Read the https://pairtools.readthedocs.io/en/latest/ and https://cooler.readthedocs.io/en/latest/index.html for more information

In [1]:
```
!pip install -q condacolab
import condacolab
condacolab.install()
```

```
⏬ Downloading https://github.com/conda-forge/miniforge/releases/download/23.1.0-1/Mambaforge-23.1.0-1-Linux-x86_64.sh...
📦 Installing...
📌 Adjusting configuration...
🩹 Patching environment...
⏲ Done in 0:00:26
🔁 Restarting kernel...
```

Homework report should include:

1. scaling plot in log-log coordinates with description; create correct labels for scaling plot, including units of measurement; make comments on operations in cell starting with ##!!!
2. replicates clusterization for all files (in directory for the lecture) with dendrogram and description; make comments on operations in cells starting with ##!!!

In [2]:
```
%%bash
pip install cooler
pip install hicrep
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting cooler
  Downloading cooler-0.9.1-py2.py3-none-any.whl (103 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 103.9/103.9 kB 3.8 MB/s eta 0:00:00
Collecting pyyaml
  Downloading PyYAML-6.0-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (682 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 682.2/682.2 kB 10.4 MB/s eta 0:00:00
Collecting numpy>=1.9
  Downloading numpy-1.24.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (17.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 17.3/17.3 MB 35.4 MB/s eta 0:00:00
Collecting click>=7
  Downloading click-8.1.3-py3-none-any.whl (96 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 96.6/96.6 kB 5.3 MB/s eta 0:00:00
Collecting h5py>=2.5
  Downloading h5py-3.8.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.6/4.6 MB 28.2 MB/s eta 0:00:00
Collecting multiprocess
  Downloading multiprocess-0.70.14-py310-none-any.whl (134 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.3/134.3 kB 3.6 MB/s eta 0:00:00
Collecting pandas>1.0
  Downloading pandas-2.0.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.3/12.3 MB 54.0 MB/s eta 0:00:00
Collecting asciitree
  Downloading asciitree-0.3.3.tar.gz (4.0 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Collecting scipy>=0.16
  Downloading scipy-1.10.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (34.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 34.4/34.4 MB 28.7 MB/s eta 0:00:00
Collecting cytoolz
  Downloading cytoolz-0.12.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 69.9 MB/s eta 0:00:00
Collecting simplejson
  Downloading simplejson-3.19.1-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (137 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 137.9/137.9 kB 15.5 MB/s eta 0:00:00
Collecting pyfaidx
  Downloading pyfaidx-0.7.2.1-py3-none-any.whl (28 kB)
Collecting tzdata>=2022.1
  Downloading tzdata-2023.3-py2.py3-none-any.whl (341 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 341.8/341.8 kB 33.5 MB/s eta 0:00:00
Collecting pytz>=2020.1
  Downloading pytz-2023.3-py2.py3-none-any.whl (502 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 502.3/502.3 kB 41.9 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.2
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 247.7/247.7 kB 26.4 MB/s eta 0:00:00
Requirement already satisfied: toolz>=0.8.0 in /usr/local/lib/python3.10/site-packages (from cytoolz->cooler) (0.12.0)
Collecting dill>=0.3.6
  Downloading dill-0.3.6-py3-none-any.whl (110 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 110.5/110.5 kB 14.0 MB/s eta 0:00:00
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/site-pack
```

```
ages (from pyfaidx->cooler) (65.6.3)
Collecting six
  Downloading six-1.16.0-py2.py3-none-any.whl (11 kB)
Building wheels for collected packages: asciitree
  Building wheel for asciitree (setup.py): started
  Building wheel for asciitree (setup.py): finished with status 'done'
  Created wheel for asciitree: filename=asciitree-0.3.3-py3-none-any.whl size=50
34 sha256=b0cc5495e8725dbead3a4eaab73520532c04b52bae94218d397cdb2e4907376b
  Stored in directory: /root/.cache/pip/wheels/7f/4e/be/1171b40f43b918087657ec57
