

Задание 1.

Загрузите данные:

[https://drive.google.com/file/d/1WdPY4uGesOXrhCSCHVfma85zNS-RobiZ/view?usp=share\\_link](https://drive.google.com/file/d/1WdPY4uGesOXrhCSCHVfma85zNS-RobiZ/view?usp=share_link)

Определите missingness каждого обследуемого при помощи plink:

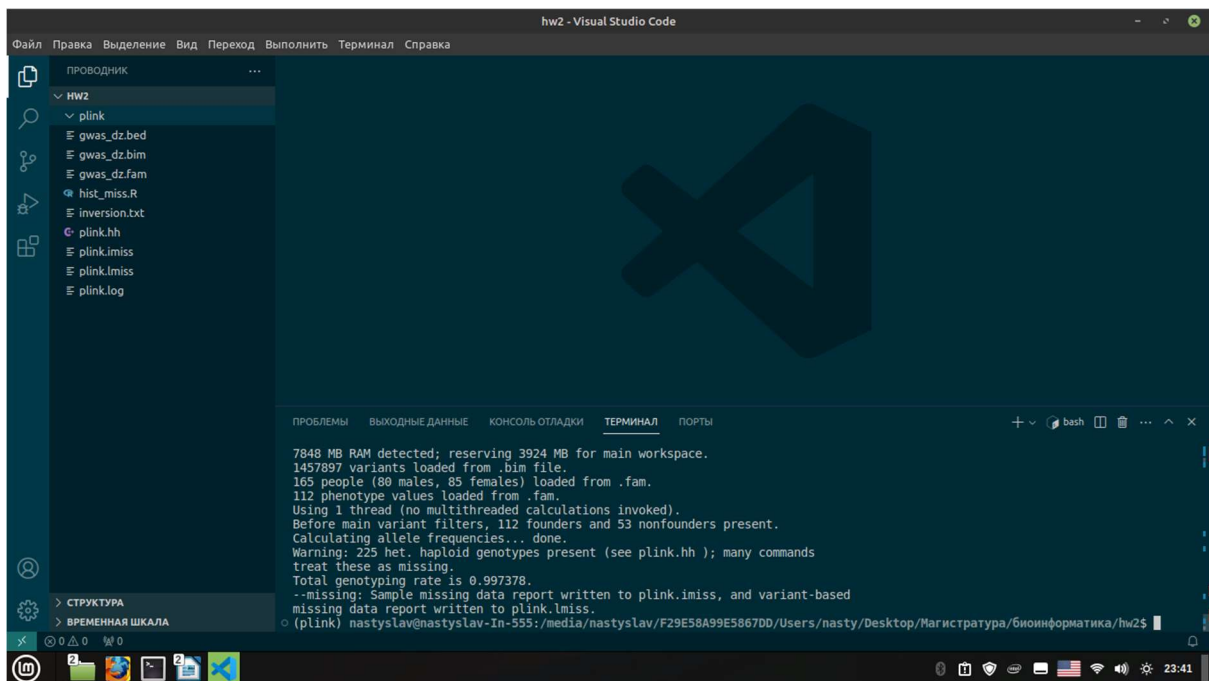
```
plink --bfile HapMap_3_r3_1 --missing
```

При помощи скрипта hist\_miss.R, который находится в одной директории с заданием, сгенерируйте графики missingness. Выберите пороговое значение данного параметра.

Почему получилось два графика? Изучите их. Если понимание вызывает у Вас трудности, просмотрите содержимое скрипта.

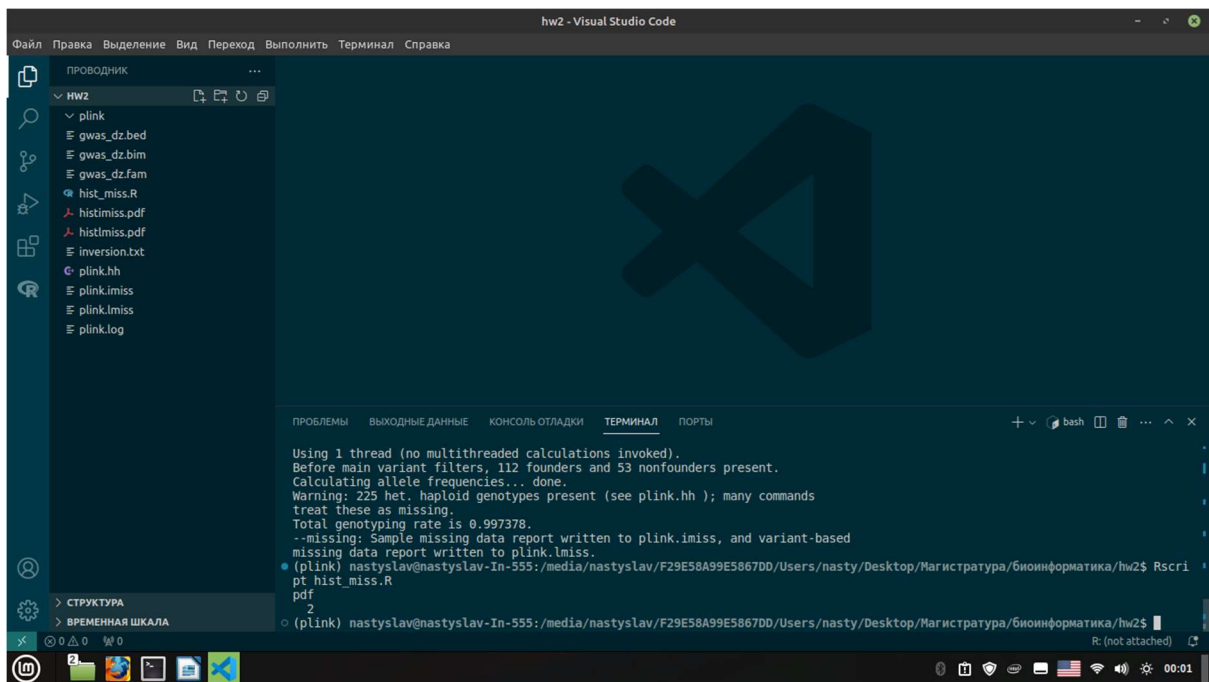
Какое значение Вы выбрали? Почему?

