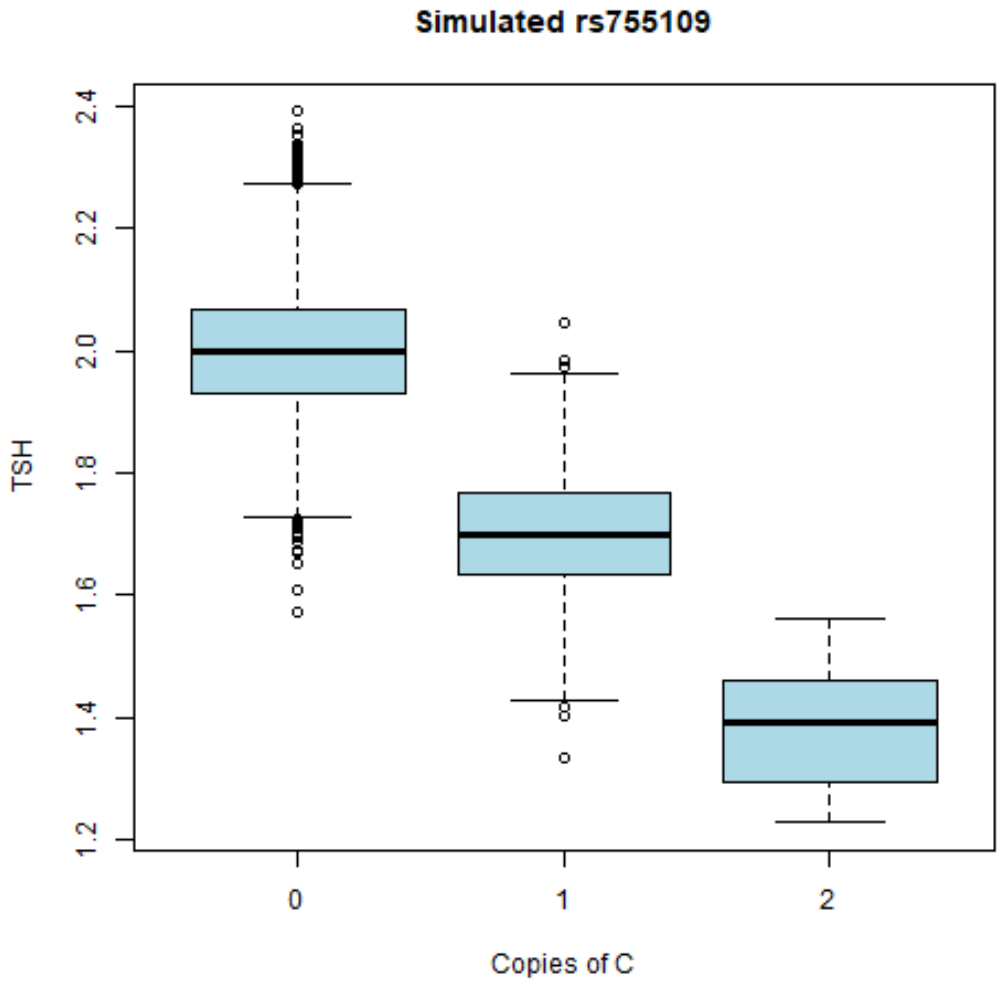
Dvije glavne vrste GWAS-a proučavaju:

- kvantitativne fenotipove (npr. visina, težina, razina kolesterola u krvi): **qt-GWAS**
- kategoričke fenotipove (bolesti), obično predstavljene kao binarne varijable: **case-control GWAS**

Poligenske bolesti i poligenski fenotipovi su pod utjecajem velikog broja genetičkih varijanti koje imaju mali efekt, u kombinaciji s čimbenicima iz okoliša.

GWAS

Pogledajmo primjer simuliranog fenotipa: razine TSH.



Iz grafa vidimo da svaka dodatna kopija alela C smanjuje razinu TSH.