cf3b81fa1a2e027d8755baa184
Successfully built asciitree
Installing collected packages: pytz, asciitree, tzdata, six, simplejson, pyyaml,
numpy, dill, cytoolz, click, scipy, python-dateutil, pyfaidx, multiprocess, h5p
y, pandas, cooler
Successfully installed asciitree-0.3.3 click-8.1.3 cooler-0.9.1 cytoolz-0.12.1 d
ill-0.3.6 h5py-3.8.0 multiprocess-0.70.14 numpy-1.24.3 pandas-2.0.1 pyfaidx-0.7.
2.1 python-dateutil-2.8.2 pytz-2023.3 pyyaml-6.0 scipy-1.10.1 simplejson-3.19.1
six-1.16.0 tzdata-2023.3
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-whe
els/public/simple/
Collecting hicrep
  Downloading hicrep-0.2.6.tar.gz (22 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Collecting Deprecated
  Downloading Deprecated-1.2.13-py2.py3-none-any.whl (9.6 kB)
Requirement already satisfied: numpy>=1.17.0 in /usr/local/lib/python3.10/site-p
ackages (from hicrep) (1.24.3)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/site-packages
(from hicrep) (1.10.1)
Requirement already satisfied: cooler in /usr/local/lib/python3.10/site-packages
(from hicrep) (0.9.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/site-packages
(from hicrep) (2.0.1)
Requirement already satisfied: h5py in /usr/local/lib/python3.10/site-packages
(from hicrep) (3.8.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.10/site-pa
ckages (from cooler->hicrep) (0.70.14)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.10/site-packages
(from cooler->hicrep) (6.0)
Requirement already satisfied: simplejson in /usr/local/lib/python3.10/site-pack
ages (from cooler->hicrep) (3.19.1)
Requirement already satisfied: pyfaidx in /usr/local/lib/python3.10/site-package
s (from cooler->hicrep) (0.7.2.1)
Requirement already satisfied: cytoolz in /usr/local/lib/python3.10/site-package
s (from cooler->hicrep) (0.12.1)
Requirement already satisfied: asciitree in /usr/local/lib/python3.10/site-packa
ges (from cooler->hicrep) (0.3.3)
Requirement already satisfied: click>=7 in /usr/local/lib/python3.10/site-packag
es (from cooler->hicrep) (8.1.3)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/site-pa
ckages (from pandas->hicrep) (2023.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.
10/site-packages (from pandas->hicrep) (2.8.2)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/site-
packages (from pandas->hicrep) (2023.3)
Collecting wrapt<2,>=1.10
  Downloading wrapt-1.15.0-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.ma
nylinux_2_17_x86_64.manylinux2014_x86_64.whl (78 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 78.4/78.4 kB 6.6 MB/s eta 0:00:00
```

```
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/site-packag
es (from python-dateutil>=2.8.2->pandas->hicrep) (1.16.0)
Requirement already satisfied: toolz>=0.8.0 in /usr/local/lib/python3.10/site-pa
ckages (from cytoolz->cooler->hicrep) (0.12.0)
Requirement already satisfied: dill>=0.3.6 in /usr/local/lib/python3.10/site-pac
kages (from multiprocess->cooler->hicrep) (0.3.6)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/site-pack
ages (from pyfaidx->cooler->hicrep) (65.6.3)
Building wheels for collected packages: hicrep
  Building wheel for hicrep (setup.py): started
  Building wheel for hicrep (setup.py): finished with status 'done'
  Created wheel for hicrep: filename=hicrep-0.2.6-py3-none-any.whl size=34756 sh
a256=a3156108648ff66d31dd117a9a20baa34f6dc0ba791d1d453e1f45c2bafef696
  Stored in directory: /root/.cache/pip/wheels/21/75/06/6354db4851e5edd4899df43e
ded6240031885d7df3d5184d7a
Successfully built hicrep
Installing collected packages: wrapt, Deprecated, hicrep
Successfully installed Deprecated-1.2.13 hicrep-0.2.6 wrapt-1.15.0
```

```
WARNING: Running pip as the 'root' user can result in broken permissions and con
flicting behaviour with the system package manager. It is recommended to use a v
irtual environment instead: https://pip.pypa.io/warnings/venv
WARNING: Running pip as the 'root' user can result in broken permissions and con
flicting behaviour with the system package manager. It is recommended to use a v
irtual environment instead: https://pip.pypa.io/warnings/venv
```