Проведите QC по параметру missingness: устранили SNPs и индивидов, которые превышают избранное пороговое значение (используйте параметры geno и mind в plink).



```
hw2 - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка

PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
7848 MB RAM detected; reserving 3924 MB for main workspace.
1457897 variants loaded from .bim file.
165 people (80 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 53 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see plink.hh ); many commands
treat these as missing.
Total genotyping rate is 0.997378.
--missing: Sample missing data report written to plink.lmiss, and variant-based
missing data report written to plink.lmiss.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

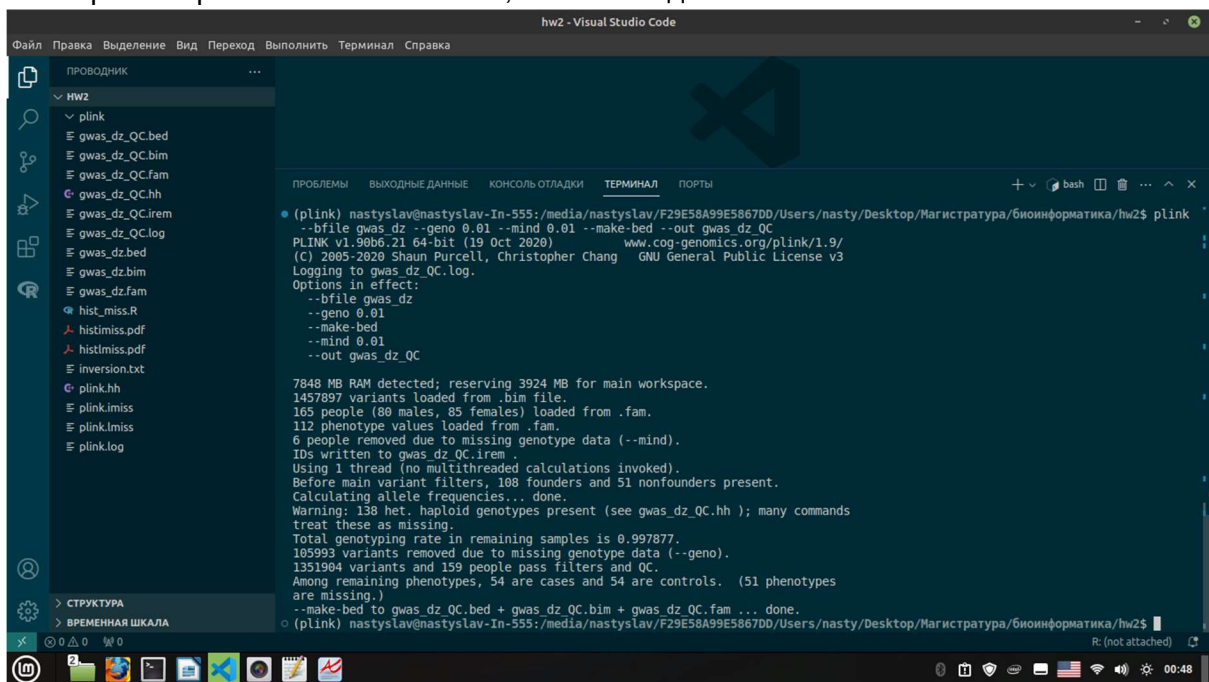


The screenshot shows the Visual Studio Code interface with a file explorer on the left containing files like `plink`, `gwas_dz.bed`, `gwas_dz.bim`, `gwas_dz.fam`, `hist_miss.R`, `histmiss.pdf`, `inversion.txt`, `plink.hh`, `plink.imiss`, `plink.lmiss`, and `plink.log`. The terminal window at the bottom displays the following output:

```
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 112 founders and 53 nonfounders present.
Calculating allele frequencies... done.
Warning: 225 het. haploid genotypes present (see plink.hh ); many commands
treat these as missing.
Total genotyping rate is 0.997378.
--missing: Sample missing data report written to plink.imiss, and variant-based
missing data report written to plink.lmiss.
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ Rscript hist_miss.R
pdf
2
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

Получилось 2 графика, потому что скрипт считал данные о пропусках из двух файлов (для индивидов и для SNP).

