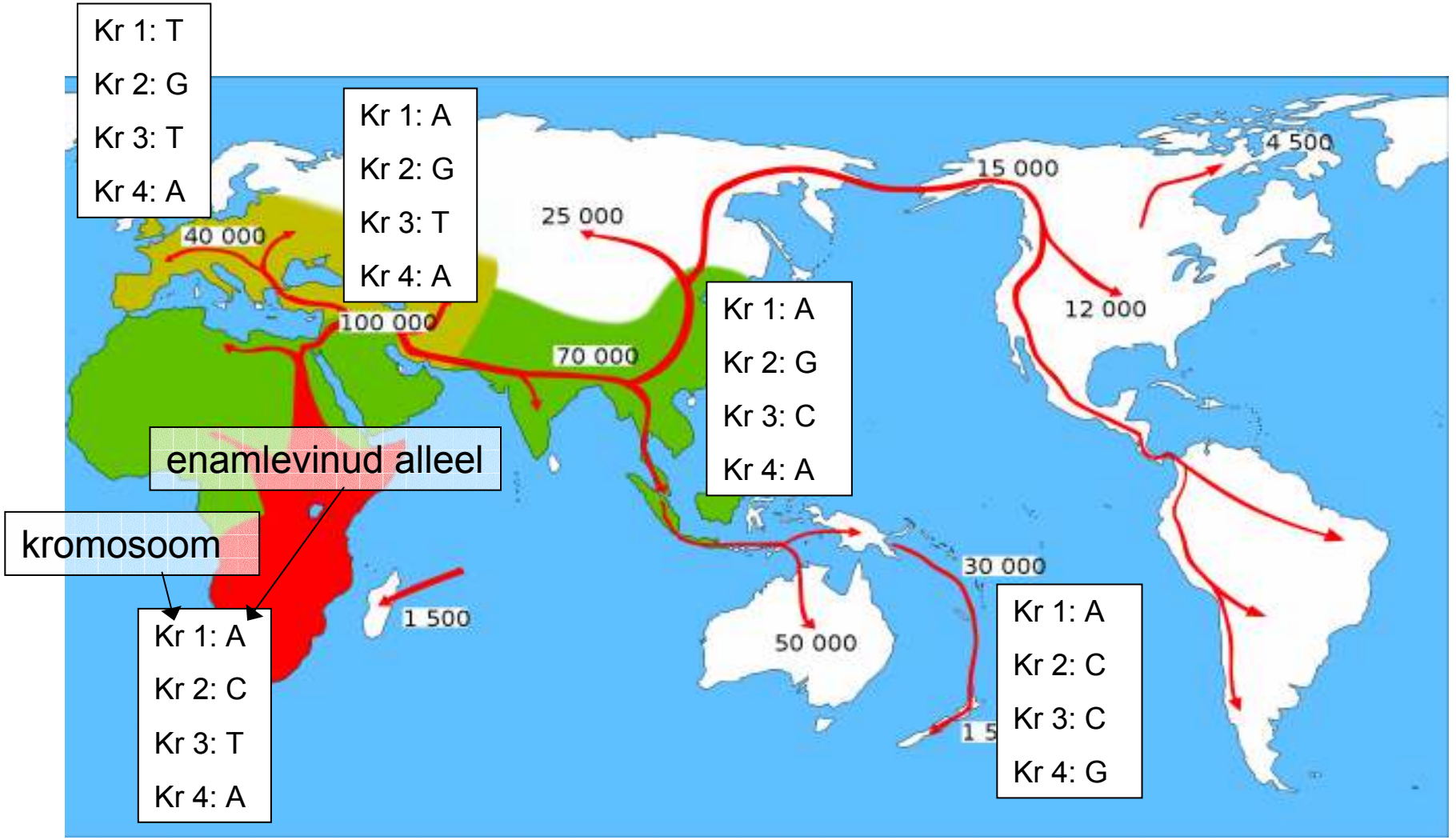
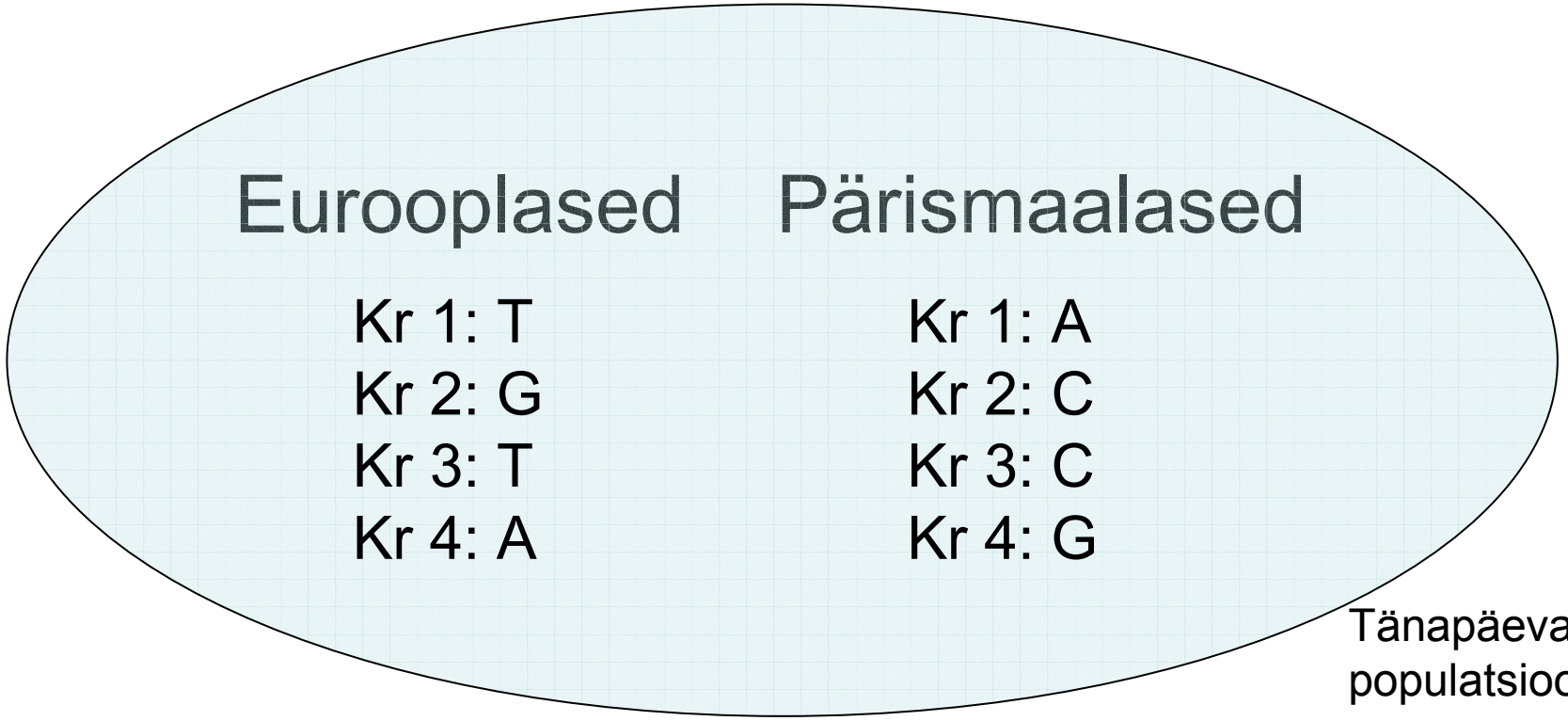
Populatsiooni struktuuri arvessevõtmine ülegenoomsetes uuringutes

