Содержание

1 Burrows-Wheeler transformation (5 points)

Create an index of a given word using an algorithm of Burrows-Wheeler transformation. Provide all the rotations you considered. The word is:

GENOMICS
rontation
1. GENOMICS$
2. ENOMICS$G
3. NOMICS$GE
4. OMICS$GEN
5. MICS$GENO
6. ICS$GENOM
7. CS$GENOMI
8. S$GENOMIC
9. $GENOMICS

alphabet sort
1. GENOMICS$
2. ENOMICS$G
3. NOMICS$GE
4. OMICS$GEN
5. MICS$GENO
6. ICS$GENOM
7. CS$GENOMI

8. S$GENOMIC
9. $GENOMICS

last symbols
SIGMONEC$

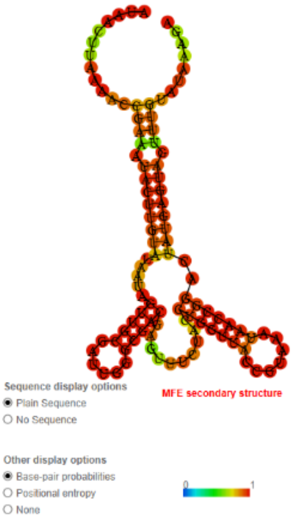
2 RNA-secondary structure (15 points)

You are provided with a file that contains 10 RNA sequences. You would need to predict its secondary structure using RNAfold, LocARNA, compare them with R-Chie and make a conclusion if these sequences contain conserved secondary structure.
Data: /srv/common/exam/rna_secondary_structure/V1.txt

2.1 RNAfold (4 points)

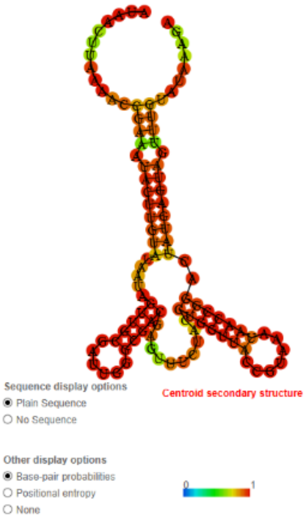
http://nibiru.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
Take the sequence with id >AL591981.1/205922-205823. Visualise with RNAfold. Provide RNA sequence, Vienna format sequence and Minimum Fold Energy plot.

Result:
AUAACUUAAAACCGAAAUACUUGUAUAAUAGUUGCGAUUGGGCGACGAGUUUCUAC
CUGGUUACCGUAAAUAACCGGACUAUGAGUAGUUUGUAUAAAGA
............((((.((((((((.....(((((......))))).........(((((((.......)))))))..)))))))).))))........


Sequence display options
● Plain Sequence
○ No Sequence

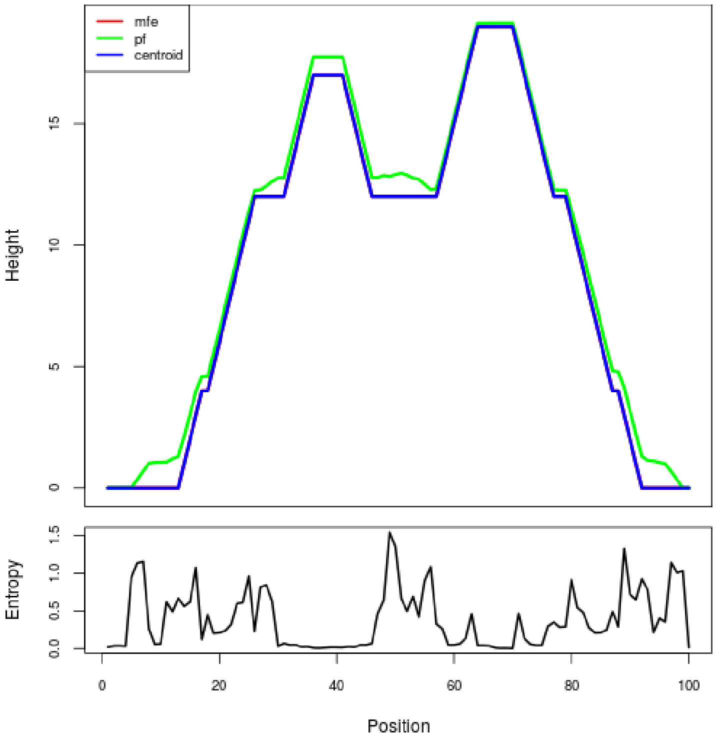
MFE secondary structure

Other display options
● Base-pair probabilities
○ Positional entropy
○ None


Sequence display options
● Plain Sequence
○ No Sequence

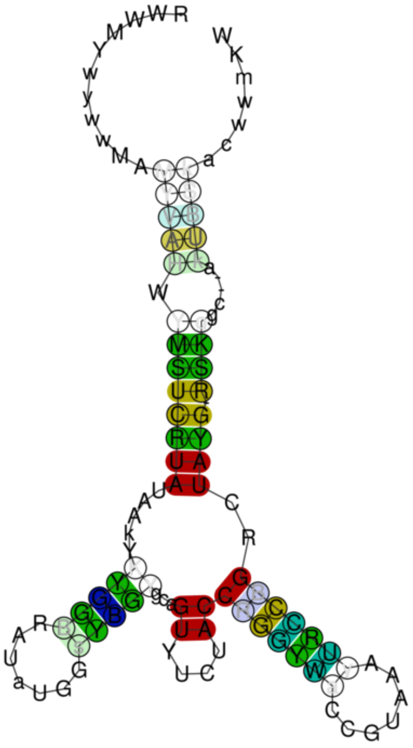
Centroid secondary structure

Other display options
● Base-pair probabilities
○ Positional entropy
○ None

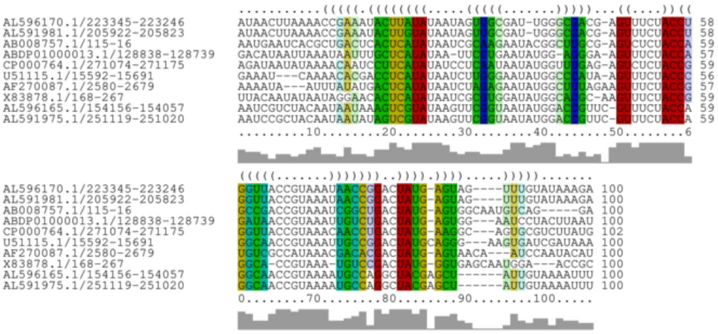


2.2 LocaRNA (4 points)

https://rna.informatik.uni-freiburg.de/LocARNA/Input.jsp
Upload the V1.txt file with your data. Provide the RNAalifold consensus structure. How many red type base-pairs can you see?