Я выбрала пороговое значение 0.01, т. к. наблюдается заметное изменение частоты.



The screenshot shows the Visual Studio Code interface with a file explorer on the left containing files like `plink`, `gwas_dz_QC.bed`, `gwas_dz_QC.bim`, `gwas_dz_QC.fam`, `gwas_dz_QC.hh`, `gwas_dz_QC.irem`, `gwas_dz_QC.log`, `gwas_dz.bed`, `gwas_dz.bim`, `gwas_dz.fam`, `hist_miss.R`, `histmiss.pdf`, `inversion.txt`, `plink.hh`, `plink.imiss`, `plink.lmiss`, and `plink.log`. The terminal window at the bottom displays the following output:

```
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink --bfile gwas_dz --geno 0.01 --mind 0.01 --make-bed --out gwas_dz_QC
PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to gwas_dz_QC.log.
Options in effect:
--bfile gwas_dz
--geno 0.01
--make-bed
--mind 0.01
--out gwas_dz_QC

7848 MB RAM detected; reserving 3924 MB for main workspace.
1457897 variants loaded from .bim file.
165 people (80 males, 85 females) loaded from .fam.
112 phenotype values loaded from .fam.
6 people removed due to missing genotype data (--mind).
IDs written to gwas_dz_QC.irem.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 138 het. haploid genotypes present (see gwas_dz_QC.hh ); many commands
treat these as missing.
Total genotyping rate in remaining samples is 0.997877.
105993 variants removed due to missing genotype data (--geno).
1351904 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--make-bed to gwas_dz_QC.bed + gwas_dz_QC.bim + gwas_dz_QC.fam ... done.
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

Задание 2.

Проверьте пол индивидов: женщины определяются как индивиды с F-значением гомозиготности совокупности SNP, локализованных на X-хромосоме, менее 0.2, мужчины — более 0.8.

Выберите одну из двух стратегий:

1) Устраняем проблемных индивидов.

Используйте команду:

`plink --bfile [Ваш последний файл с фильтрацией по missingness] --check-sex`

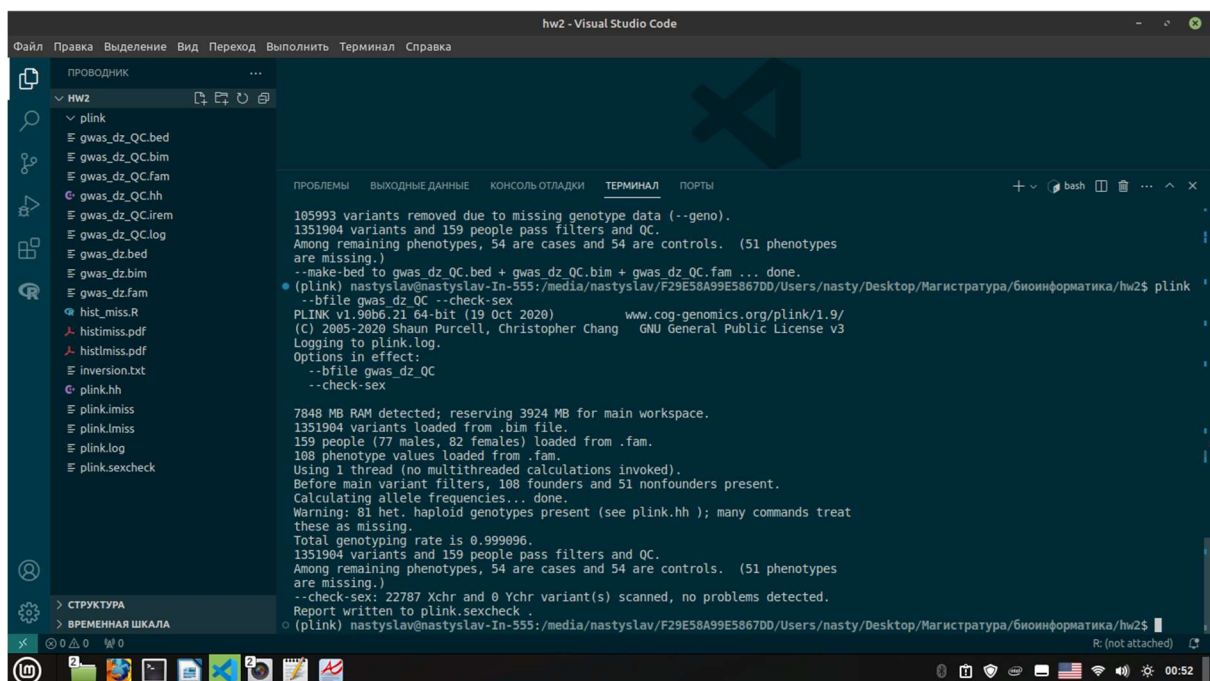
Посмотрите файл с расширением `".sexcheck"`, сгенерированный последней командой `plink`. В нём хранятся указанные F-значения.

Напишите скрипт, позволяющий отобразить гистограмму распределения F-значений для: (1) всех индивидов (ожидаем увидеть, что основная их часть сгруппирована в областях значений от 0 до 0.2 и от 0.8 до 1.0); (2) мужчин; (3) женщин.

В файле с расширением `".sexcheck"` индивиды с неправильным назначением пола имеют статус `"PROBLEM"`. Удалите индивидов со статусом `"PROBLEM"`, поместив их идентификаторы (FID и IID) в отдельный файл и применив на нём параметр `--remove` в `plink`.

2) Присваиваем пол согласно F-значению.

Используем команду `--impute-sex` в `plink`. Эта команда автоматически рассчитывает F-значения и присваивает пол согласно им.