Märt Möls

Populatsiooni struktuuri mõjust

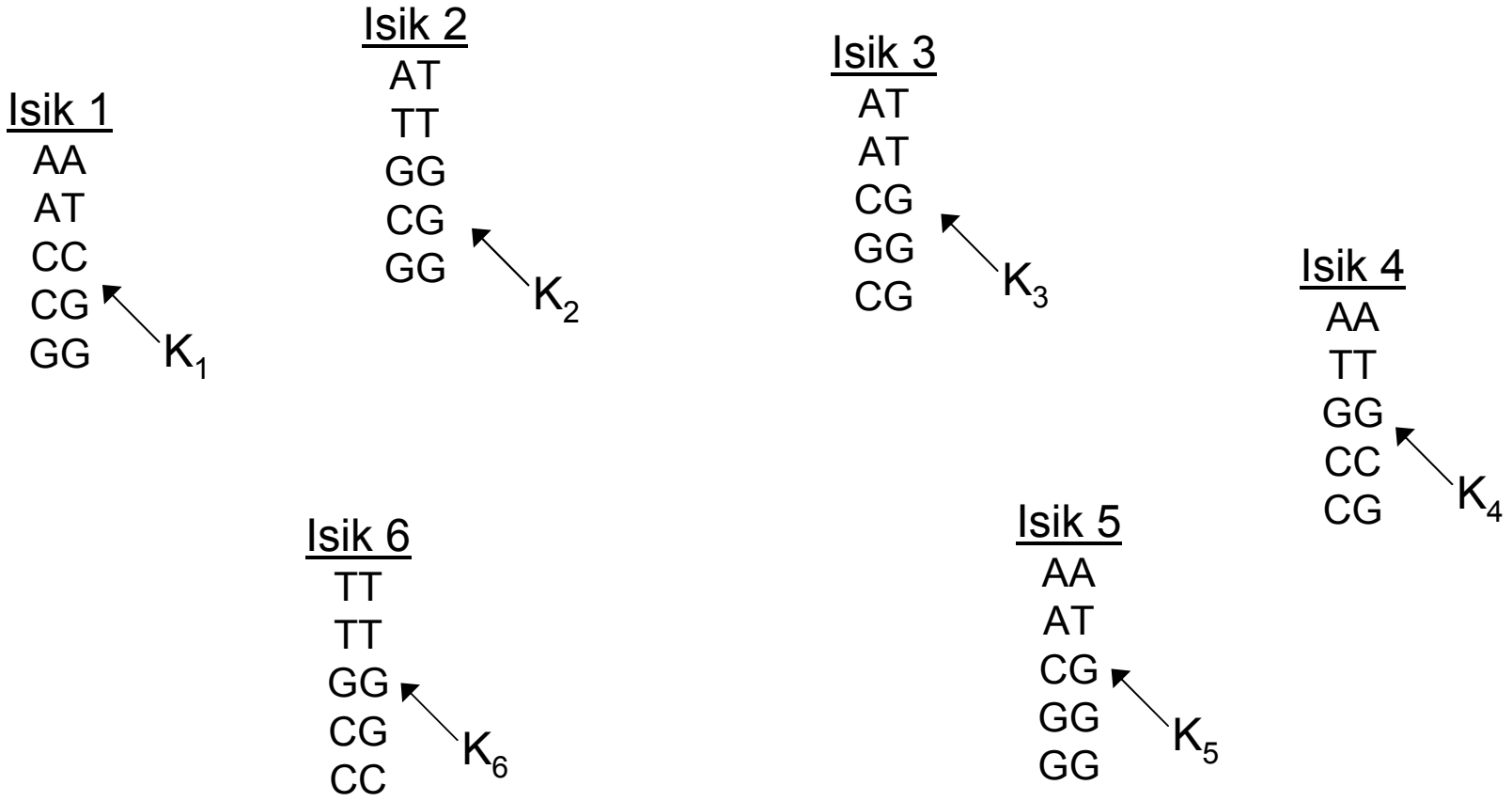


Populatsiooni struktuuri mõjust

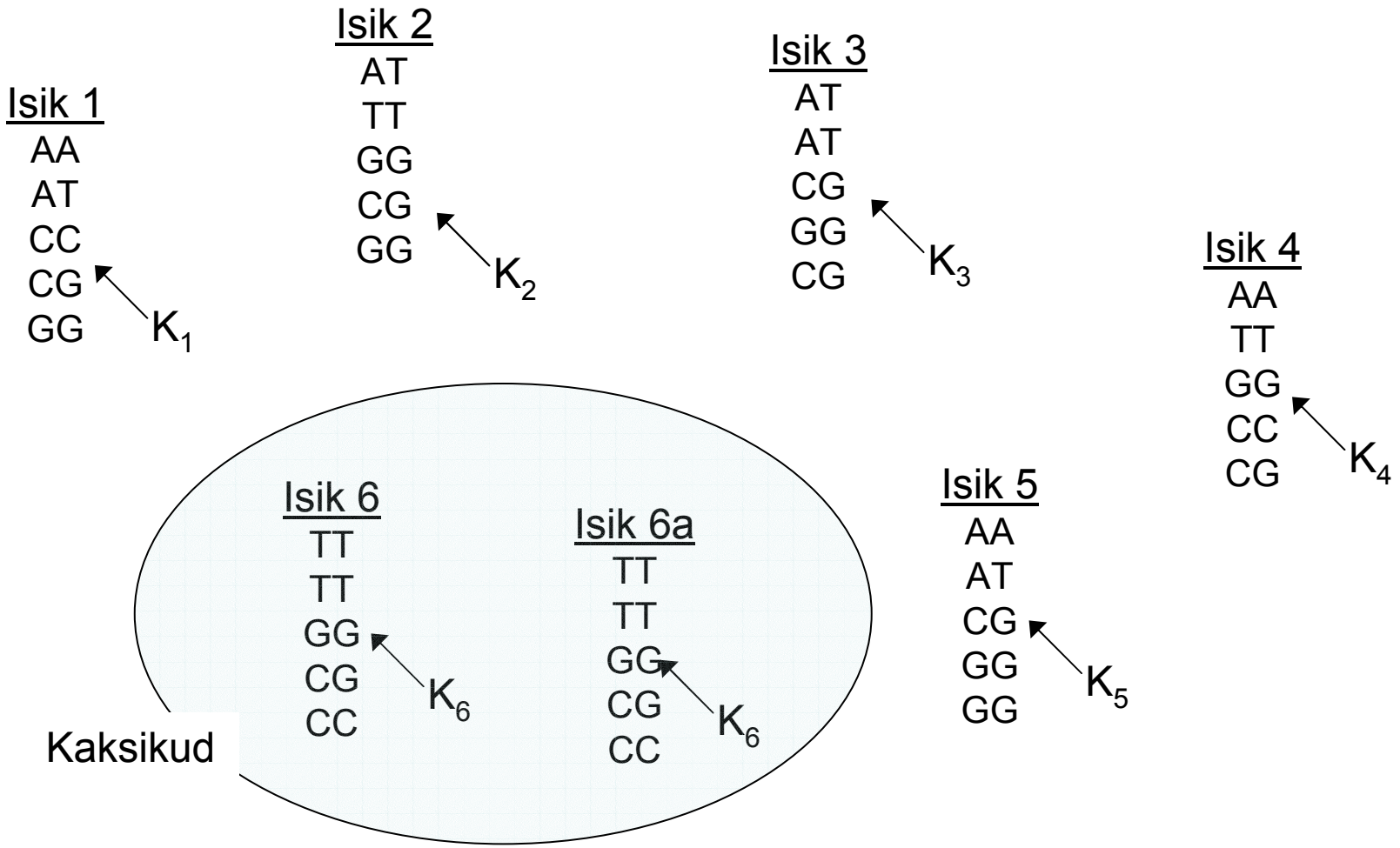


Ükskõik, milline SNP ka ei mõjutaks fenotüüpi, “signaali” annavad kõik ülaltoodud, erinevates kromosoomides paiknevad SNP’d

Populatsiooni struktuuri mõjust II



Populatsiooni struktuuri mõjust II



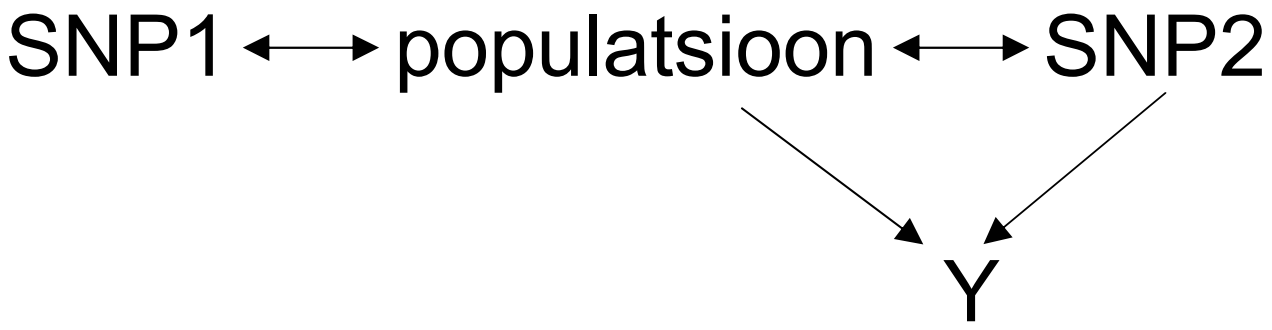
Lihtsustatud näide

	Alpi mägiküla	Eestlased
Lookus 1	100% AA	100% GG
Lookus 2	100% CC	100% AA
Lookus 3	100% GG	100% CC
	elavad kõrgel	elavad madalal
	elavad kaua	ei ela nii kaua
	....	...

Lihtsustatud näide

	sisserännanud	pärismaalased
Lookus 1	100% AA	100% GG
Lookus 2	100% CC	100% AA
Lookus 3	100% GG	100% CC
	elavad madalal	elavad madalal
	ei söö X'i	ei söö X'i
	elavad kaua	ei ela nii kaua
	....	...

Genereeritud andmetega näide



Kas avastame “süüdlase”?

Y=f(populatsioon)+1.5*SNP2+rnorm(n)

```
> m_SNP1=lm(Y~SNP1)
> summary(m_SNP1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.86530    0.03831 100.907  < 2e-16 ***
SNP1        -0.22180    0.02960  -7.493 9.04e-14 ***
```

Asjasse mittepuutuv SNP1 osutub stat. oluliseks...

Kas avastame “süüdlase”?