In [1]:
```
import matplotlib.pyplot as plt
```

In [2]:
```
import cooler
```

In [3]:
```
import numpy as np
```

In [4]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [5]:
```
import seaborn as sns
import pandas as pd
import hicrep
from hicrep import hicrepSCC
from hicrep.utils import readMcool
```

In [6]:
```
!gdown 1dtuPlh4PR6kJPmwReRKWcOhAcAx-YyBE
```

```
Downloading...
From: https://drive.google.com/uc?id=1dtuPlh4PR6kJPmwReRKWcOhAcAx-YyBE
To: /content/HiC1.dm3.mapq_30.1000.mcool
100% 124M/124M [00:02<00:00, 50.6MB/s]
```

In [7]:
```
mcool ='HiC1.dm3.mapq_30.1000.mcool'
```

In [8]:
```
resolution = 20000
clr = cooler.Cooler(f'{mcool}::resolutions/{resolution}')
```

Tasks for seminar:

1. get info and attributes of Hi-C matrix with cooler.info
2. open cooler object as balanced matrix for intrachromosomal contacts
3. open cooler as unbalanced matrix for interchromosomal contacts
4. get table with coordinates and contacts, are they raw or balanced?
5. get the table in command line with command *cooler dump*
6. look at the table with bins, which columns present there?
7. plot a piece of map (log)
8. scaling plot (in log - log coordinates)
9. replicates clusterization

In [9]:
```
m=clr.matrix(balance=True).fetch('chrX')
```

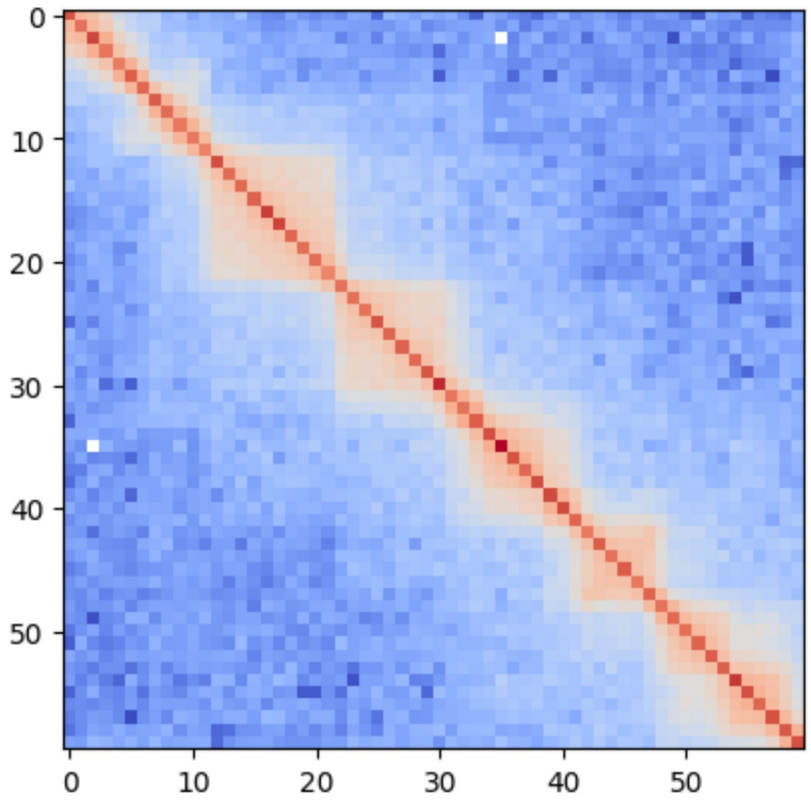
In [10]:
```
pix=clr.pixels(join=True)[:]
```