Red Base-Pairs: 5

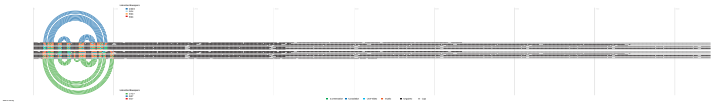


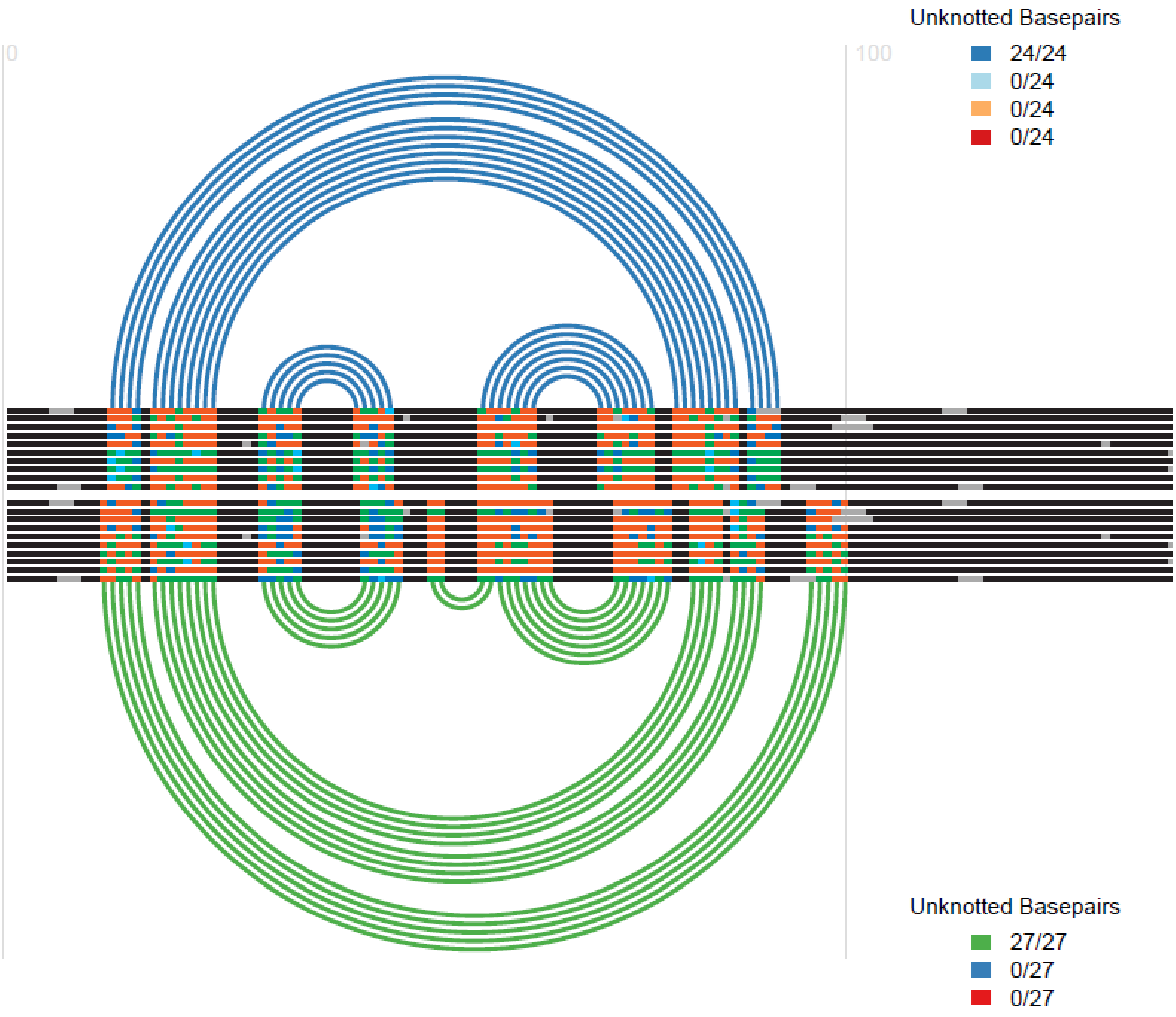
```
                                   ...........(((((.((((((((.....(((((.......)))))...((....))((
AL596170.1/223345-223246           AUAACUUAAAACCGAAAUACUUAUAUAAUAGUUGCGAU-UGGGCGACG-AGUUUCUACCU 58
AL591981.1/205922-205823           AUAACUUAAAACCGAAAUACUUGUAUAAUAGUUGCGAU-UGGGCGACG-AGUUUCUACCU 58
AB008757.1/115-16                  AAUGAAUCACGCUGACUCACUCAUAUAAUCGCAAGAAUACGGCUUGCG-AGUCUCUACCA 59
ABDP01000013.1/128838-128739       GACAUAAUUAAAUAAUUUGCUCAUAUAA-UUCCUGAAUAUGG-AGGGA-AGUUUCUACCA 57
CP000764.1/271074-271175           AGAUAAUAUAAAACAAUCCUUCAUAUAUCCUCAAUAAUAUGGUUUGAG-AGUCUCUACCG 59
U51115.1/15592-15691               GAAAU---CAAAACACGACCUCAUAUAAUCUUGGGAAUAUGGCCCAUA-AGUUUCUACCC 56
AF270087.1/2580-2679               AAAAUA---AUUUAUAUGACUCAUAUAAUCUAGAGAAUAUGGCUUUAGAAGUUUCUACCG 57
X83878.1/168-267                   UUACAAUAUAAUAGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGC-AAGUUUCUACCG 59
AL596165.1/154156-154057           AAUCGUCUACAAUAAUAAAGUCGUAUAAGUUCGGUAAUAUGGACCGUUC-GUUUCUACCA 59
AL591975.1/251119-251020           AAUCCGCUACAAUAAUAUAGUCGUAUAAGUUCGGUAAUAUGGACCGUUC-GUUUCUACCA 59
                                   .........10........20........30........40........50........6

                                   (((((.......)))))))..)))).)))).....)))))......
AL596170.1/223345-223246           GGUUACCGUAAAUAACCGGACUAUG-AGUAG----UUUGUAUAAAGA 100
AL591981.1/205922-205823           GGUUACCGUAAAUAACCGGACUAUG-AGUAG----UUUGUAUAAAGA 100
AB008757.1/115-16                  GCCGACCGUAAAUCGGCUGACUAUG-AGUGGCAAUGUCAG-----GA 100
ABDP01000013.1/128838-128739       GAUAACCGUAAAUUGUCUGACUAUG-AGUGG---AAUCCUACUUAAU 100
CP000764.1/271074-271175           GGUUACCGUAAACAACCUGACUAUG-AAGGC---AGUGCGUCUUAUG 102
U51115.1/15592-15691               GGCAACCGUAAAUUGCCGGACUAUGCAGGG---AAGUGAUCGAUAAA 100
AF270087.1/2580-2679               UGUCGCCAUAAACGACACGACUAUG-AGUAACA---AUCCAAUACAU 100
X83878.1/168-267                   GGCA-CCGUAAA-UGUCCGACUAUG-GGUGAGCAAUGGA---ACCGC 100
AL596165.1/154156-154057           GGCAACCGUAAAAUGCCAGGCUACGAGCU------AUUGUAAAAUUU 100
AL591975.1/251119-251020           GGCAACCGUAAAAUGCCAGGCUACGAGCU------AUUGUAAAAUUU 100
                                   0........70........80........90........100.....
```