Y=f(populatsioon)+1.5*SNP2+rnorm(n)

```
> m_SNP2=lm(Y~SNP2)
> summary(m_SNP2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.74376    0.03974  94.211  < 2e-16 ***
SNP2        -0.08886    0.02959  -3.003  0.00270 **
```

Asjasse puutuv SNP2 osutub ka statistiliselt oluliseks, aga tema p-väärtus on märksa suurem kui SNP1-l, ja tema mõju on tagurpidiseks hinnatud...

Lihtne lahendus?

```
> m_SNP1=lm(Y~SNP1+factor(populatsioon))
> drop1(m_SNP1, test="F")
Model:
Y ~ SNP1 + factor(populatsioon)
                     Df Sum of Sq    RSS    AIC F value  Pr(F)
<none>                             4168.6 1192.7
SNP1                  1      0.35 4168.9 1190.9  0.2236 0.6363
factor(populatsioon)  8    719.90 4888.5 1606.8 58.0694 <2e-16 ***

> m_SNP2=lm(Y~SNP2+factor(populatsioon))
> drop1(m_SNP2, test="F")
Model:
Y ~ SNP2 + factor(populatsioon)
                     Df Sum of Sq    RSS     AIC F value     Pr(F)
<none>                             2693.0   12.99
SNP2                  1    1475.9 4168.9 1190.89 1474.28 < 2.2e-16 ***
factor(populatsioon)  8    2280.6 4973.6 1653.40  284.76 < 2.2e-16 ***

> summary(m_SNP2)
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
SNP2                   1.59594    0.04157  38.396  < 2e-16 ***
            [.................................]
```

Uute vaatluste prognoosimine?

Prognoosivead on suuremad, kui kasutame õiget SNP'i:

```
> Y_uus=f(populatsioon_uus)+1.5*SNP2_uus+rnorm(n)
> m_SNP1=lm(Y~SNP1)
> m_SNP2=lm(Y~SNP2)

> yprog_SNP1=predict(m_SNP1,
                data.frame(SNP1=SNP1_uus))
> yprog_SNP2=predict(m_SNP2,
                data.frame(SNP2=SNP2_uus))

> mean((Y_uus-yprog_SNP1)**2)
[1] 2.369676 <- keskmine prognoosiviga vale SNPga
> mean((Y_uus-yprog_SNP2)**2)
[1] 2.535556 <- keskmine prognoosiviga õige SNPga
```

Keskmine prognoosiviga kasutades “õiget” väärtust:

```
> Y_uus_prog=k+1.5*SNP2_uus

> mean(Y_uus_prog)
[1] 4.113815
> mean(Y_uus)
[1] 4.113815

> mean((Y_uus-Y_uus_prog)**2)
[1] 3.775324 <- kõige ebatäpsem prognoos!
```

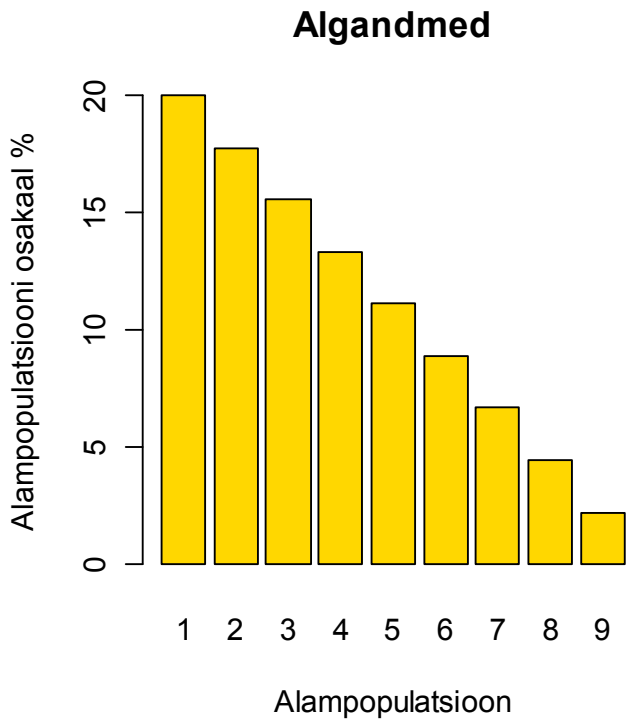
Täpne väärtus viib täpsema tulemuseni vaid siis, kui teame ka alampopulatsiooni-spetsiifilisi kordajaid:

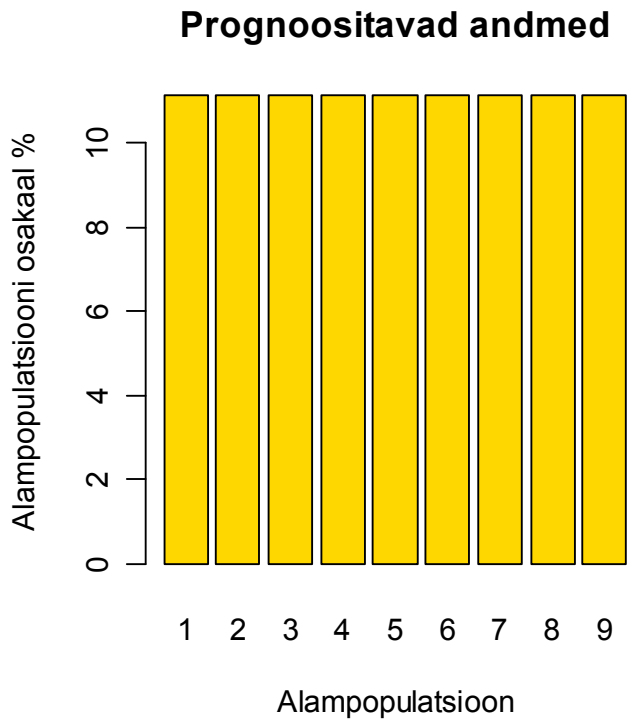
```
> m_SNP1=lm(Y~SNP1+factor(populatsioon))
> m_SNP2=lm(Y~SNP2+factor(populatsioon))

> yprog_SNP1_ja_pop = predict(m_SNP1,
      data.frame(SNP1=SNP1_uus,
      populatsioon=populatsioon_uus))
> yprog_SNP2_ja_pop = predict(m_SNP2,
      data.frame(SNP2=SNP2_uus,
      populatsioon=populatsioon_uus))

> mean((Y_uus-yprog_SNP1_ja_pop)**2)
[1] 1.715935
> mean((Y_uus-yprog_SNP2_ja_pop)**2)
[1] 1.014909 <-     SNP2 kasutav mudel lõpuks parem
```

Tähelepanekuid

- Uuritavat tunnust võib paremini prognoosida asjasse mittepuutuv SNP
- Tegeliku (näiteks molekulaarse) toimemehaanismi täpne teadmine aitab jõuda täpsema prognoosini üksnes siis, kui oskame sisse viia populatsioonispetsiifilisi korrektsioone

Kusjuures alampopulatsioonide osakaal võib vahepeal olla märkimisväärselt muutunud...