Aditivni genetički model

Najjednostavniji način za statističku analizu naših podataka je korištenje *aditivnog modela* koji pretpostavlja da srednje vrijednosti skupina ovise aditivno o broju alela 1 (obično manje frekventnog, tj. efektnog alela) u genotipu, te da su standardne devijacije skupina genotipova konstantne. Dakle, koristimo linearni model:

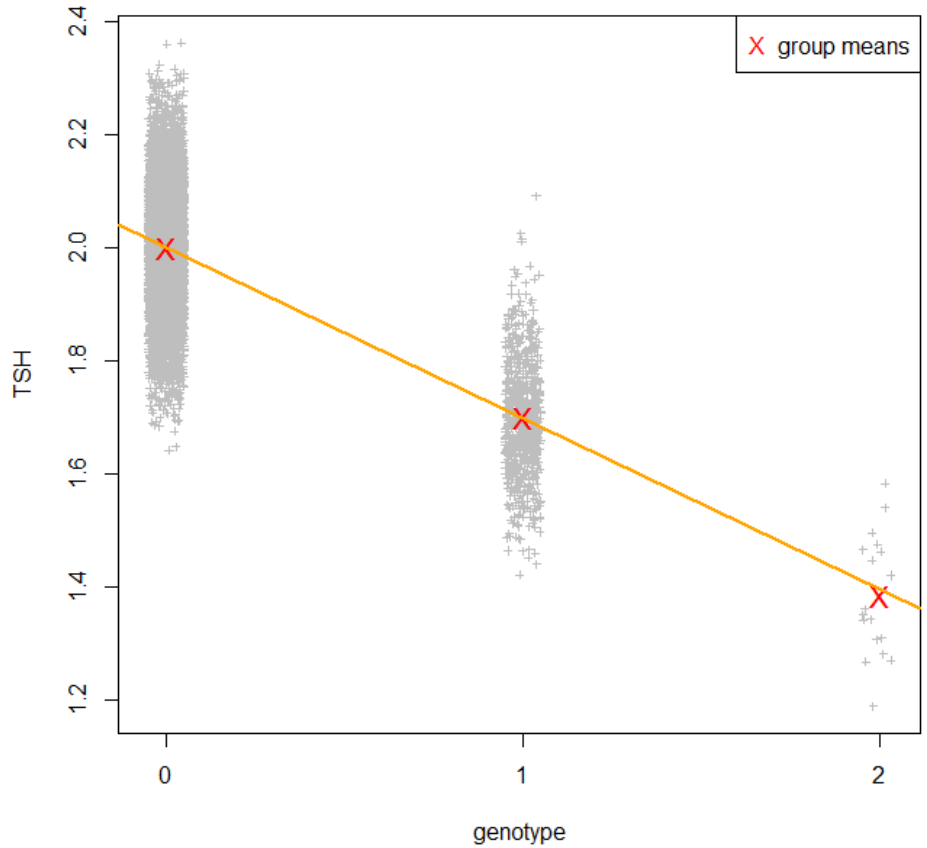
$$y = \mu + \beta x + \varepsilon, \tag{1}$$

gdje je y fenotip, x je genotip (0,1 ili 2) i parametri koji trebaju biti procijenjeni su:

- μ, srednja vrijednost genotipa 0
- β, učinak svake kopije alela 1 na srednju vrijednost fenotipa.

Pretpostavlja se da greške ε imaju identičnu normalnu distribuciju $N(0, \sigma^2)$ gdje σ^2 nije poznata i bit će procijenjena iz podataka.

GWAS



Možemo vidjeti statistički značajnu povezanost između genotipa i fenotipa gdje svaka kopija alela C snižava TSH za 0.3 jedinice.

GWAS

U praksi se GWAS analize za studije slučajeva i kontrola (case-control) provode pomoću regresijskih modela jer imaju mogućnost uzimanja u obzir zbunjujućih kovarijata.