2.3 R-Chie (7 points)

https://e-rna.org/r-chie/

You have to match the dot-bracket formula from RNAfold with LocaRNA consensus structure by length. Do it using this colab code (the instructions are inside):
https://colab.research.google.com/drive/1dQZQgV2Y63VPjoqrqPC1B8Y5xG_HfzPv?usp=sharing





Result: https://e-rna.org/r-chie/results.cgi?id=VGhtPHSdux

Provide a plot with comparison of two structures. Do structures contain different secondary structure elements? Does our data contain conserved secondary structure?

На представленной визуализации видно, что структуры РНК для обеих последовательностей демонстрируют высокий уровень консервативности. Консервативные элементы, обозначенные зелёным цветом, указывают на области, которые сохраняют свою структуру в обеих РНК, что свидетельствует о значимой роли этих участков.

На графике также отображены "неперекрещенные"парные основания, что подтверждает успешное отображение общих черт структуры благодаря совместной обработке файлов, полученных на этапе Align&Fold. Эти пары демонстрируют устойчивые структурные элементы, такие как шпильки, которые сохраняются между последовательностями.

Для улучшения восприятия графического вывода программы было выполнено обрезание изображения, чтобы точнее рассмотреть количество спаренных нуклеотидов и их расположение. На основе визуализации можно отметить следующее:

- Верхняя часть отображает неперекрещенные парные основания, представляющие первую последовательность из предыдущего задания.

- Нижняя часть показывает результаты выравнивания для 14 РНК-последовательностей, что позволяет увидеть общие элементы, характерные для данной группы.

Также отмечены unknotted base pairs, свидетельствуя о том, что совместная загрузка файлов прошлого этапа, отразила общие черты структуры.

3 Imputation (20 points)

You have a vcf file with 10 samples for the region chr1:1000000-5000000. We masked a part of the variants as a simulation of “missing” variants.

Requirements:
Beagle jar
Data: ground_truth_hg38.vcf.gz
Data: test_hg38.vcf.gz
Reference panel: 1kG_hg38_chr1.vcf.gz
Genetic maps: beagle_hg38_chr1.map

3.1 Imputation 1

Use the given script to run a correct phasing/imputation command and impute the masked file. Provide the number of variants in the output file (10 points)

Code

```
java -Xmx4g -jar beagle.29Oct24.c8e.jar \
  gt=test_hg38.vcf.gz \
  ref=1kG_hg38_chr1.vcf.gz \
  map=beagle_hg38_chr1.map \
  out=imputed_output
```

Ответ: 135164

3.2 Imputation 2

Study the effect of filtration of rare alleles (that have MAF lower than 5%) by calculating concordance matrix of the imputation WITH and WITHOUT filtration using the ground truth. (5 points)

Query Sample EGAN0000	Genotyped Sample EGAN0000	Discordance	Average -log P(HWE)	Number of sites compared	Number of matching genotypes
1375753	1375753	8.289306e+02	7.469652e−02	133930	133930
1375753	1375750	6.714338e+04	2.290539e−02	133930	133930
1375753	1375749	5.901065e+04	2.658607e−02	133930	133930
1375753	1375754	5.705806e+04	2.992230e−02	133930	133930
1375753	1375745	7.246696e+04	2.039288e−02	133930	133930
1375753	1375751	6.560525e+04	2.417863e−02	133930	133930
1375753	1375747	6.938149e+04	2.009476e−02	133930	133930
1375753	1375755	6.855256e+04	2.258329e−02	133930	133930
1375753	1375743	6.787100e+04	2.233790e−02	133930	133930
1375753	1375748	6.633287e+04	2.491584e−02	133930	133930
1375750	1375753	6.734601e+04	2.288812e−02	133930	133930
1375750	1375750	1.086820e+03	6.842761e−02	133930	133930
1375750	1375749	6.044746e+04	2.427207e−02	133930	133930
1375750	1375754	6.589999e+04	2.233813e−02	133930	133930