In [11]:
```
bins=clr.bins()[:]
```

In [12]:
```
plt.imshow(np.log(m[120:180,120:180]),cmap='coolwarm')
```

```
<ipython-input-12-57c1b47ce89e>:1: RuntimeWarning: divide by zero encountered in
log
  plt.imshow(np.log(m[120:180,120:180]),cmap='coolwarm')
```

Out[12]:
```
<matplotlib.image.AxesImage at 0x7f651800c6d0>
```



In [13]:
```
##!!! write comments for each row
z=np.zeros(len(m)) # создаём пустой массив нулей размера |m|
for i in range(len(m)): # для каждой i-ой диагонали
  z[i]=np.nanmean(np.diagonal(m,i)) # добавим в массив среднее число count-ов в
```

```
<ipython-input-13-08370e28c84e>:4: RuntimeWarning: Mean of empty slice
  z[i]=np.nanmean(np.diagonal(m,i)) # добавим в массив среднее число count-ов в диагонали
```

In [16]:
```
##!!! write comments for each row
##!!! why do we paste 20000 below?
# 20000 - размер бина матрицы Hi-C. i-ый бин соответствует расстоянию i*20000
plt.plot(np.arange(len(m))*20000,z)
# строим зависимость логарифма среднего числа count-ов диагоналей в зависимости
plt.xscale('log') # лог. шкала абсцисс
plt.yscale('log') # лог. шкала ординат

plt.title('Scaling plot') # Заголовок графика
plt.xlabel('Distance, b.p.') # подпись оси абсцисс
plt.ylabel('Contact Probability'); # подпись оси ординат
```