Logistička regresija

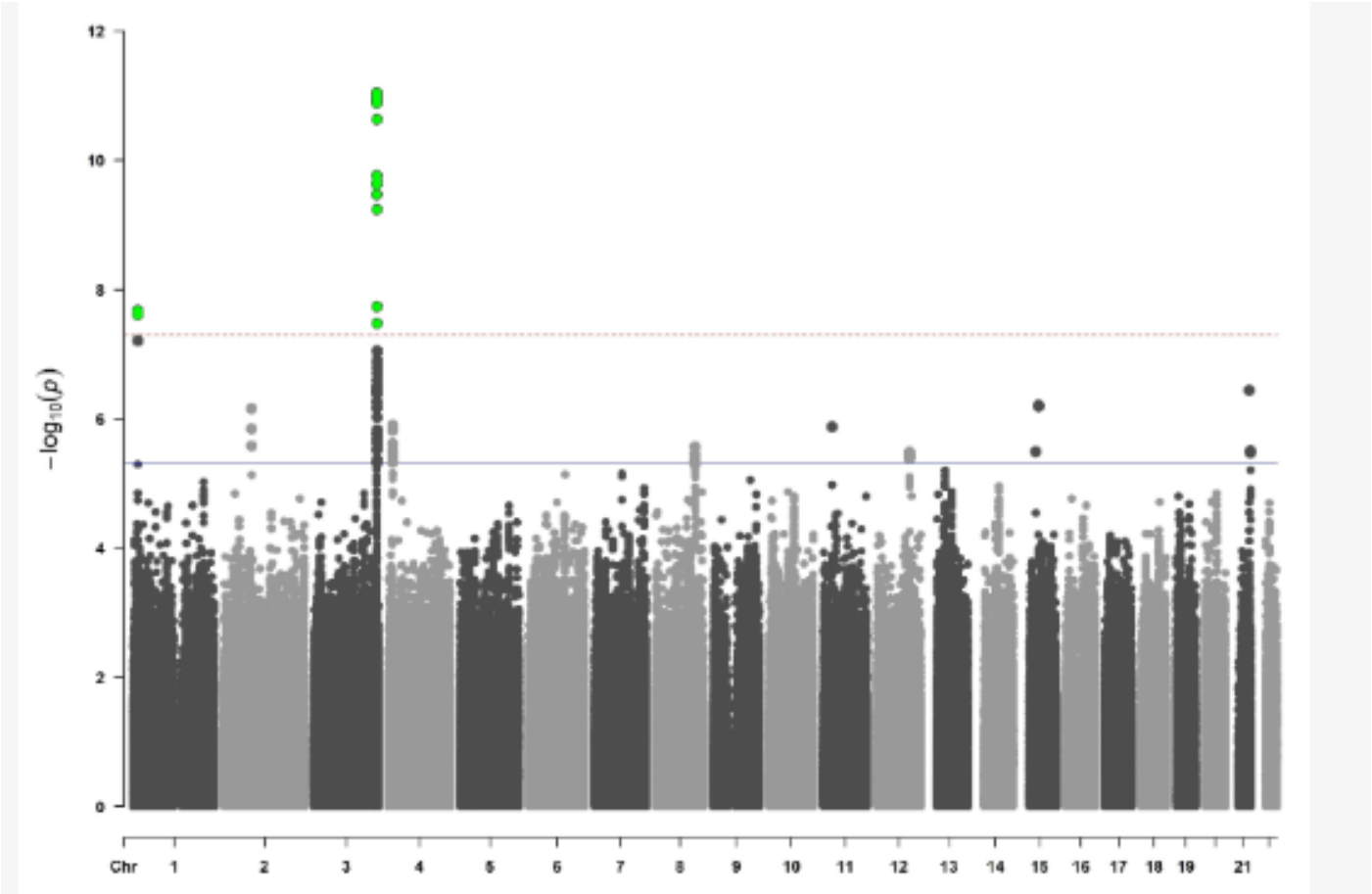
Model logističke regresije zamjenjuje linearnu regresiju kao osnovni GWAS model kada je fenotip binaran. Zavisna varijabla je sada logaritam relativnog rizika. Najjednostavniji model je aditivni model:

$$\log\left(\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)}\right) = \mu + \beta x.$$

Stoga je μ logaritam izgleda za genotip 0, a β je logaritam omjera izgleda (logOR) između genotipa 1 i 0 (a $\exp(\beta)$ je odgovarajući omjer izgleda). Slično tome, 2β je logOR između genotipa 2 i 0.