1375750	1375745	$6.088035e+04$	$2.352153e-02$	133930	133930
1375750	1375751	$5.668964e+04$	$2.621802e-02$	133930	133930
1375750	1375747	$6.171849e+04$	$2.473032e-02$	133930	133930
1375750	1375755	$5.928696e+04$	$2.699354e-02$	133930	133930
1375750	1375743	$6.586314e+04$	$2.109568e-02$	133930	133930
1375750	1375748	$6.963017e+04$	$2.018064e-02$	133930	133930
1375749	1375753	$5.901065e+04$	$2.648532e-02$	133930	133930
1375749	1375750	$6.024484e+04$	$2.422458e-02$	133930	133930
1375749	1375749	$8.473513e+02$	$6.670613e-02$	133930	133930
1375749	1375754	$6.581709e+04$	$2.141872e-02$	133930	133930
1375749	1375745	$5.479231e+04$	$2.742881e-02$	133930	133930
1375749	1375751	$6.080667e+04$	$2.396158e-02$	133930	133930
1375749	1375747	$5.679096e+04$	$2.545256e-02$	133930	133930
1375749	1375755	$6.869072e+04$	$2.073788e-02$	133930	133930
1375749	1375743	$6.264874e+04$	$2.308044e-02$	133930	133930
1375749	1375748	$6.039220e+04$	$2.589115e-02$	133930	133930
1375754	1375753	$5.705806e+04$	$2.974744e-02$	133930	133930
1375754	1375750	$6.560525e+04$	$2.230935e-02$	133930	133930
1375754	1375749	$6.574341e+04$	$2.145109e-02$	133930	133930
1375754	1375754	$1.068399e+03$	$7.062692e-02$	133930	133930
1375754	1375745	$6.676576e+04$	$2.240849e-02$	133930	133930
1375754	1375751	$6.887493e+04$	$2.076450e-02$	133930	133930
1375754	1375747	$6.384608e+04$	$2.290571e-02$	133930	133930
1375754	1375755	$6.450922e+04$	$2.368917e-02$	133930	133930
1375754	1375743	$5.951722e+04$	$2.636853e-02$	133930	133930
1375754	1375748	$7.002622e+04$	$2.078129e-02$	133930	133930
1375745	1375753	$7.220907e+04$	$2.052660e-02$	133930	133930
1375745	1375750	$6.082509e+04$	$2.340652e-02$	133930	133930
1375745	1375749	$5.497652e+04$	$2.732630e-02$	133930	133930
1375745	1375754	$6.696838e+04$	$2.234889e-02$	133930	133930
1375745	1375745	$8.657720e+02$	$6.800868e-02$	133930	133930
1375745	1375751	$5.821856e+04$	$2.496488e-02$	133930	133930
1375745	1375747	$6.718943e+04$	$2.021466e-02$	133930	133930
1375745	1375755	$6.947360e+04$	$1.990031e-02$	133930	133930
1375745	1375743	$5.960011e+04$	$2.596605e-02$	133930	133930
1375745	1375748	$6.251979e+04$	$2.524018e-02$	133930	133930
1375751	1375753	$6.553157e+04$	$2.418215e-02$	133930	133930
1375751	1375750	$5.650544e+04$	$2.617158e-02$	133930	133930
1375751	1375749	$6.058562e+04$	$2.416972e-02$	133930	133930
1375751	1375754	$6.887493e+04$	$2.084222e-02$	133930	133930
1375751	1375745	$5.810804e+04$	$2.503001e-02$	133930	133930
1375751	1375751	$1.252606e+03$	$7.098667e-02$	133930	133930
1375751	1375747	$6.555920e+04$	$2.159184e-02$	133930	133930
1375751	1375755	$6.915124e+04$	$2.066809e-02$	133930	133930
1375751	1375743	$6.795389e+04$	$2.143524e-02$	133930	133930
1375751	1375748	$6.637892e+04$	$2.212453e-02$	133930	133930
1375747	1375753	$6.921571e+04$	$2.018664e-02$	133930	133930
1375747	1375750	$6.153428e+04$	$2.475522e-02$	133930	133930
1375747	1375749	$5.701201e+04$	$2.540516e-02$	133930	133930
1375747	1375754	$6.393818e+04$	$2.290049e-02$	133930	133930

1375747	1375745	6.709733e+04	2.029168e−02	133930	133930
1375747	1375751	6.576183e+04	2.154580e−02	133930	133930
1375747	1375747	1.049979e+03	6.871131e−02	133930	133930
1375747	1375755	4.923848e+04	3.251167e−02	133930	133930
1375747	1375743	6.613024e+04	2.012106e−02	133930	133930
1375747	1375748	6.777889e+04	2.188202e−02	133930	133930
1375755	1375753	6.866309e+04	2.256276e−02	133930	133930
1375755	1375750	5.928696e+04	2.694500e−02	133930	133930
1375755	1375749	6.905913e+04	2.061847e−02	133930	133930
1375755	1375754	6.471185e+04	2.372003e−02	133930	133930
1375755	1375745	6.949202e+04	2.000474e−02	133930	133930
1375755	1375751	6.926176e+04	2.071774e−02	133930	133930
1375755	1375747	4.940427e+04	3.243383e−02	133930	133930
1375755	1375755	1.123662e+03	7.144853e−02	133930	133930
1375755	1375743	6.425133e+04	2.234665e−02	133930	133930
1375755	1375748	5.995011e+04	2.669008e−02	133930	133930
1375743	1375753	6.799994e+04	2.237249e−02	133930	133930
1375743	1375750	6.582630e+04	2.117671e−02	133930	133930
1375743	1375749	6.292505e+04	2.306354e−02	133930	133930
1375743	1375754	5.977511e+04	2.647914e−02	133930	133930
1375743	1375745	5.989484e+04	2.586450e−02	133930	133930
1375743	1375751	6.815652e+04	2.147074e−02	133930	133930
1375743	1375747	6.635129e+04	2.013774e−02	133930	133930
1375743	1375755	6.428818e+04	2.232827e−02	133930	133930
1375743	1375743	1.234186e+03	7.099688e−02	133930	133930
1375743	1375748	6.916045e+04	2.122739e−02	133930	133930
1375748	1375753	6.619472e+04	2.494398e−02	133930	133930
1375748	1375750	6.936307e+04	2.025944e−02	133930	133930
1375748	1375749	6.062246e+04	2.579988e−02	133930	133930
1375748	1375754	6.999859e+04	2.087327e−02	133930	133930
1375748	1375745	6.262110e+04	2.517580e−02	133930	133930
1375748	1375751	6.629603e+04	2.220326e−02	133930	133930
1375748	1375747	6.775126e+04	2.189212e−02	133930	133930
1375748	1375755	5.957248e+04	2.685303e−02	133930	133930
1375748	1375743	6.893019e+04	2.123661e−02	133930	133930
1375748	1375748	1.132872e+03	7.361196e−02	133930	133930