```
hw2 - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка

PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
+ v  bash  [icon]  [icon]  [icon]  [icon]  [icon]

105993 variants removed due to missing genotype data (--geno).
1351904 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--make-bed to gwas_dz_QC.bed + gwas_dz_QC.bim + gwas_dz_QC.fam ... done.
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink
  -bfile gwas_dz_QC --check-sex
  PLINK v1.90b6.21 64-bit (19 Oct 2020)          www.cog-genomics.org/plink/1.9/
  (C) 2005-2020 Shaun Purcell, Christopher Chang  GNU General Public License v3
  Logging to plink.log.
  Options in effect:
    --bfile gwas_dz_QC
    --check-sex

7848 MB RAM detected; reserving 3924 MB for main workspace.
1351904 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 81 het. haploid genotypes present (see plink.hh ); many commands treat
these as missing.
Total genotyping rate is 0.999996.
1351904 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--check-sex: 22787 Xchr and 0 Ychr variant(s) scanned, no problems detected.
Report written to plink.sexcheck .
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

```
plink.sexcheck - hw2 (рабочая область) - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка

PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
+  bash  ...  ^  x

Error: Failed to open plink.sexcheck.bed.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink
--bfile gwas_dz_QC --impute-sex --make-bed --out plink_with_sex
PLINK v1.90b6.21 64-bit (19 Oct 2020)      www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink_with_sex.log.
Options in effect:
--bfile gwas_dz_QC
--impute-sex
--make-bed
--out plink_with_sex

7848 MB RAM detected; reserving 3924 MB for main workspace.
1351904 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 81 het. haploid genotypes present (see plink_with_sex.hh ); many
commands treat these as missing.
Total genotyping rate is 0.999096.
1351904 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--impute-sex: 22787 Xchr and 0 Ychr variant(s) scanned, all sexes imputed.
Report written to plink_with_sex.sexcheck .
--make-bed to plink_with_sex.bed + plink_with_sex.bim + plink_with_sex.fam ...
done.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

	FID	IID	PEDSEX	SNPSEX	STATUS	F
1	1328	NA06989	2	2	OK	-0.01672
2	1349	NA11843	1	1	OK	1
4	1338	NA12341	2	2	OK	-0.01169

### Задание 3.

Отберите только аутосомные SNP: отберите из файла ".bim" те SNP, которые находятся на хромосомах 1-22, и отберите только их при помощи параметра `--extract`.

Выведите частоту встречаемости SNP в популяции в отдельный файл при помощи параметра `freq` в `plink` и отобразите гистограмму распределения популяционных частот SNP.

Выберите пороговое значение для минимального допустимого MAF (т.е. популяционной частоты) и примените его при помощи параметра `--maf` в `plink`. Обычно выбирают значение от 0.01 до 0.05.

Также отберите при помощи параметра `--hwe` только те SNPs, р-значение которых по закону Харди-Вайнберга не превышает  $1e-6$ .





```
autosomal_snps.bim - hw2 (рабочая область) - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка
PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
(пlink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink
--bfile autosomal_snps --maf 0.01 --make-bed --out autosomal_snps_maf_filtered
PLINK v1.90b6.21 64-bit (19 Oct 2020)      www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang  GNU General Public License v3
Logging to autosomal_snps_maf_filtered.log.
Options in effect:
--bfile autosomal_snps
--maf 0.01
--make-bed
--out autosomal_snps_maf_filtered