Vizualizacija GWAS rezultata

Primjer manhattan grafa:



Pleić, N.; Babić Leko, M.; Gunjača, I.; Boutin, T.; Torlak, V.; Matana, A.; Punda, A.; Polašek, O.; Hayward, C.; Zemunik, T. Genome-Wide Association Analysis and Genomic Prediction of Thyroglobulin Plasma Levels. Int. J. Mol. Sci. 2022, 23, 2173. https://doi.org/10.3390/ijms23042173

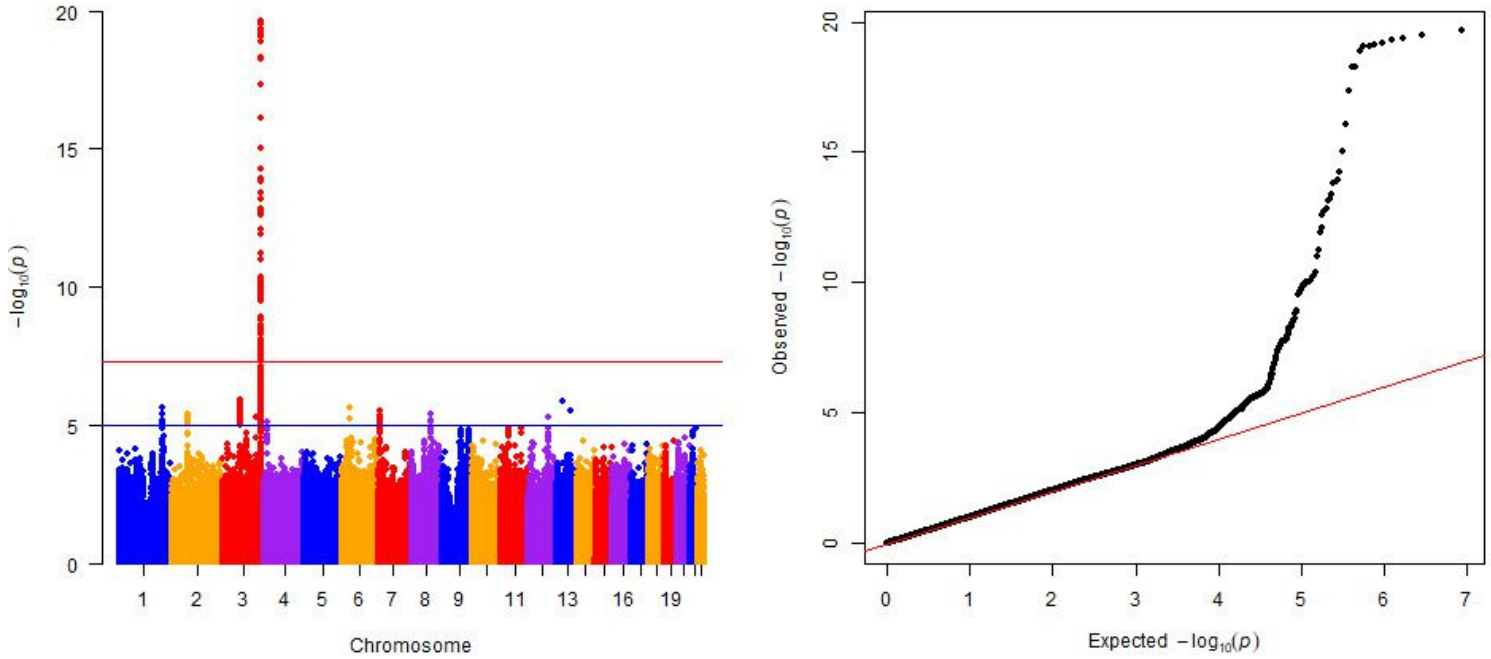
GWAS

Os x predstavlja kromosomski položaj SNP-ova, a os y predstavlja njihove $-log_{10}$(p-vrijednost) dobivene analizom.

Svaka točka na Manhattan grafu označava SNP. Budući da najsnažnije povezani SNPovi imaju najmanje p-vrijednosti (npr. 10^{-12}), njihovi negativni logaritmi bit će najveći (npr. 12).

Crvena horizontalna linija označava prag značajnosti (genome-wide significance level) ($p = 5 \times 10^{-8}$), dok plava horizontalna linija označava indikativni prag značajnosti ($p = 5 \times 10^{-6}$).

GWAS



Slika: **Manhattan graf i Q-Q graf**

Na Q-Q grafu, vidimo snažnu devijaciju od nulte distribucije (distribucija p-vrijednosti kada je nulta hipoteza o nepostojanju asocijacije istinita, koja je označena crvenom linijom).