Kas siis?

- Mujal maailmas avaldatud populatsiooni järgi korrigeeritud tulemused on kasutud prognoosimiseks siin?
- Siin leitud populatsiooni-struktuuri järgi korrigeeritud tulemused ei kõlba kasutamiseks mujal maailmas prognooside tegemiseks?
- Siin leitud “põhjuslikud” SNP’d võivad mõjutada uuritavat tunnust ka mujal maailmas ...

  ... kuid võivad osutuda kasutuks uuritava tunnuse väärtuste (jääb haigeks / kui pikaks inimene kasvab / kui targaks saab) prognoosimisel?

Kuidas arvestada populatsiooni struktuuriga, probleemiks ainult (halvasti segunenud) alampopulatsioonid

- Identifitseeri alampopulatsioonid ja kasuta alumpopulatsiooni identifikaatorit mudelis;
- Klasterda geeniinfo järgi inimesed alampopulatsioonideks;
- Kasuta teatavat arvu abitunnuseid, mis kirjeldaksid enam-vähem ära geneetilise “tausta” (peakomponendid)
- Geneetilist kaugust saab kasutada muukski kui lihtsalt klasterdamiseks?

Variant A. Lihtsalt tea alampopulatsioone.

```
> m_SNP1=lm(Y~SNP1+factor(populatsioon))
> drop1(m_SNP1, test="F")
                    Df Sum of Sq    RSS    AIC  F value  Pr(F)
SNP1                 1      0.42 4735.8 1535.1   0.2383 0.6255
factor(populatsioon)  8   1707.77 6443.2 2352.4 121.2656 <2e-16 ***

> m_SNP2=lm(Y~SNP2+factor(populatsioon))
> drop1(m_SNP2, test="F")
                    Df Sum of Sq    RSS     AIC F value     Pr(F)
SNP2                 1    2008.1 4735.8 1535.13 1980.30 < 2.2e-16 ***
factor(populatsioon)  8    4160.9 6888.6 2532.86  512.92 < 2.2e-16 ***
```

Töötab muljetavaldavalt hästi, kui populatsioonid täpselt ja asjakohaselt on eraldatavad.

Peakomponendid

- Leia k<<n abitunnust, peakomponenti, mida teades oleks suhteliselt hästi võimalik rekonstrueerida esialgsed n-tunnust.

Genereerimisskeem

Tekitan 9 “varjatud” alampopulatsiooni;

Genereerin 1000 SNP’i, igaühel neist sõltub alleelisagedus populatsioonist;

Leian 10 peakomponenti, mis peaksid võimalikult hästi kirjeldama kõiki neid 1000-t SNP’i;

Kohandan analüüsi võttes arvesse 9 peakomponenti.

Kuidas töötab?

- **SNP1** (tegelik SNP ei asu samas kromosoomis)

  p-väärtus < 0.05 : 97% valimitest

  keskmine p-väärtus: 0,01

- **SNP2** (tegelik)

  p-väärtus < 0.05 : 65% valimitest

  keskmine p-väärtus: 0,12

Klasterdame endale populatsioonid

Genetic Epidemiology 32: 215–226 (2008)

Improved Correction for Population Stratification in Genome-wide Association Studies by Identifying Hidden Population Structures

Qizhai Li[1,2] and Kai Yu[1*]

[1]Division of Cancer Epidemiology and Genetics, NCI, NIH, Bethesda, Maryland
[2]Academy of Mathematics and Systems Science, CAS, Beijing, China

- Step 1: Based on a chosen similarity metric, calculate the genetic correlation between two subjects and form a similarity matrix.
- Step 2: Based on the similarity matrix, use the MDS method [Mardia et al., 2003] to represent each subject by a vector of L coordinates (called principal coordinates).
- Step 3: Group points represented by vectors of principal coordinates into a variable number of clusters using the k-medoids clustering method. For each subject, determine its membership in each of the k clusters.
- Step 4: To test a marker's association with the disease, use a logistic regression model, treating the L principal coordinates and the membership in k clustered groups as covariates.

Hindame geneetilise kauguse – suguluskoefitsient (kinship) – kahe suvalise inimese vahel kasutades informatsiooni markerite kohta:

$$\hat{f}_{ij} = \frac{1}{L} \sum_{l=1}^{L} \frac{(g_{l,i} - p_l)(g_{l,j} - p_l)}{p_l(1 - p_l)}$$