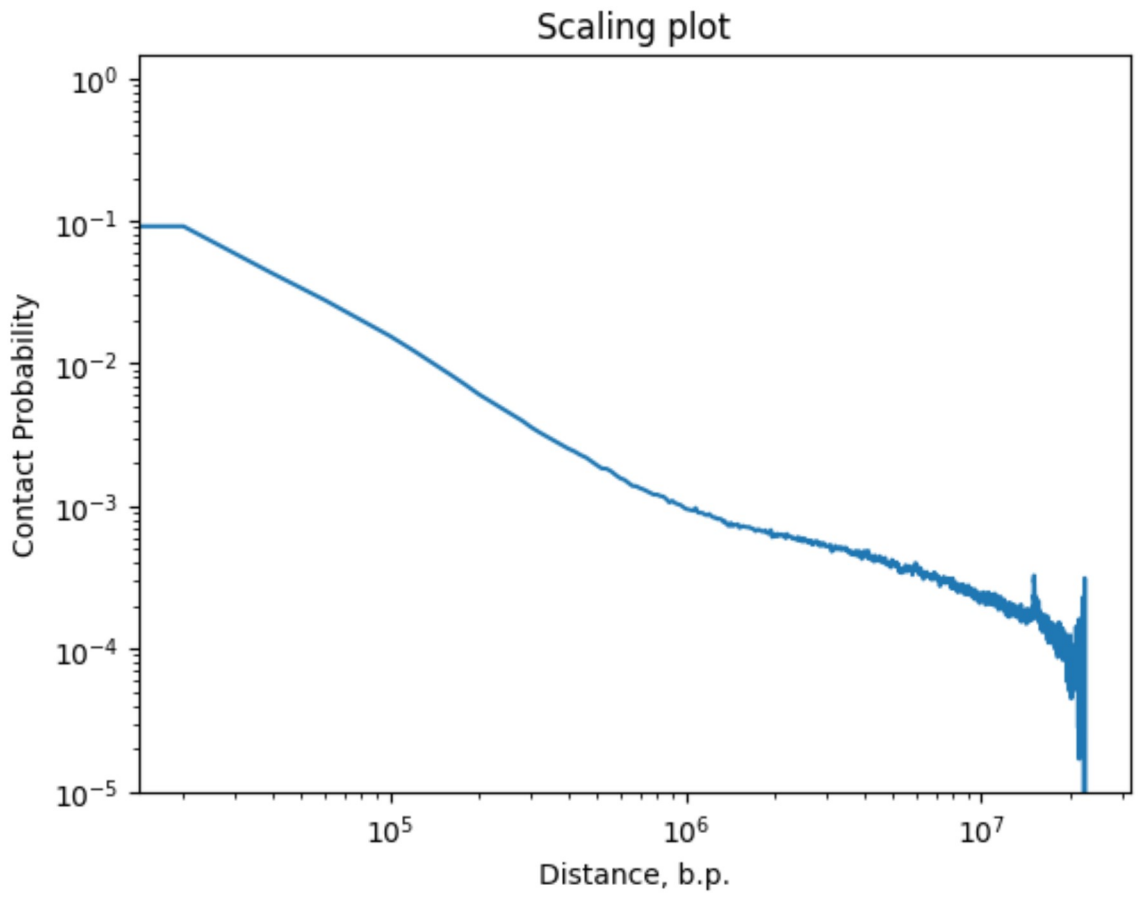


График показывает частоту (вероятность) контактов между участками в зависимости от расстояния между ними. По оси абсцисс находится расстояние между участками, по оси ординат - частота контактов. Для хороших данных мы ожидаем увидеть плавно спадающий график.

In [ ]:

Replicates clusterization with stratum-adjusted correlation coefficient (scc)

We have replicates for 2 *drosophila* cell lines: Bg3 and Kc167.
Bg3 - nervous cell line (HiC1...., HiC2... files)
Kc167 - embryonic cell line (HiC3...., HiC4... files)
The aim is to conduct replicates clusterization, using scc and demonstrate that replicates of same cell line tend to be closer to each other comparing with different cell types.
Hicrep can only calculate scc for each chromosome separately.
This is why you obtain several values with hicrepSCC function (run code below). The number of the values is equal to the number of chromosomes in cool file. So lets imagine, you take 2 mcool (or cool) files and decide to calculate scc between Hi-C matrices storing in these files only for chromosome 'chr2L', than for 'chr2R', etc. You will obtain as many scc as there are chromosomes in the Hi-C map. This is exactly what hicrepSCC function gives as an output. But then, to get single general scc value for 2 Hi-C maps (2 replicates) you should calculate average value of scc across all the chromosomes. So you have an scc (averaged across chrs) for each pair of samples (HiC1-HiC2,HiC1-HiC3,HiC1-HiC4,HiC2-HiC3,HiC2-HiC4,HiC3-HiC4). Now you can use these values as the measure of similarity between each 2 samples and build a dendrogram.
To do this, you should construct symmetric matrix of similarity from calculated SCCs with ones on the diagonal. This matrix should be used as an input for 'linkage' function (see below)

In [17]:
```
!gdown 1dtuPlh4PR6kJPmwReRKWcOhAcAx-YyBE
!gdown 1MPzzTmu3jNymHpLNv9oQkkOQR_r3C1AM
```

```
Downloading...
From: https://drive.google.com/uc?id=1dtuPlh4PR6kJPmwReRKWcOhAcAx-YyBE
To: /content/HiC1.dm3.mapq_30.1000.mcool
100% 124M/124M [00:00<00:00, 168MB/s]
Downloading...
From: https://drive.google.com/uc?id=1MPzzTmu3jNymHpLNv9oQkkOQR_r3C1AM
To: /content/HiC2.dm3.mapq_30.1000.mcool
100% 127M/127M [00:04<00:00, 28.1MB/s]
```