Таблица 1: Discordance and Genotype Comparison for Unfiltered Data

Query Sample EGAN0000	Genotyped Sample EGAN0000	Discordance	Average -log P(HWE)	Number of sites compared	Number of matching genotypes
1375753	1375753	7.184065e+02	7.016587e−01	14237	14237
1375753	1375750	6.705128e+04	2.143966e−01	14237	14237
1375753	1375749	5.884486e+04	2.491032e−01	14237	14237
1375753	1375754	5.681859e+04	2.805530e−01	14237	14237
1375753	1375745	7.238406e+04	1.907537e−01	14237	14237

1375753	1375751	6.539342e+04	2.264754e−01	14237	14237
1375753	1375747	6.928018e+04	1.879865e−01	14237	14237
1375753	1375755	6.840520e+04	2.114401e−01	14237	14237
1375753	1375743	6.775126e+04	2.090873e−01	14237	14237
1375753	1375748	6.613945e+04	2.334127e−01	14237	14237
1375750	1375753	6.723548e+04	2.142878e−01	14237	14237
1375750	1375750	9.947168e+02	6.426322e−01	14237	14237
1375750	1375749	6.028168e+04	2.273350e−01	14237	14237
1375750	1375754	6.566052e+04	2.092073e−01	14237	14237
1375750	1375745	6.079746e+04	2.201855e−01	14237	14237
1375750	1375751	5.647781e+04	2.456604e−01	14237	14237
1375750	1375747	6.161718e+04	2.315941e−01	14237	14237
1375750	1375755	5.913960e+04	2.529281e−01	14237	14237
1375750	1375743	6.574341e+04	1.974016e−01	14237	14237
1375750	1375748	6.943676e+04	1.888678e−01	14237	14237
1375749	1375753	5.890013e+04	2.481273e−01	14237	14237
1375749	1375750	6.015273e+04	2.268064e−01	14237	14237
1375749	1375749	6.815652e+02	6.265198e−01	14237	14237
1375749	1375754	6.557762e+04	2.005582e−01	14237	14237
1375749	1375745	5.470942e+04	2.569420e−01	14237	14237
1375749	1375751	6.059483e+04	2.244336e−01	14237	14237
1375749	1375747	5.668964e+04	2.383882e−01	14237	14237
1375749	1375755	6.854335e+04	1.940800e−01	14237	14237
1375749	1375743	6.252900e+04	2.160725e−01	14237	14237
1375749	1375748	6.019878e+04	2.425876e−01	14237	14237
1375754	1375753	5.694753e+04	2.788147e−01	14237	14237
1375754	1375750	6.551315e+04	2.087896e−01	14237	14237
1375754	1375749	6.557762e+04	2.007975e−01	14237	14237
1375754	1375754	8.289306e+02	6.634686e−01	14237	14237
1375754	1375745	6.668286e+04	2.097150e−01	14237	14237
1375754	1375751	6.866309e+04	1.943581e−01	14237	14237
1375754	1375747	6.374477e+04	2.144296e−01	14237	14237
1375754	1375755	6.436186e+04	2.218434e−01	14237	14237
1375754	1375743	5.939749e+04	2.470042e−01	14237	14237
1375754	1375748	6.983280e+04	1.945182e−01	14237	14237
1375745	1375753	7.209854e+04	1.920726e−01	14237	14237
1375745	1375750	6.073298e+04	2.191108e−01	14237	14237
1375745	1375749	5.481074e+04	2.560667e−01	14237	14237
1375745	1375754	6.672892e+04	2.093085e−01	14237	14237
1375745	1375745	7.828789e+02	6.386841e−01	14237	14237
1375745	1375751	5.800672e+04	2.338719e−01	14237	14237
1375745	1375747	6.708812e+04	1.891144e−01	14237	14237
1375745	1375755	6.932623e+04	1.862009e−01	14237	14237
1375745	1375743	5.948038e+04	2.432180e−01	14237	14237
1375745	1375748	6.232637e+04	2.364637e−01	14237	14237
1375751	1375753	6.542105e+04	2.264611e−01	14237	14237
1375751	1375750	5.641333e+04	2.451222e−01	14237	14237
1375751	1375749	6.041983e+04	2.263721e−01	14237	14237
1375751	1375754	6.863546e+04	1.951350e−01	14237	14237
1375751	1375745	5.802514e+04	2.343761e−01	14237	14237