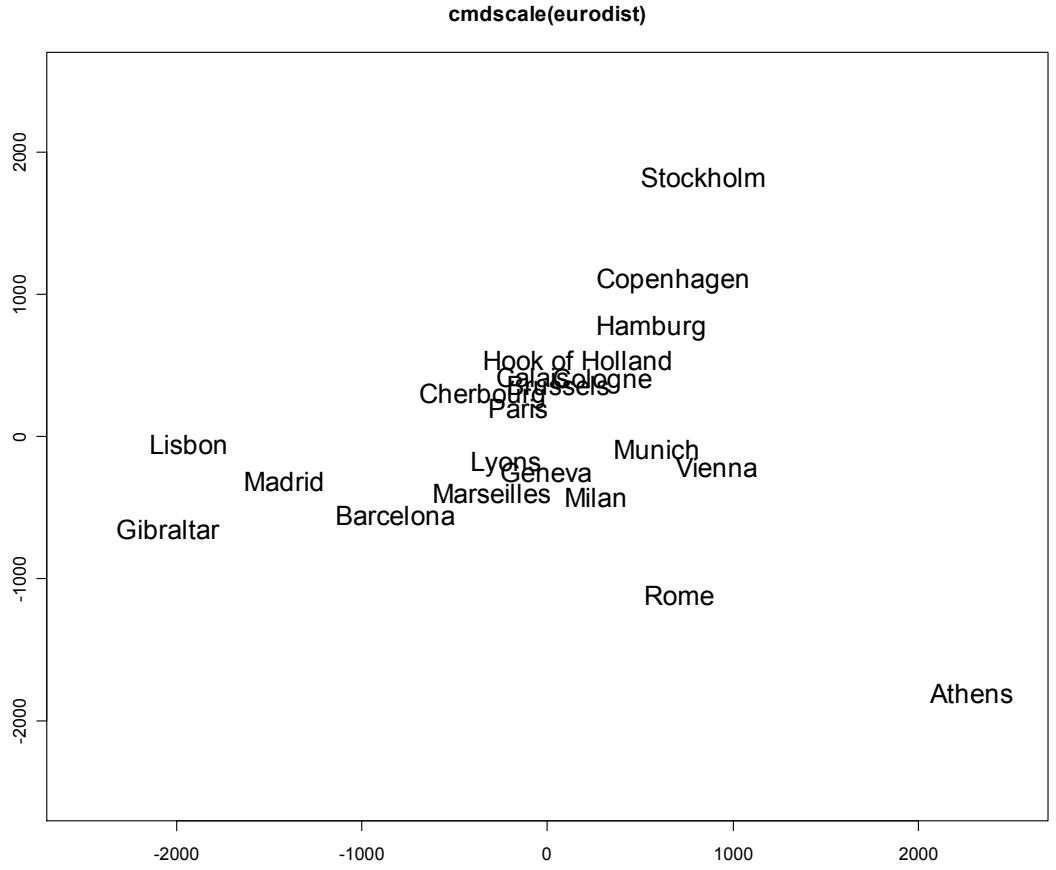
Järgmine samm – kasutades kauguseid, leia peotäis tunnuseid (koordinaate), nii et inimestevaheline kaugus kajastuks võimalikult täpselt leitud koordinaatide vahelise kaugusena (kasutades Multidimensional Scalingut, MDS)

MDS

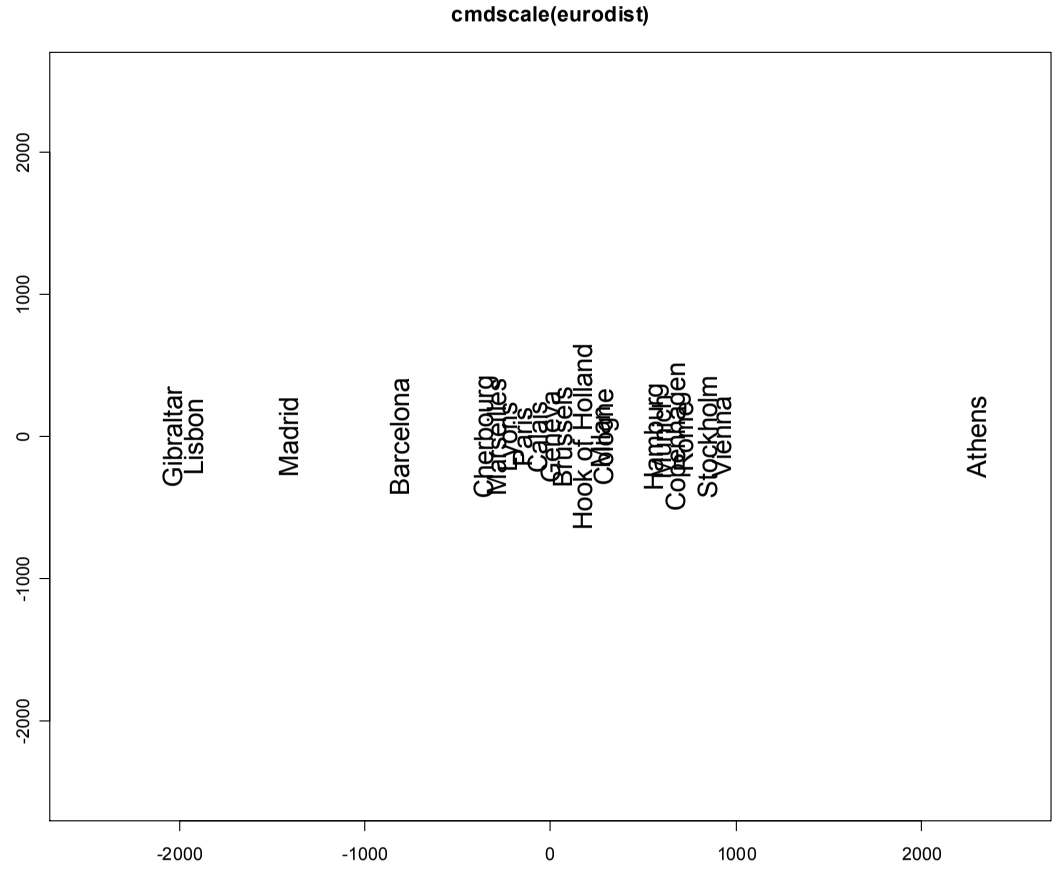
On antud kaugused:

	Athens	Barcelona	Brussels	Calais	Cherbourg…
Barcelona	3313				
Brussels	2963	1318			
Calais	3175	1326	204		
Cherbourg	3339	1294	583	460	
Cologne	2762	1498	206	409	785
……	……		……		……