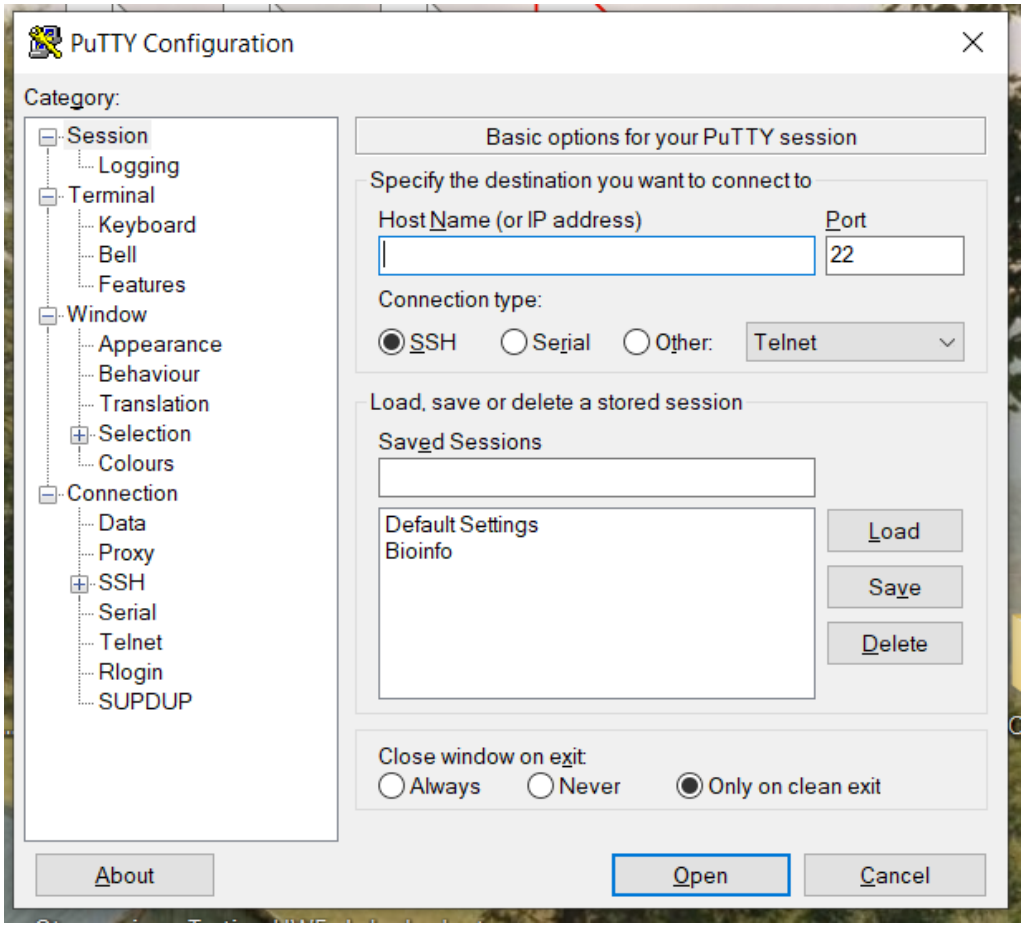
Kratak uvod u Unix/Linux i komandnu liniju

Komandna linija (eng. command line interface, CLI) je način upravljanja računalom unosom tekstualnih naredbi koje se predaju operacijskom sustavu za izvršavanje.

Prvo ćete naučiti kako se prijaviti na Linux poslužitelj koji ćemo koristiti tijekom tečaja. Nakon toga proći ćemo kroz neke osnovne Linux naredbe za navigaciju kroz mape i manipulaciju datotekama.