1375751	1375751	1.040768e+03	6.668070e-01	14237	14237
1375751	1375747	6.545789e+04	2.020698e-01	14237	14237
1375751	1375755	6.900387e+04	1.934235e-01	14237	14237
1375751	1375743	6.783416e+04	2.005959e-01	14237	14237
1375751	1375748	6.618551e+04	2.071543e-01	14237	14237
1375747	1375753	6.910518e+04	1.888745e-01	14237	14237
1375747	1375750	6.144218e+04	2.317982e-01	14237	14237
1375747	1375749	5.684622e+04	2.379942e-01	14237	14237
1375747	1375754	6.369871e+04	2.144975e-01	14237	14237
1375747	1375745	6.701444e+04	1.898018e-01	14237	14237
1375747	1375751	6.554999e+04	2.017079e-01	14237	14237
1375747	1375747	9.486651e+02	6.453310e-01	14237	14237
1375747	1375755	4.909111e+04	3.048382e-01	14237	14237
1375747	1375743	6.601051e+04	1.882331e-01	14237	14237
1375747	1375748	6.758548e+04	2.048730e-01	14237	14237
1375755	1375753	6.855256e+04	2.112271e-01	14237	14237
1375755	1375750	5.919486e+04	2.523979e-01	14237	14237
1375755	1375749	6.889335e+04	1.929649e-01	14237	14237
1375755	1375754	6.447238e+04	2.222071e-01	14237	14237
1375755	1375745	6.940913e+04	1.871024e-01	14237	14237
1375755	1375751	6.904992e+04	1.939183e-01	14237	14237
1375755	1375747	4.930295e+04	3.040623e-01	14237	14237
1375755	1375755	9.762961e+02	6.711242e-01	14237	14237
1375755	1375743	6.413160e+04	2.091697e-01	14237	14237
1375755	1375748	5.975669e+04	2.501032e-01	14237	14237
1375743	1375753	6.788942e+04	2.094373e-01	14237	14237
1375743	1375750	6.573420e+04	1.981345e-01	14237	14237
1375743	1375749	6.275926e+04	2.159661e-01	14237	14237
1375743	1375754	5.953564e+04	2.481625e-01	14237	14237
1375743	1375745	5.981195e+04	2.422262e-01	14237	14237
1375743	1375751	6.794468e+04	2.010018e-01	14237	14237
1375743	1375747	6.624998e+04	1.883908e-01	14237	14237
1375743	1375755	6.414081e+04	2.090412e-01	14237	14237
1375743	1375743	1.114451e+03	6.668311e-01	14237	14237
1375743	1375748	6.896703e+04	1.987148e-01	14237	14237
1375748	1375753	6.608419e+04	2.336278e-01	14237	14237
1375748	1375750	6.927097e+04	1.895057e-01	14237	14237
1375748	1375749	6.045667e+04	2.417073e-01	14237	14237
1375748	1375754	6.975912e+04	1.954271e-01	14237	14237
1375748	1375745	6.253821e+04	2.357475e-01	14237	14237
1375748	1375751	6.608419e+04	2.078928e-01	14237	14237
1375748	1375747	6.764995e+04	2.048946e-01	14237	14237
1375748	1375755	5.942512e+04	2.516064e-01	14237	14237
1375748	1375743	6.881045e+04	1.987273e-01	14237	14237
1375748	1375748	9.394547e+02	6.915058e-01	14237	14237

Таблица 2: Discordance and Genotype Comparison for Filtered Data

Где информацию о каждом столбце можно получить из результата скрипта:

- Genotyped sample
- Discordance, given either as an abstract score or number of mismatches, see the options -E/-u in man page for details. Note that samples with high missingness have fewer sites compared, which results in lower overall discordance. Therefore it is advisable to use the average score per site rather than the absolute value, i.e. divide the value by the number of sites compared (smaller value = better match)
- Average negative log of HWE probability at matching sites, attempts to quantify the following intuition: rare genotype matches are more informative than common genotype matches, hence two samples with similar discordance can be further stratified by the HWE score (bigger value = better match, the observed concordance was less likely to occur by chance)
- Number of sites compared for this pair of samples (bigger = more informative)
- Number of matching genotypes

3.3 Imputation 3

What is the genotype of the samples EGAN00001375753 and EGAN00001375754 at the site "4100418"? If you know that the imputation is totally correct at this position for both samples, what do you expect the fields of DS and GP must be for each sample? (5 points)

First line - unfiltered, second line - filtered

Query Sample EGAN0000	Genotyped Sample EGAN0000	Discordance	Average -log P(HWE)	Number of sites compared	Number of matching genotypes
1375753	1375754	$5.705806e{+}04$	$2.992230e{-}02$	133930	133930
1375753	1375754	$5.681859e{+}04$	$2.805530e{-}01$	14237	14237

1375753 gt:0|0 imputed:0|1
1375754 gt:0|0 imputed:0|0

4 Differential Expression (20 points)

The Cancer Genome Atlas (TCGA) is one of the biggest biological databases for cancer. It contains different projects and tissues. Lung Adinocarcinoma (LUAD) is one of the classical datasets that are studied for expression. You are provided 10 samples (5 controls and 5 cancer patients) with raw counts of 1000 genes.

4.1 Differential Expression 1

Apply the protocol of differential gene expression using DESeq2 (or edgeR if you want). Note: DON'T FORGET to normalize manually or by whatever metric (TPM, etc..), or built-in flag you want, but state how you normalized. What is the p-value for the most significant result? (10 points)

4.2 Differential Expression 2

Adjust your results for multiple testing using FDR. and provide the top10 genes after adjustment with alpha = 0.01 (confidence level of 99%) (5 points)

4.3 Differential Expression 3

Provide a volcano plot for your experimental output and Top10 (ordered by log2 fold change) genes downregulated and upregulated. (5 points)

Python and R solutions:

Python

Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.multitest import multipletests
from scipy import stats

data = pd.read_csv('TCGA_LUAD_10_samples.csv')
data.columns = [f'Gene_{i+1}' for i in range(data.shape[1])]

controls = data.iloc[:, 5:10].values
cancer = data.iloc[:, 0:5].values

normalized_controls = np.log1p(controls)
normalized_cancer = np.log1p(cancer)

normalized_data = np.concatenate((normalized_controls, normalized_cancer), axis=0)

p_values = []
log2_fold_changes = []
```

Code

```
for gene in range(normalized_data.shape[1]):
    control_mean = np.mean(normalized_controls[:, gene])
    cancer_mean = np.mean(normalized_cancer[:, gene])
    fold_change = cancer_mean - control_mean  # log2 fold change
    t_stat, p_val = stats.ttest_ind(normalized_controls[:, gene],
                                    normalized_cancer[:, gene],
                                    equal_var=False)

    p_values.append(p_val)
    log2_fold_changes.append(fold_change)
results_df = pd.DataFrame({
    ’Gene’: [f’Gene_{i+1}’ for i in range(len(p_values))],
    ’Log2_Fold_Change’: log2_fold_changes,
    ’P_Value’: p_values
})

results_df[’Adj_P_Value’] = multipletests(results_df[’P_Value’], method=’fdr_bh’)[1]

most_significant_p_value = results_df[’P_Value’].min()
top_genes = results_df[results_df[’Adj_P_Value’] < 0.01].nlargest(10,
                                                   ’Log2_Fold_Change’)

print(f’Most significant p-value: {most_significant_p_value}’)
print(f’Top 10 genes after FDR adjustment:\n{top_genes}’)