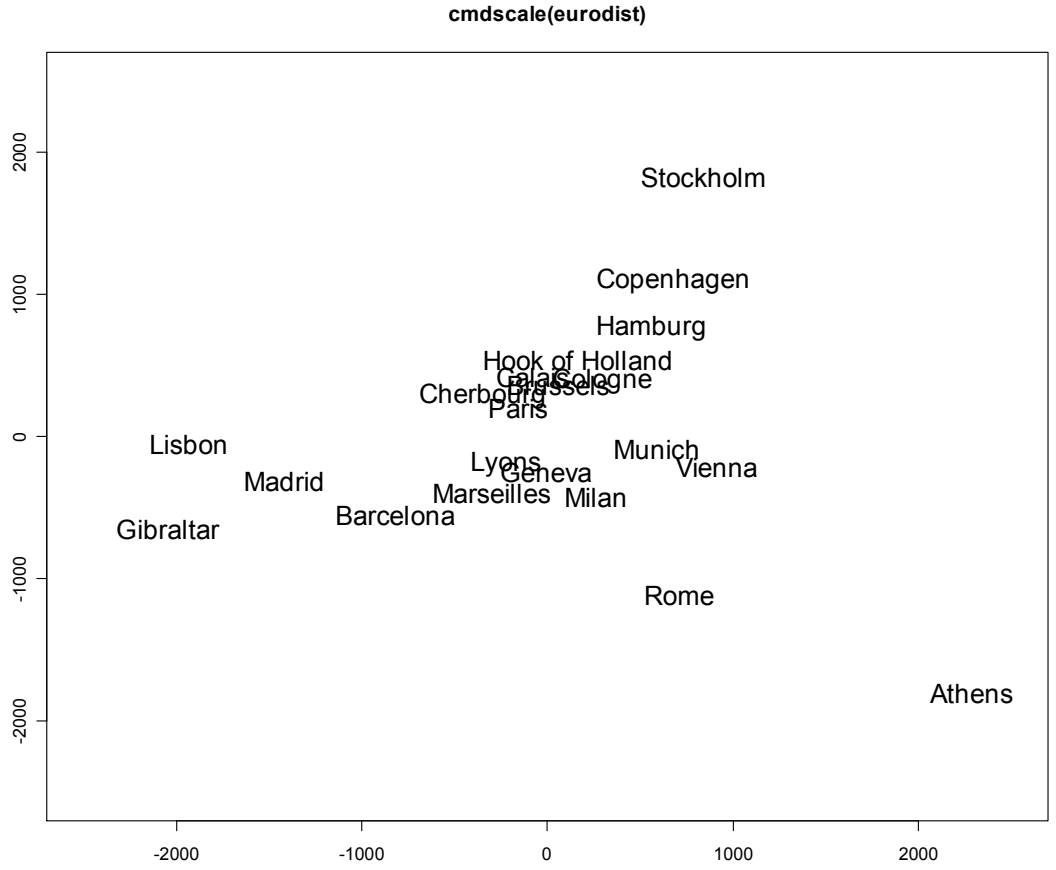
MDS – rekonstrueerib koordinaadid



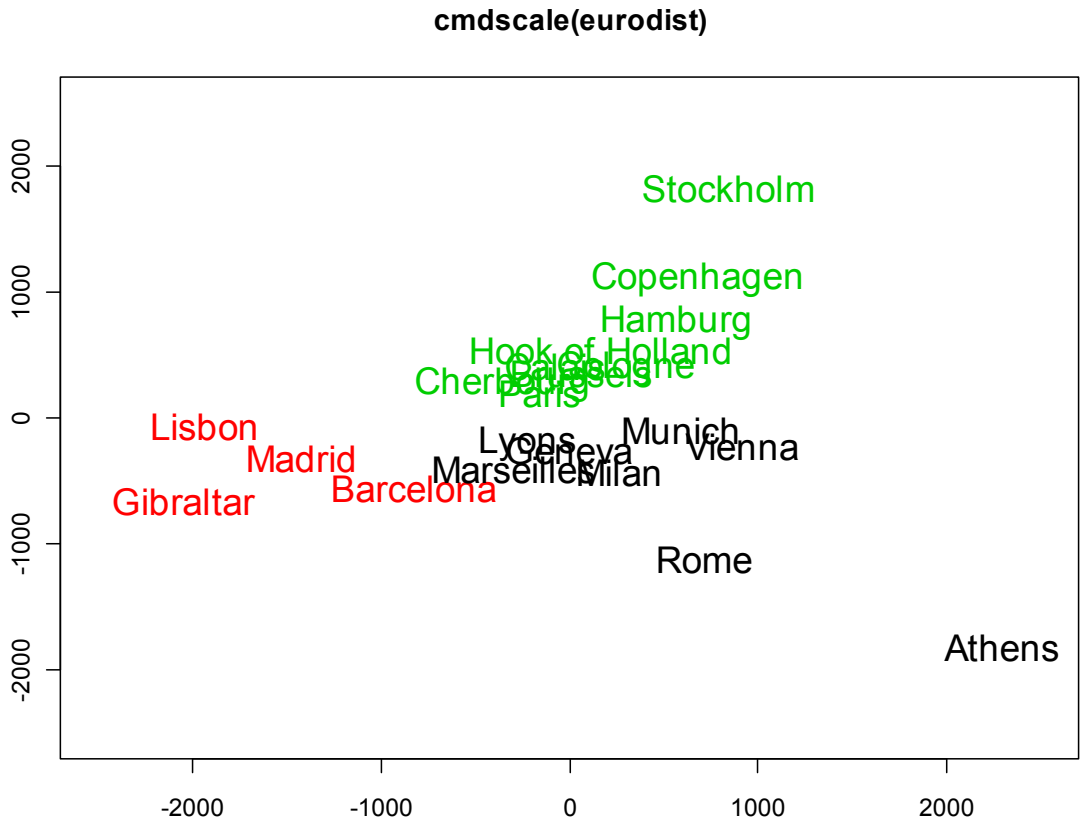
MDS – rekonstrueerib koordinaadid



MDS – rekonstrueerib koordinaadid



Tekitatakse klastrid



Kuidas töötab MDS-klasterdus?

- **SNP1** (tegelik SNP ei asu samas kromosoomis)

  p-väärtus < 0.05 : 70% valimitest

  keskmine p-väärtus: 0,11

- **SNP2** (tegelik)

  p-väärtus < 0.05 : 100% valimitest

  keskmine p-väärtus: <0,001

Kuidas töötab tavaline klasterdus?