Spajanje na poslužitelj

Ukoliko nemate instaliran PuTTY, možete ga preuzeti s:
https://www.putty.org/
IP adresa: 161.53.133.104



Linux naredbe

1. Navigacija kroz Linux mape i kreiranje novih

Kao prvi zadatak kreirat ćemo novu mapu za pohranjivanje datoteka današnje praktične vježbe. Unesite sljedeću naredbu:

```
mkdir tutorial
```

Kao sljedeći korak, pređite u novu mapu:

```
cd tutorial
```

Linux

Ovim ste korakom ušli u svoju novo-kreiranu mapu. Da provjerite gdje se trenutno nalazite, unesite:

```
pwd
```

Ova naredba ispisuje vaš radni direktorij. I trebala bi izgledati ovako:

```
npleic@bioinformatics-kruzok:~$ pwd
/home/npleic/tutorial
```

S naredbom ls izlistavamo sve datoteke iz trenutnog direktorija:

```
npleic@bioinformatics-kruzok:~/tutorial$ ls
```

PLINK

PLINK je razvijen 2007. godine, a predstavlja besplatan alat (open-source) namijenjen provođenju cjelogenomskih asocijacijskih studija (eng. genome-wide association study (GWAS)).

Razvitkom genotipizacijskih čipova, volumen podataka za obradu naglo je porastao. Ovaj veliki obujam podataka predstavljao je izazov za mnoge istraživače. To je utjecalo na razvoj PLINK-a. Glavni cilj Purcella i kolega bio je razviti sveobuhvatan paket koji kombinira upravljanje podacima, kontrolu kvalitete i cjelogenomsku analizu.

Iako postoji više programa za analizu, samo ih nekoliko pruža upravljanje i kontrolu kvalitete (QC) genotipskih podataka, što PLINK čini popularnim izborom.

Plink

Službeni web i dokumentacija:
`https://www.cog-genomics.org/plink/`

Podaci za današnji praktikum:
`https://zzz.bwh.harvard.edu/plink/res.shtml#hapmap`

Description	***File size***	***File name***
Entire HapMap (release 23, 270 individuals, 3.96 million SNPs)	120M	hapmap_r23a.zip
CEU (release 23, 90 individuals, 3.96 million SNPs)	59M	hapmap_CEU_r23a.zip
YRI (release 23, 90 individuals, 3.88 million SNPs)	65M	hapmap_YRI_r23a.zip
JPT+CHB (release 23, 90 individuals, 3.99 million SNPs)	58M	hapmap_JPT_CHB_r23a.zip
CEU founders (release 23, 60 individuals, filtered 2.3 million SNPs)	31M	hapmap_CEU_r23a_filtered.zip
YRI founders (release 23, 60 individuals, filtered 2.6 million SNPs)	38M	hapmap_YRI_r23a_filtered.zip
JPT+CHB founders (release 23, 90 individuals, filtered 2.2 million SNPs)	33M	hapmap_JPT_CHB_r23a_filtered.zip

"The HapMap genotype data (the latest is release 23) are available here as PLINK binary filesets. The SNPs are currently coded according NCBI build 36 coordinates on the forward strand."

Plink

Prvo moramo preuzeti genotipske podatke s Plinkove stranice:

```
wget https://zzz.bwh.harvard.edu/plink/dist/
    hapmap_CEU_r23a.zip
```

Zatim raspakiramo komprimirani file:

```
unzip hapmap_CEU_r23a.zip
```

Naredba ls nam izlistava sve datoteke u trenutnom direktoriju:

```
npleic@bioinformatics-kruzok:~/tutorial$ ls
hapmap_CEU_r23a.bed  hapmap_CEU_r23a.bim
    hapmap_CEU_r23a.fam
```

Plink

Opis PLINK-format datoteka:

.bed datoteka: binarna datoteka koja sadrži informacije o genotipu u komprimiranom formatu.

.bim datoteka: proširena MAP datoteka koja sadrži informacije o SNPovima, uključujući kromosom, ID SNPa, genetsku udaljenost, poziciju baznih parova i imena alela.

.fam datoteka: je ekvivalent ulaznoj .ped datoteci, ali bez informacija o genotipu. Sadrži identifikatore za obitelji i pojedince, zajedno s fenotipom i drugim informacijama.