# Volcano plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x=’Log2_Fold_Change’,
                y=-np.log10(results_df[’P_Value’]),
                data=results_df,
                alpha=0.6)
plt.axhline(y=-np.log10(0.01), color=’red’, linestyle=’--’)
plt.axvline(x=0, color=’grey’, linestyle=’--’)
plt.title(’Volcano Plot of Differential Expression’)
plt.xlabel(’Log2 Fold Change’)
plt.ylabel(’-Log10 P-Value’)
plt.show()
```

it can be done in R too (and result would be much better):

Code

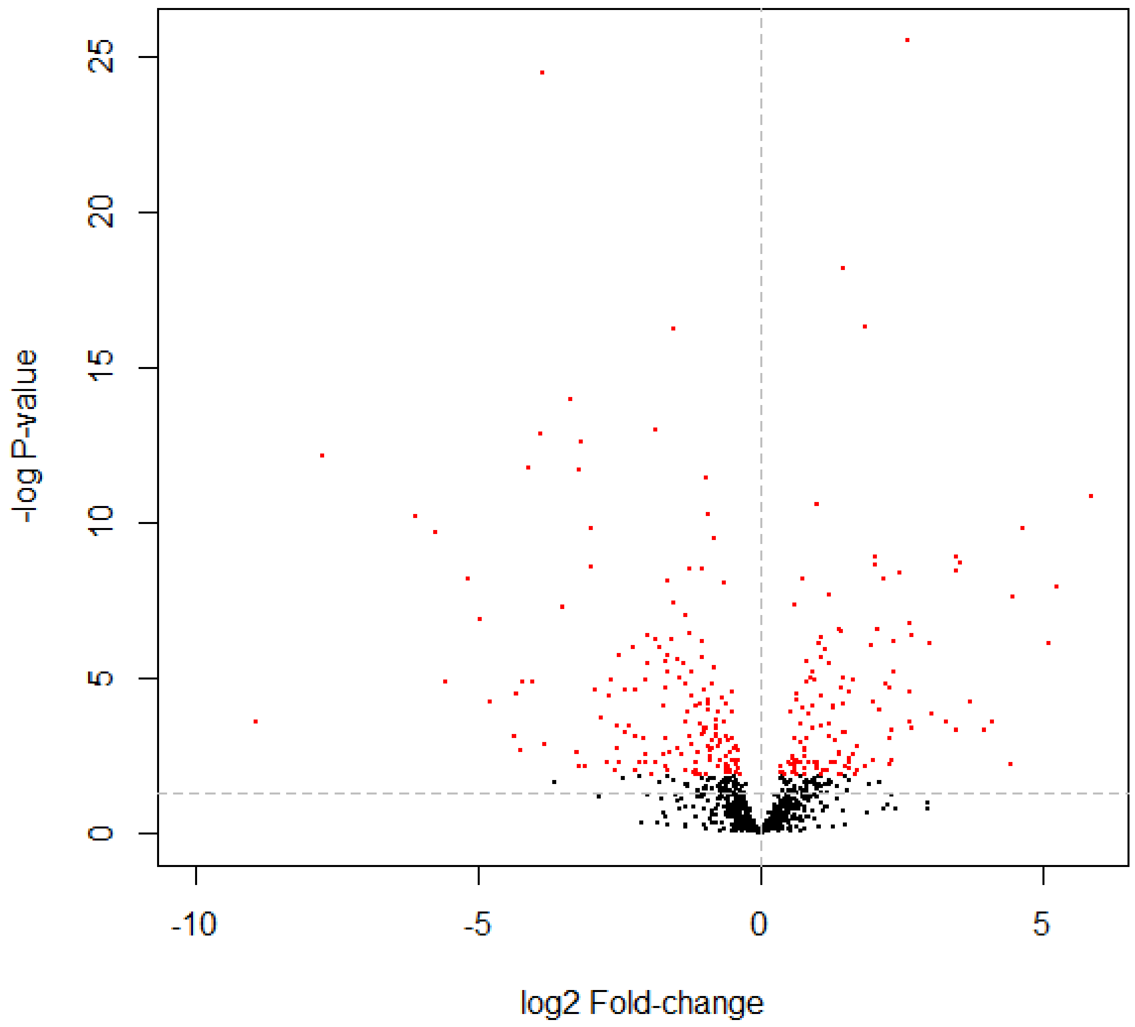
```
library("DESeq2")

sample_names <- c("Cancer1", "Cancer2", "Cancer3", "Cancer4", "Cancer5",
                  "Normal1", "Normal2", "Normal3", "Normal4", "Normal5")

count_table <- read.csv("C://Users//vlad2//Downloads//exam
                        //Diff_Expr//TCGA_LUAD_10_samples.csv")
count_matrix <- as.matrix(count_table)

col_data <- data.frame(
  row.names = colnames(count_matrix),
  condition = rep(c("Cancer", "Normal"), each = 5)
)

dds <- DESeqDataSetFromMatrix(countData = count_matrix,
                              colData = col_data,
                              design = ~ condition)
dds <- DESeq(dds)

res_alpha_01 <- results(dds, alpha = 0.01)
ordered_res_alpha_01 <- res_alpha_01[order(res_alpha_01$padj), ]

top10_foldchange <- head(ordered_res_alpha_01[order(
                            abs(ordered_res_alpha_01$log2FoldChange),
                                decreasing = TRUE), ], 10)
print(top10_foldchange)

top_significant_genes <- ordered_res_alpha_01[!is.na(ordered_res_alpha_01$padj)
                                               & ordered_res_alpha_01$padj < 0.05, ]
top_significant_genes <- head(top_significant_genes[order(abs(
                                      top_significant_genes$log2FoldChange),
                                      decreasing = TRUE), ], 10)
print(top_significant_genes)

plot(res_alpha_01$log2FoldChange,
     -log10(res_alpha_01$pvalue),
     xlab = "log2 Fold-change",
     ylab = "-log P-value", pch = 20, cex = 0.5)
points(res_alpha_01$log2FoldChange[res_alpha_01$padj < 0.05],
       -log10(res_alpha_01$pvalue[res_alpha_01$padj < 0.05]),
       col = "red", pch = 20, cex = 0.5)
abline(v = 0, h = -log10(0.05), lty = "dashed", col = "grey")
```

got:

Code