- **SNP1** (tegelik SNP ei asu samas kromosoomis)

  p-väärtus < 0.05 : 77% valimitest

  keskmine p-väärtus: 0,08

- **SNP2** (tegelik)

  p-väärtus < 0.05 : 94% valimitest

  keskmine p-väärtus: 0,02

Hinnangute korrigeerimine

Testime lineaarse mudeli parameetrit näiteks hii-ruut testi abil:

$$T^2 = \frac{\hat{\beta}^2}{\hat{D}\left(\hat{\beta}\right)}$$

Kui kõik inimesed pärineksid samast segunevast populatsioonist, ja H_0 kehtiks, siis:

$$T^2 \overset{H_0}{\sim} \chi^2$$

Hinnangute korrigeerimine

Kui populatsioonil on struktuur (alampopulatsioonid, inbriiding), siis

$$T^2 \overset{H_0}{\sim} \lambda \cdot \chi^2$$

Yurii Aulchenko, GenAbeli tutorial

Lisaparameeter lambda on hinnatav ja hinnanguid on võimalik vastavalt korrigeerida. Meetod võib leevendada ka muid hädasid?

Kuidas töötab hr. Tambovi koefitsient?

- **SNP1** (tegelik SNP ei asu samas kromosoomis)

  p-väärtus < 0.05 : 57% valimitest

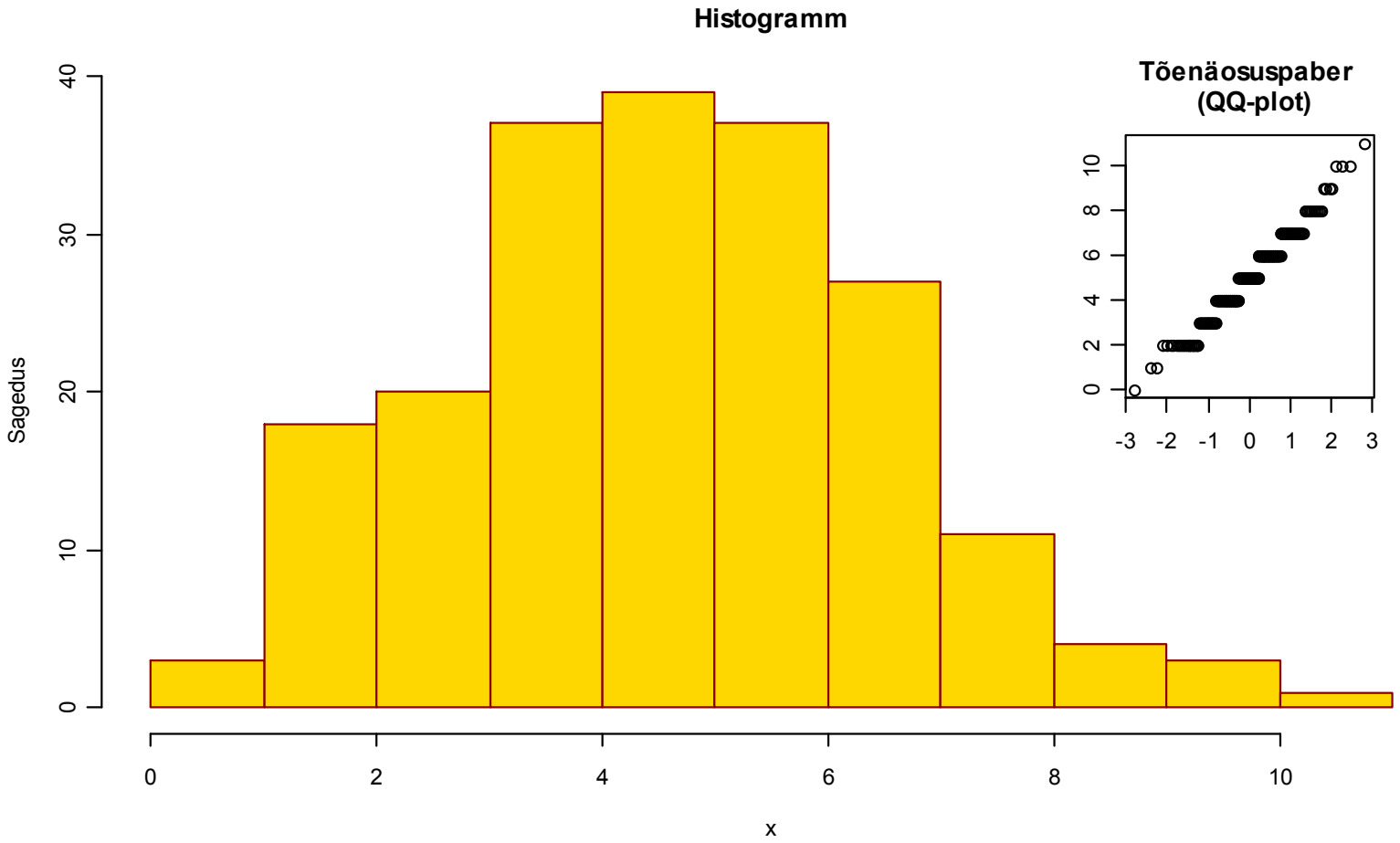
  keskmine p-väärtus: 0,10

- **SNP2** (tegelik)

  p-väärtus < 0.05 : 12% valimitest

  keskmine p-väärtus: 0,57

Peaaegu normaaljaotus?



Mittetöötava rusikareegli näide: t-test, “suur valim” (*n*=100)

kasutatud olulisuse nivoo	vale testitulemuse (I liiki vea) tõenäosus
0,05	0,0504
0,0005	0,00053
0,000005	0,0000073
0,000005 (ühepoolne)	0,000011

$$0{,}000011 * 10000 \neq 0{,}05$$

Bonferroni meetod võimendab üles ka kõige pisema eksimuse eeldustes suureks probleemiks.

Segamudelite lähenemine

$$Y_i = X_i\beta + G_i + e_i$$

$$Cov(G_i, G_j) = 2f_{ij}\sigma_G^2$$