Plink

Ograničimo analizu samo na autosome (1-22):

```
plink --bfile hapmap_CEU_r23a --chr 1-22 --make-bed --
    out hapmap_CEU_r23a_autosomal
```

```
npleic@bioinformatics-kruzok:~/tutorial$ plink --bfile hapmap_CEU_r23a --chr 1-22 --make-bed --out hapmap_CEU_r23a_autosomal
PLINK v1.90b7 64-bit (16 Jan 2023)            www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hapmap_CEU_r23a_autosomal.log.
Options in effect:
  --bfile hapmap_CEU_r23a
  --chr 1-22
  --make-bed
  --out hapmap_CEU_r23a_autosomal

496702 MB RAM detected; reserving 248351 MB for main workspace.
3849034 out of 3967651 variants loaded from .bim file.
90 people (44 males, 46 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 60 founders and 30 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.983268.
3849034 variants and 90 people pass filters and QC.
Note: No phenotypes present.
--make-bed to hapmap_CEU_r23a_autosomal.bed + hapmap_CEU_r23a_autosomal.bim +
hapmap_CEU_r23a_autosomal.fam ... done.
```

Plink

Sortiramo genotipski file:

```
plink --bfile hapmap_CEU_r23a_autosomal --make-bed --
    out hapmap_CEU_sorted
```

```
npleic@bioinformatics-kruzok:~/tutorial$ plink --bfile hapmap_CEU_r23a_autosomal --make-bed --out hapmap_CEU_sorted
PLINK v1.90b7 64-bit (16 Jan 2023)            www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hapmap_CEU_sorted.log.
Options in effect:
  --bfile hapmap_CEU_r23a_autosomal
  --make-bed
  --out hapmap_CEU_sorted

496702 MB RAM detected; reserving 248351 MB for main workspace.
3849034 variants loaded from .bim file.
90 people (44 males, 46 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 60 founders and 30 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.983268.
3849034 variants and 90 people pass filters and QC.
Note: No phenotypes present.
--make-bed to hapmap_CEU_sorted.bed + hapmap_CEU_sorted.bim +
hapmap_CEU_sorted.fam ... done.
```

Plink

Nakon toga iz zajedničkog /opt/shared/srce direktorija kopiramo datoteku koja sadržava simulirane fenotipove, prvi je kvantitativni (qt-GWAS), a drugi binarni (case-control GWAS) simulirani fenotip:

```
cp /opt/shared/srce/CEU_simulated_phenotypes.list /
   home/npleic/tutorial
```

Plink

Sada spajamo genotipske s fenotipskim podacima:

```
plink --bfile hapmap_CEU_sorted --pheno
    CEU_simulated_phenotypes.list --make-bed --out
    CEU_hapmap_sorted_pheno
```

```
npleic@bioinformatics-kruzok:~/tutorial$ plink --bfile hapmap_CEU_sorted --pheno CEU_simulated_phenotypes.list --make-bed --out CEU_hapmap_sorte
d_pheno
PLINK v1.90b7 64-bit (16 Jan 2023)            www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to CEU_hapmap_sorted_pheno.log.
Options in effect:
  --bfile hapmap_CEU_sorted
  --make-bed
  --out CEU_hapmap_sorted_pheno
  --pheno CEU_simulated_phenotypes.list

496702 MB RAM detected; reserving 248351 MB for main workspace.
3849034 variants loaded from .bim file.
90 people (44 males, 46 females) loaded from .fam.
90 phenotype values present after --pheno.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 60 founders and 30 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.983268.
3849034 variants and 90 people pass filters and QC.
Among remaining phenotypes, 46 are cases and 44 are controls.
--make-bed to CEU_hapmap_sorted_pheno.bed + CEU_hapmap_sorted_pheno.bim +
CEU_hapmap_sorted_pheno.fam ... done.
```

Plink QC

Sada filtriramo podatke na:

- minor allele frequency (MAF)>5% (isključuje sve SNPove kojima je frekvencija minor alela manja od 5%)
- genotyping call rate >95% (isključuje SNPove kojima nedostaje podatak o genotipu za više od 5% ispitanika u uzorku)
- mind 0.1 (isključuje sve ispitanike kojima nedostaje podatak o genotipu za više od 10% SNPova)

```
plink --bfile CEU_hapmap_sorted_pheno --maf 0.05 --
    geno 0.05 --mind 0.1 --make-bed --out
    hapmap_CEU_filtered
```