```
     baseMean log2FoldChange     lfcSE      stat     pvalue       padj
    <numeric>      <numeric> <numeric> <numeric>  <numeric>  <numeric>
1     47.5961       -8.93650  2.462102  -3.62962 2.83834e-04 1.96673e-03
2   3022.7178       -7.73918  1.078292  -7.17726 7.11245e-13 6.86352e-11
3    112.6118       -6.11392  0.936973  -6.52518 6.79219e-11 3.85557e-09
4  25628.8541        5.86533  0.868220   6.75558 1.42261e-11 9.80586e-10
5   1094.2735       -5.75766  0.905267  -6.36018 2.01518e-10 9.72325e-09
6    145.9469       -5.57844  1.284207  -4.34388 1.39991e-05 1.55277e-04
7    585.2159        5.25078  0.919493   5.71052 1.12633e-08 3.01920e-07
8    129.8693       -5.20071  0.897241  -5.79633 6.77801e-09 2.03536e-07
9     22.0635        5.11229  1.034799   4.94037 7.79731e-07 1.28926e-05
10  2203.8164       -4.96443  0.943136  -5.26375 1.41147e-07 3.16761e-06
```

with the following plot:



5 Microbiome (20 points)

Data for this tasks came from Fecal Microbiome Transplant (FMT) study involving children under 18 with autism and gastrointestinal disorders, assessed through the Autism Diagnostic Interview-Revised (ADI-R) and Gastrointestinal Symptom Rating Scale (GSRS). The study aimed to reduce behavioral and gastrointestinal symptoms by tracking microbiome changes, autism severity metrics, and GSRS scores over an 18-week period. Data includes five treated with FMT individuals and five controls, each contributing six to sixteen samples, along with five samples of the transplanted fecal material. The data were sequenced on two Illumina MiSeq runs. Follow the steps and answer the questions.
Data: /srv/common/exam/microbiome/
* Steps of sequence quality control, generation of phylogenetic tree and calculation of core-metrics were performed for you. Perform the alpha and beta diversity analysis with QIIME2.
Activate environment with Qiime2
1. configure conda
/opt/tljh/user/bin/conda init
2. restart session
source /.bashrc
3. activate session
conda activate /opt/tljh/user/envs/qiime2-amplicon-2024.10
Calculate alpha-beta diversity
Substitute “???” with correct file names. For visualization files you can use any name you want.
Use statistic of Faith’s Phylogenetic Diversity in core-metrics-results

Code

```
qiime diversity alpha-group-significance \
  --i-alpha-diversity core-metrics-results/???.qza \
  --m-metadata-file ???.tsv \
  --o-visualization ???.qzv
```

Use statistic of unweighted UniFrac distance in core-metrics-results

Code

```
qiime diversity beta-group-significance \
  --i-distance-matrix core-metrics-results/???.qza \
  --m-metadata-file ???.tsv \
  --m-metadata-column treatment-group \
  --o-visualization ???.qzv \
  --p-pairwise
```

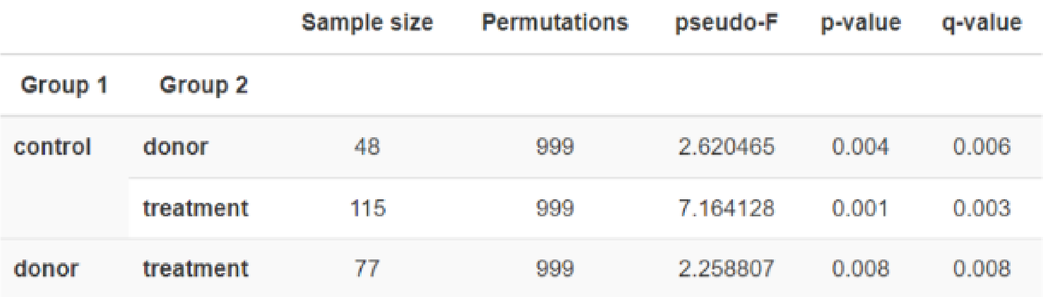
Visualize results using Galaxy website: https://usegalaxy.eu/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fiuc%2Fqiime_extract_viz%2Fqiime_extract_viz%2F0.1.0%2Bgalaxy0&version=latest
You would need to upload .qzv files and Galaxy will output .html file.

5.1 Alpha-diversity (4 points)

What is the p-value for Kruskal-Wallis statistic between control and treatment group in column treatment-group? Is it significant? Provide screenshot.

5.2 Beta-diversity (4 points)

Analyze pairwise permanova results. What are the pseudo-F and p-value between control and treatment? Provide screenshot.



		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
control	donor	48	999	2.620465	0.004	0.006
	treatment	115	999	7.164128	0.001	0.003
donor	treatment	77	999	2.258807	0.008	0.008

The transplantation of Fecal Microbiome in the treatment group significantly alters the structure compared to the control group ($p = 0.001$, pseudo-F $= 7.16$). The differences between donors and recipients are also significant, but less pronounced (pseudo-F $= 2.26$). This indicates a partial transfer of bacteria from the donor; however, the Microbiome of the recipients retains its individuality and differs from both the donors and the control group.

5.3 Experimental design (8 points)

Do samples differ in composition by subject-id (i.e., across individuals)?

5.4 Explanation of beta-diversity (8 points)

What does this suggest to you about what is changing in the microbiome with fecal microbiota transplant?