7848 MB RAM detected; reserving 3924 MB for main workspace.
1351904 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 81 het. haploid genotypes present (see autosomal_snps_maf_filtered.hh
); many commands treat these as missing.
Total genotyping rate is 0.999096.
214502 variants removed due to minor allele threshold(s)
(-maf/-max-maf/-mac/-max-mac).
1137402 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--make-bed to autosomal_snps_maf_filtered.bed + autosomal_snps_maf_filtered.bim
+ autosomal_snps_maf_filtered.fam ... done.
(пlink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

```
autosomal_snps.bim - hw2 (рабочая область) - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка
PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
(пlink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink
--bfile autosomal_snps_maf_filtered --hwe 1e-6 --make-bed --out final_autosomal_snps
PLINK v1.90b6.21 64-bit (19 Oct 2020)      www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang  GNU General Public License v3
Logging to final_autosomal_snps.log.
Options in effect:
--bfile autosomal_snps_maf_filtered
--hwe 1e-6
--make-bed
--out final_autosomal_snps

7848 MB RAM detected; reserving 3924 MB for main workspace.
1137402 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 5 het. haploid genotypes present (see final_autosomal_snps.hh ); many
commands treat these as missing.
Total genotyping rate is 0.999104.
Warning: --hwe observation counts vary by more than 10%, due to the X
chromosome. You may want to use a less stringent --hwe p-value threshold for X
chromosome variants.
--hwe: 0 variants removed due to Hardy-Weinberg exact test.
1137402 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
--make-bed to final_autosomal_snps.bed + final_autosomal_snps.bim +
final_autosomal_snps.fam ... done.
(пlink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

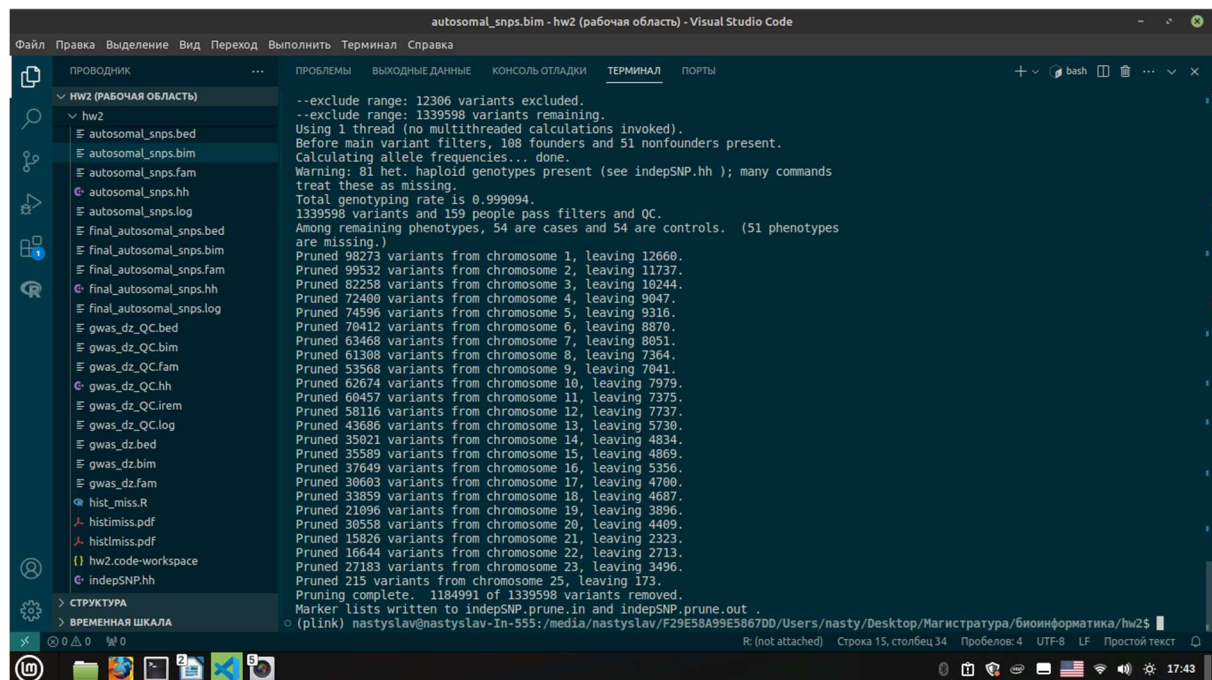
#### Задание 4.

Оставьте только те SNP, которые не коррелируют в высокой степени благодаря LD. Для этого нужно убрать регионы с заранее высоким LD (тестовый файл `inversion.txt`, изучите его содержимое), примените команду:

```
plink --bfile [Ваш файл] --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP
```

В команде `--indep-pairwise 50 5 0.2` числа - это, соответственно, размер окна (в SNPs), количество SNPs, на которые окно перемещается в каждой итерации и пороговый коэффициент корреляции.

Файл indepSNP.prune.in будет содержать SNPs, которые годятся для дальнейшего анализа.



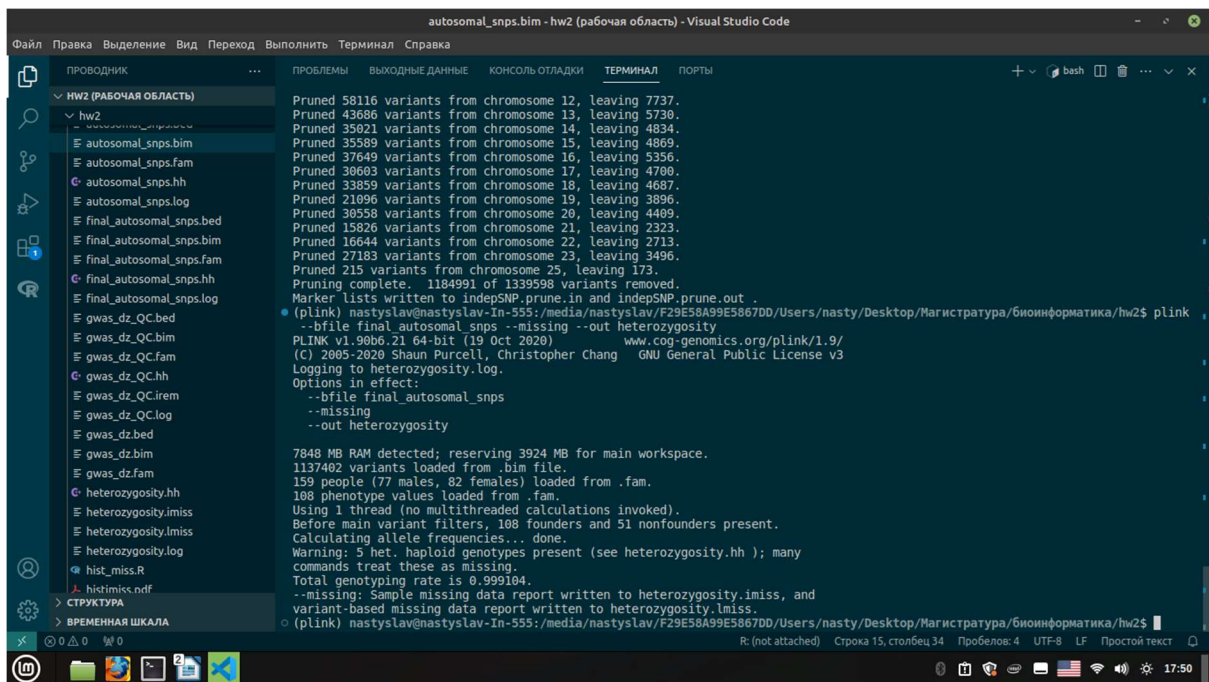
```
autosomal_snps.bim -hw2 (рабочая область) - Visual Studio Code
Файл  Правка  Выделение  Вид  Переход  Выполнить  Терминал  Справка
PROBLEMY  ВЫХОДНЫЕ ДАННЫЕ  КОНСОЛЬ ОТЛАДКИ  ТЕРМИНАЛ  ПОРТЫ
PROVODNIK
  HW2 (РАБОЧАЯ ОБЛАСТЬ)
    autosomal_snps.bed
    autosomal_snps.bim
    autosomal_snps.fam
    autosomal_snps.hh
    autosomal_snps.log
    final_autosomal_snps.bed
    final_autosomal_snps.bim
    final_autosomal_snps.fam
    final_autosomal_snps.hh
    final_autosomal_snps.log
    gwas_dz_QC.bed
    gwas_dz_QC.bim
    gwas_dz_QC.fam
    gwas_dz_QC.hh
    gwas_dz_QC.irem
    gwas_dz_QC.log
    gwas_dz.bed
    gwas_dz.bim
    gwas_dz.fam
    hist_miss.R
    histmiss.pdf
    histmiss.pdf
    hw2.code-workspace
    indepSNP.hh
СТРУКТУРА
ВРЕМЕННАЯ ШКАЛА
--exclude range: 12386 variants excluded.
--exclude range: 1339598 variants remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 81 het, haploid genotypes present (see indepSNP.hh ); many commands
treat these as missing.
Total genotyping rate is 0.999994.
1339598 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
Pruned 98273 variants from chromosome 1, leaving 12660.
Pruned 99532 variants from chromosome 2, leaving 11737.
Pruned 82258 variants from chromosome 3, leaving 10244.
Pruned 72480 variants from chromosome 4, leaving 9047.
Pruned 74596 variants from chromosome 5, leaving 9316.
Pruned 70412 variants from chromosome 6, leaving 8870.
Pruned 63468 variants from chromosome 7, leaving 8051.
Pruned 61308 variants from chromosome 8, leaving 7364.
Pruned 53568 variants from chromosome 9, leaving 7041.
Pruned 62674 variants from chromosome 10, leaving 7979.
Pruned 60457 variants from chromosome 11, leaving 7375.
Pruned 58116 variants from chromosome 12, leaving 7737.
Pruned 43686 variants from chromosome 13, leaving 5730.
Pruned 35021 variants from chromosome 14, leaving 4834.
Pruned 35589 variants from chromosome 15, leaving 4869.
Pruned 37649 variants from chromosome 16, leaving 5356.
Pruned 38083 variants from chromosome 17, leaving 4709.
Pruned 33059 variants from chromosome 18, leaving 4687.
Pruned 21096 variants from chromosome 19, leaving 3896.
Pruned 30558 variants from chromosome 20, leaving 4409.
Pruned 15826 variants from chromosome 21, leaving 2323.
Pruned 16644 variants from chromosome 22, leaving 2713.
Pruned 27183 variants from chromosome 23, leaving 3496.
Pruned 215 variants from chromosome 25, leaving 173.
Pruning complete. 1184991 of 1339598 variants removed.
Marker lists written to indepSNP.prune.in and indepSNP.prune.out .
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E38A99E38670D/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

## Задание 5.

Доли гетерозиготных SNPs каждого индивида должны не сильно отличаться друг от друга (исключаем инбридинг). Примем допустимую частоту гетерозиготности как не более 3 стандартных отклонений от среднего значения.