Plink

```
npleic@bioinformatics-kruzok:~/tutorial$ plink --bfile CEU_hapmap_sorted_pheno --maf 0.05 --geno 0.05 --mind 0.1 --make-bed --out hapmap_CEU_fil
tered
PLINK v1.90b7 64-bit (16 Jan 2023)            www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hapmap_CEU_filtered.log.
Options in effect:
  --bfile CEU_hapmap_sorted_pheno
  --geno 0.05
  --maf 0.05
  --make-bed
  --mind 0.1
  --out hapmap_CEU_filtered

496702 MB RAM detected; reserving 248351 MB for main workspace.
3849034 variants loaded from .bim file.
90 people (44 males, 46 females) loaded from .fam.
90 phenotype values loaded from .fam.
0 people removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 60 founders and 30 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.983268.
407149 variants removed due to missing genotype data (--geno).
1454485 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
1987400 variants and 90 people pass filters and QC.
Among remaining phenotypes, 46 are cases and 44 are controls.
--make-bed to hapmap_CEU_filtered.bed + hapmap_CEU_filtered.bim +
hapmap_CEU_filtered.fam ... done.
```

Plink

Kako bismo mogli kontrolirati na populacijsku stratifikaciju, moramo provesti PCA analizu. Prvo trebamo provesti pročišćavanje (pruning) varijanti:

```
plink --bfile hapmap_CEU_filtered \
      --indep 50 5 2 \
      --out pruned_data
```

Zatim provodimo PCA analizu:

```
plink --bfile hapmap_CEU_filtered \
      --extract pruned_data.prune.in \
      --pca 3 \
      --out pca_result
```

Trebamo dodati naslovni redak (header) na rezultirajuću datoteku:

```
echo -e "FID\tIID\tPC1\tPC2\tPC3" | cat - pca_result.
   eigenvec > pca_result_header.eigenvec
```

Plink

Konačno, provodimo GWAS analizu: U slučaju binarnog fenotipa:

```
plink --bfile hapmap_CEU_filtered \
      --logistic sex hide-covar\
      --pheno CEU_simulated_phenotypes.list \
      --pheno-name Phenotype_binary \
      --allow-no-sex \
      --missing-phenotype -9 \
      --covar pca_result_header.eigenvec \
      --covar-name PC1-PC3 \
      --ci 0.95 \
      --out CEU_binary_results
```

Plink

Za domaći rad kvantitativni fenotip:

```
plink --bfile hapmap_CEU_filtered \
      --linear sex hide-covar\
      --pheno CEU_simulated_phenotypes.list \
      --pheno-name Phenotype_qt \
      --allow-no-sex \
      --missing-phenotype -9 \
      --covar pca_result_header.eigenvec \
      --covar-name PC1-PC3 \
      --ci 0.95 \
      --out CEU_qt_results
```

Vizualizacija rezultata

Sada pokrećemo R ili RStudio:

```
install.packages("data.table")
library(data.table)

##ucitavamo rezultate
path<-"/home/npleic/tutorial/CEU_binary_results.assoc.logistic
results<-as.data.frame(fread(path,  header=T, sep=" ",
stringsAsFactors=FALSE))

str(results)

install.packages("qqman")
library(qqman)
```

Vizualizacija rezultata

```
##Izbacimo p-vrijednosti koje su NA (missing)
results <- results[!is.na(results$P), ]

##sljedeće tri linije pokrenuti zajedno:
png("Manhattan plot of binary phenotype GWAS in CEU.png")
manhattan(results, chr="CHR", bp="BP", snp="SNP", p="P",
main="Manhattan Plot of CEU Data",
col = c("blue","orange","red","purple"), chrlabs = NULL,
suggestiveline = -log10(1e-06), genomewideline = -log10(5e-08)
highlight = NULL, logp = TRUE,ylim = c(0, 8.5))
dev.off()
```

Vizualizacija rezultata

```
png("QQ plot of the binary phenotype GWAS in CEU.png")
qq(results$P, main="QQ Plot of CEU Data",ylim = c(0, 5),
xlim=c(0,5) )
dev.off()

#Sortiramo rezultate da dobijemo SNPove s
#najnižom p-vrijednosti:

results_sorted <- results[order(results$P), ]
top10<-results_sorted[1:10,]
install.packages("openxlsx")
library(openxlsx)

##Zapisujemo 10 najznačajnijih SNPova u excel file:
write.xlsx(top10, file = "Top10_variants.xlsx")
```

Pitanja?

Kontakt: npleic@mefst.hr

LinkedIn: linkedin.com/in/nikolinapleic/