In [18]:
```
!gdown 1hXry50UpQwOkr6kCeLVzL_Qca4o6BwN-
!gdown 1cG2rfyjV0Mx-VO3ftXZSJ6tZG9MJXF7B
```

```
Downloading...
From: https://drive.google.com/uc?id=1hXry50UpQwOkr6kCeLVzL_Qca4o6BwN-
To: /content/HiC3.dm3.mapq_30.1000.mcool
100% 104M/104M [00:01<00:00, 54.5MB/s]
Downloading...
From: https://drive.google.com/uc?id=1cG2rfyjV0Mx-VO3ftXZSJ6tZG9MJXF7B
To: /content/HiC4.dm3.mapq_30.1000.mcool
100% 126M/126M [00:04<00:00, 31.3MB/s]
```

In [19]:
```
### the code is for calculation of scc between HiC1.dm3.mapq_30.1000.mcool and H
##!!! describe the next four parameter (as comments)
binSize = 20000 # значение размера бина
dBPMax = 5000000 # макс. расстояние, на котором будут учитываться контакты
bDownSample = True # делать ли уменьшение семплирования для несовпадающих по ра
h=0 # параметр для фильтра внутри алгоритма  при работе с матрицами

fmcool1 = 'HiC1.dm3.mapq_30.1000.mcool'
fmcool2 = 'HiC2.dm3.mapq_30.1000.mcool'
cool1, binSize1 = readMcool(fmcool1, binSize)
cool2, binSize2 = readMcool(fmcool2, binSize)
scc=hicrepSCC(cool1, cool2, h, dBPMax, bDownSample)
```

```
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
```

In [21]:
```
scc
```

Out[21]:
```
array([0.6129397 , 0.5768267 , 0.76475088, 0.5767983 , 0.88104427,
       0.62127888,        nan])
```

In [31]:
```
# Now calculate scc for each pair of samples, average across chromosomes and cre

def scc_pairwise(cool1, cool2):
    cool1, binSize1 = readMcool(cool1, binSize)
    cool2, binSize2 = readMcool(cool2, binSize)
    return np.nanmean(hicrepSCC(cool1, cool2, h, dBPMax, bDownSample))


corr_matrix = np.ones((4, 4))
template = 'HiC{}.dm3.mapq_30.1000.mcool'
for i, j in ((0, 1), (0, 2), (0, 3), (1, 2), (1, 3), (2, 3)):
    val = scc_pairwise(template.format(i+1), template.format(j+1))
    corr_matrix[i][j] = val
    corr_matrix[j][i] = val

corr_matrix
```