Рассчитайте частоты гетерозиготности (для этого нужны два показателя – общее число SNPs –  $N(NM)$ , и число SNPs в гомозиготном состоянии –  $O(NOM)$ ) для каждого индивида. Постройте гистограмму распределения полученных значений.

Определите, частоты гетерозиготности каких индивидов выходят за рамки допустимых значений, удалите их из анализа при помощи параметра remove в plink.



```
autosomal_snps.bim - hw2 (рабочая область) - Visual Studio Code
Файл Правка Выделение Вид Переход Выполнить Терминал Справка
ПРОВЕРКА ПРОБЛЕМЫ ВЫХОДНЫЕ ДАННЫЕ КОНСОЛЬ ОТЛАДКИ ТЕРМИНАЛ ПОРТЫ
HW2 (РАБОЧАЯ ОБЛАСТЬ)
hw2
autosomal_snps.bim
autosomal_snps.fam
autosomal_snps.hh
autosomal_snps.log
final_autosomal_snps.bed
final_autosomal_snps.bim
final_autosomal_snps.fam
final_autosomal_snps.hh
final_autosomal_snps.log
gwas_dz_QC.bed
gwas_dz_QC.bim
gwas_dz_QC.fam
gwas_dz_QC.hh
gwas_dz_QC.lim
gwas_dz_QC.log
gwas_dz.bed
gwas_dz.bim
gwas_dz.fam
heterozygosity.hh
heterozygosity.imiss
heterozygosity.lmiss
heterozygosity.log
hist_miss.R
histmiss.ndf
СТРУКТУРА
ВРЕМЕННАЯ ШКАЛА
Pruned 58116 variants from chromosome 12, leaving 7737.
Pruned 43686 variants from chromosome 13, leaving 5730.
Pruned 35021 variants from chromosome 14, leaving 4834.
Pruned 35589 variants from chromosome 15, leaving 4869.
Pruned 37649 variants from chromosome 16, leaving 5356.
Pruned 36603 variants from chromosome 17, leaving 4709.
Pruned 33859 variants from chromosome 18, leaving 4687.
Pruned 21096 variants from chromosome 19, leaving 3896.
Pruned 38558 variants from chromosome 20, leaving 4409.
Pruned 15826 variants from chromosome 21, leaving 2323.
Pruned 16644 variants from chromosome 22, leaving 2713.
Pruned 27183 variants from chromosome 23, leaving 3496.
Pruned 215 variants from chromosome 25, leaving 173.
Pruning complete. 1184991 of 1339598 variants removed.
Marker lists written to indepSNP.prune.in and indepSNP.prune.out.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink
--bfile final_autosomal_snps --missing --out heterozygosity
PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to heterozygosity.log.
Options in effect:
--bfile final_autosomal_snps
--missing
--out heterozygosity
7848 MB RAM detected; reserving 3924 MB for main workspace.
1137402 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 5 het. haploid genotypes present (see heterozygosity.hh ); many
commands treat these as missing.
Total genotyping rate is 0.999104.
--missing: Sample missing data report written to heterozygosity.imiss, and
variant-based missing data report written to heterozygosity.lmiss.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

Задание 6.

Pihat – показатель родственности между индивидами.

Близнецы имеют Pihat=1.0.

Сиблинги – 0.5.

Сиблинги второго порядка – 0.25...

Для определения родственности примените команду:

```
plink --bfile [Ваш файл] --extract indepSNP.prune.in --genome --min 0.2 --out pihat_min0.2
```

Исследуйте получившийся файл. Примите во внимание: PO = parent-offspring, UN = unrelated individuals. Что Вы можете сказать о составе исследуемых?

Отфильтруйте данные так, чтобы в него вошли только фаундеры:

```
plink --bfile [исходный файл] --filter-founders --make-bed --out [получившийся файл]
```

Рассчитайте Pihat для полученного файла. Сколько родственных связей осталось?

Как Вы считаете, какого из двух родственных индивидов следует удалить из анализа? Подсказка: примените следующую команду:

```
plink --bfile [файл] --missing
```

Указав FID и IID в отдельном файле, примените к нему команду --remove в plink.

Поздравляю! Ваши данные прошли контроль качества.

*Я удалила из анализа сиблингов второго порядка, потому что у меня в файле оставались в основном сиблинги.*



indepSNP.prune.in - hw2 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

PROBLEMY ВЫХОДНЫЕ ДАННЫЕ КОНСОЛЬ ОТЛАДКИ **ТЕРМИНАЛ** ПОРТЫ

```
hw2 > indepSNP.prune.in
1 rs1048488
2 rs12562034
3 rs12124819
4 rs4970383

(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/Биоинформатика/hw2$ plink
k --bfile final_autosomal_snps --extract indepSNP.prune.in --genome --min 0.2 --out pihat_min0.2
PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to pihat_min0.2.log.
Options in effect:
  --bfile final_autosomal_snps
  --extract indepSNP.prune.in
  --genome
  --min 0.2
  --out pihat_min0.2

7848 MB RAM detected; reserving 3924 MB for main workspace.
1137482 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
--extract: 134018 variants remaining.
Using up to 4 threads (change this with --threads).
Before main variant filters, 108 founders and 51 nonfounders present.
Calculating allele frequencies... done.
Warning: 4 het. haploid genotypes present (see pihat_min0.2.hh ); many commands
treat these as missing.
Total genotyping rate is 0.999073.
134018 variants and 159 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls. (51 phenotypes
are missing.)
Excluding 3090 variants on non-autosomes from IBD calculation.
IBD calculations complete.
Finished writing pihat_min0.2.genome .
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/Биоинформатика/hw2$
```

R: (not attached) Строка 1, столбец 1 Пробелов: 4 UTF-8 LF Простой текст 19:05

pihat\_min0.2.genome - hw2 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

PROBLEMY ВЫХОДНЫЕ ДАННЫЕ КОНСОЛЬ ОТЛАДКИ **ТЕРМИНАЛ** ПОРТЫ

```
hw2 > pihat_min0.2.genome
1 FID1 IID1 FID2 IID2 RT EZ Z0 Z1 Z2 PI_HAT PHE DST PPC RATI
2 1349 NA11843 1349 NA10853 P0 0.5 0.0030 0.9902 0.0068 0.5019 -1 0.851734 1.0000 472.75
3 1330 NA12341 1330 NA12335 P0 0.5 0.0000 1.0000 0.0000 0.5000 0 0.850395 1.0000 1896.0
4 1463 NA12877 1463 NA12898 P0 0.5 0.0025 0.9975 0.0000 0.4988 0 0.850776 1.0000 1252.6

Excluding 3090 variants on non-autosomes from IBD calculation.
IBD calculations complete.
Finished writing pihat_min0.2.genome .
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/Биоинформатика/hw2$ plink
--bfile final_autosomal_snps --filter-founders --make-bed --out founders_data
PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to founders_data.log.
Options in effect:
  --bfile final_autosomal_snps
  --filter-founders
  --make-bed
  --out founders_data

7848 MB RAM detected; reserving 3924 MB for main workspace.
1137482 variants loaded from .bim file.
159 people (77 males, 82 females) loaded from .fam.
108 phenotype values loaded from .fam.
51 people removed due to founder status (--filter-founders).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 108 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 5 het. haploid genotypes present (see founders_data.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.999071.
1137482 variants and 108 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls.
--make-bed to founders_data.bed + founders_data.bim + founders_data.fam ...
done.
(plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/Биоинформатика/hw2$
```

R: (not attached) Строка 1, столбец 1 Пробелов: 3 UTF-8 LF Простой текст 19:07

pihat\_min0.2.genome - hw2 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

PROBЛЕМЫ ВЫХОДНЫЕ ДАННЫЕ КОНСОЛЬ ОТЛАДКИ **ТЕРМИНАЛ** ПОРТЫ

```
--make-bed to founders_data.bed + founders_data.bim + founders_data.fam ...
done.
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink --bfile founders_data --genome --min 0.2 --out founders_pihat_min0.2

PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to founders_pihat_min0.2.log.
Options in effect:
  --bfile founders_data
  --genome
  --min 0.2
  --out founders_pihat_min0.2

7848 MB RAM detected; reserving 3924 MB for main workspace.
1137402 variants loaded from .bim file.
108 people (55 males, 53 females) loaded from .fam.
108 phenotype values loaded from .fam.
Using up to 4 threads (change this with --threads).
Before main variant filters, 108 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 5 het. haploid genotypes present (see founders_pihat_min0.2.hh ); many
commands treat these as missing.
Total genotyping rate is 0.999071.
1137402 variants and 108 people pass filters and QC.
Among remaining phenotypes, 54 are cases and 54 are controls.
Excluding 21827 variants on non-autosomes from IBD calculation.
IBD calculations complete.
Finished writing founders_pihat_min0.2.genome .
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

R: (not attached) Строка 1, столбец 1 Пробелов: 3 UTF-8 LF Простой текст 19:08

founders\_data.bim - hw2 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

PROBЛЕМЫ ВЫХОДНЫЕ ДАННЫЕ КОНСОЛЬ ОТЛАДКИ **ТЕРМИНАЛ** ПОРТЫ

```
Finished writing founders_pihat_min0.2.genome .
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$ plink --bfile founders_data --remove remove_list.txt --make-bed --out filtered_data

PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to filtered_data.log.
Options in effect:
  --bfile founders_data
  --make-bed
  --out filtered_data
  --remove remove_list.txt

7848 MB RAM detected; reserving 3924 MB for main workspace.
1137402 variants loaded from .bim file.
108 people (55 males, 53 females) loaded from .fam.
108 phenotype values loaded from .fam.
--remove: 107 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 107 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 5 het. haploid genotypes present (see filtered_data.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.999077.
1137402 variants and 107 people pass filters and QC.
Among remaining phenotypes, 53 are cases and 54 are controls.
--make-bed to filtered_data.bed + filtered_data.bim + filtered_data.fam ...
done.
• (plink) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2$
```

R: (not attached) Строка 23, столбец 34 Пробелов: 4 UTF-8 LF Простой текст 19:28