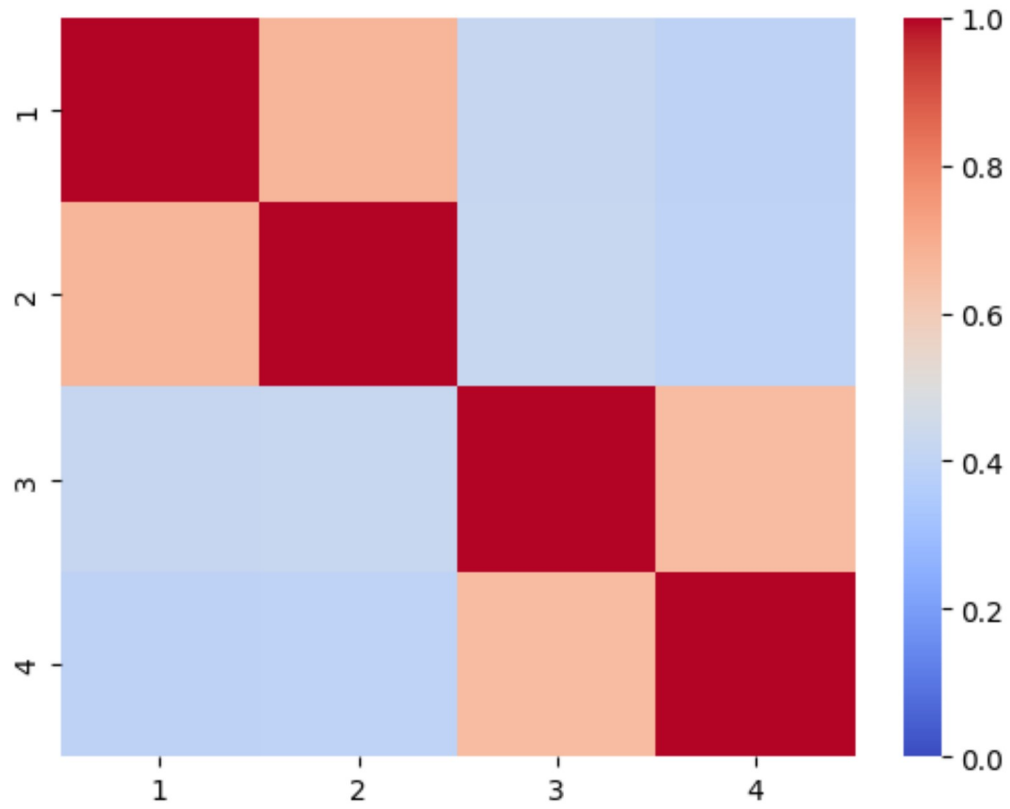
```
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
/usr/local/lib/python3.10/site-packages/hicrep/hicrep.py:91: RuntimeWarning: inv
alid value encountered in double_scalars
  return rhoNan2Zero @ wsNan2Zero / wsNan2Zero.sum()
```

Out[31]:
```
array([[1.        , 0.67293392, 0.42014761, 0.39182045],
       [0.67293392, 1.        , 0.4242003 , 0.39491033],
       [0.42014761, 0.4242003 , 1.        , 0.65523491],
       [0.39182045, 0.39491033, 0.65523491, 1.        ]])
```

In [36]:
```
import seaborn as sns

sns.heatmap(corr_matrix, cmap="coolwarm", xticklabels=[1, 2, 3, 4], yticklabels=
```

Out[36]: <Axes: >



как можно видеть, первые два образца коррелируют между собой, как и два последних. Первые два слабо коррелируют с последними двумя.

In [ ]:
```
## A piece of code for the dendrogram generation
```

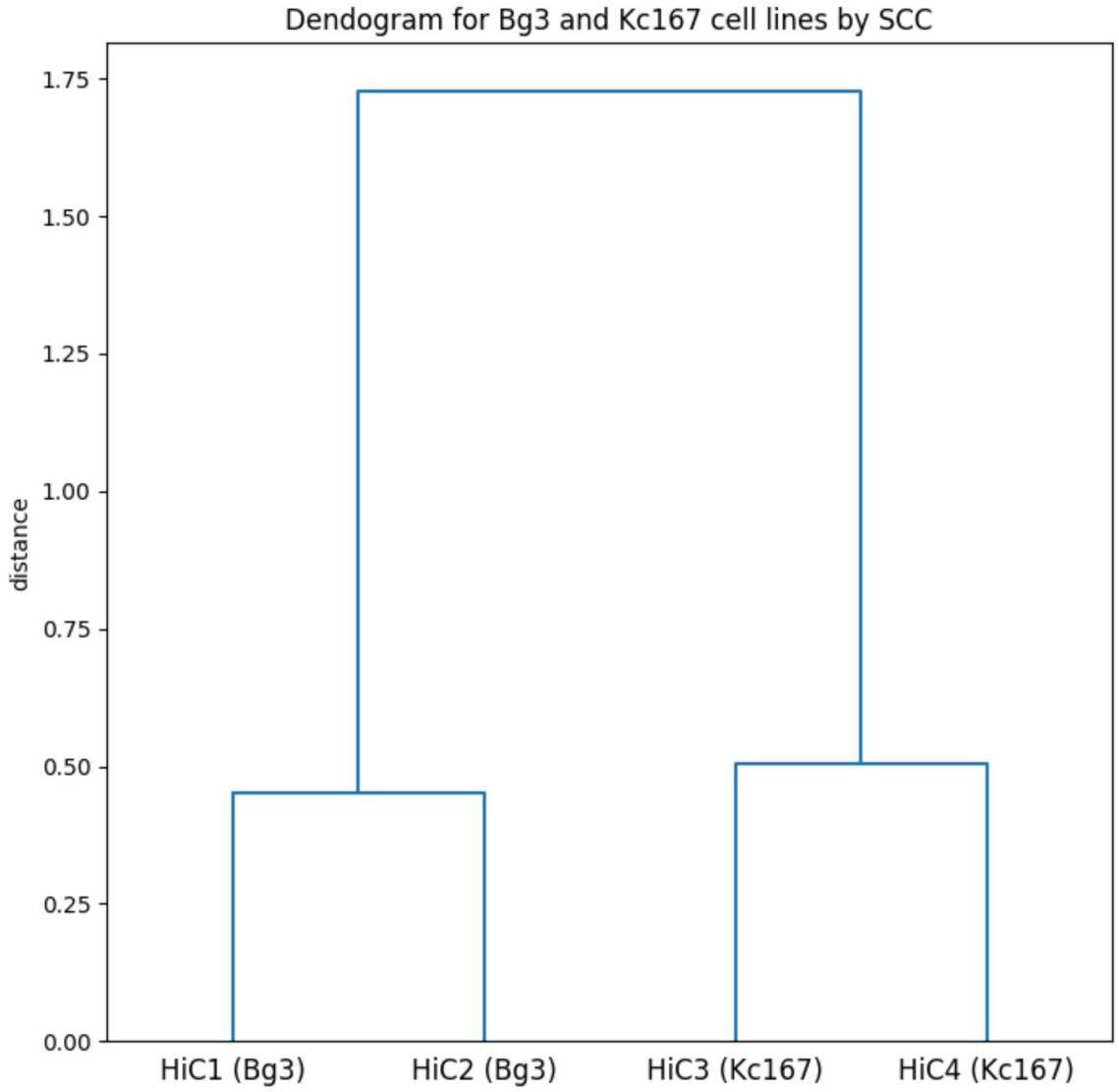
In [41]:
```
from scipy.cluster.hierarchy import linkage, dendrogram


Z=linkage(corr_matrix, 'single', 'correlation')
plt.figure(figsize=(8,8))
plt.ylabel('distance')

plt.title('Dendogram for Bg3 and Kc167 cell lines by SCC')
dendrogram(Z, color_threshold=0, labels=['HiC1 (Bg3)', 'HiC2 (Bg3)', 'HiC3 (Kc16
```

Out[41]:

```
{'icoord': [[5.0, 5.0, 15.0, 15.0],
  [25.0, 25.0, 35.0, 35.0],
  [10.0, 10.0, 30.0, 30.0]],
 'dcoord': [[0.0, 0.45007397803361926, 0.45007397803361926, 0.0],
  [0.0, 0.5049415804235364, 0.5049415804235364, 0.0],
  [0.45007397803361926,
   1.7279985591740132,
   1.7279985591740132,
   0.5049415804235364]],
 'ivl': ['HiC1 (Bg3)', 'HiC2 (Bg3)', 'HiC3 (Kc167)', 'HiC4 (Kc167)'],
 'leaves': [0, 1, 2, 3],
 'color_list': ['C0', 'C0', 'C0'],
 'leaves_color_list': ['C0', 'C0', 'C0', 'C0']}
```

In [ ]:

In [ ]: