Mgr. Ondřej Vaculík
437307@mail.muni.cz

DSIB01 Autumn 2021

05 Motif Detection

Overview

- Peak calling - brief overview
- Motif representation in biology
  - PPM
  - PWM
  - sequence logos
- Tools
  - Bedops
  - Bedtools
  - The MEME Suite
    - MEME-ChIP
    - Tomtom
- Demo on real dataset
- Homework - Individual work

Clip-seq analysis - peak calling

- a statistical procedure, which uses coverage properties of CLIP and Input samples to find regions which are enriched due to protein binding
- requires mapped reads, and outputs a set of regions, which represent the putative binding locations. Each region is usually associated with a significance score which is an indicator of enrichment
- many different tools for peak calling available:
  - **iCount**
  - **Paraclu**
  - **PureCLIP**
  - **Piranha**

Sequence motifs

- a **nucleotide or amino-acid sequence pattern** that is widespread and usually assumed to be **related to biological function** of the macromolecule

- **short, recurring patterns** in DNA/RNA that are presumed to have a biological function. Often they i**ndicate sequence-specific binding sites** for proteins such as nucleases, transcription factors, RNA-binding proteins. Others are **involved in important processes** at the RNA level, including ribosome binding, mRNA processing (splicing, editing, polyadenylation) and transcription termination.

Sequence motif representation - PPMs

- a **position probability matrix**
- in general:
  - there's one row for each symbol of the alphabet and one column for each position in the pattern
- in **PPM** each number is a **probability of nucleotide** occurrence **in given position** (sum of each column is 1)

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

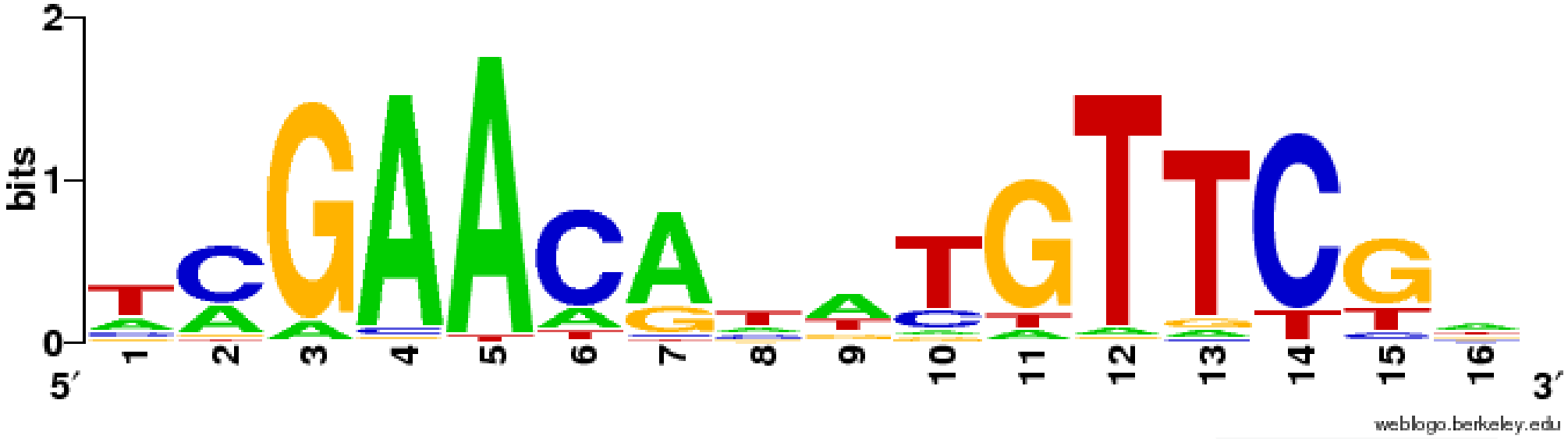
Sequence motif representation - PWMs

- a **position weight matrix**
  - also known as a position-specific weight matrix (**PSWM**) or position-specific scoring matrix (**PSSM**)
  - **the most commonly used**
- the elements in PWMs are calculated as **log likelihoods**
- PWMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}.$$

Sequence motif representation - Sequence logos

- Graphical representation of PWMs
  - **the bigger letter the higher chance for the nucleotide to appear in the position**



Tools - BEDOPS + bedtools

- **BEDOPS:**
  - open-source command-line toolkit that performs efficient and scalable Boolean and other set operations, statistical calculations, archiving, conversion and other management of genomic data of arbitrary scale
  - https://bedops.readthedocs.io/en/latest/
  - functions for today: **sort-bed**, **bedextract**
- **bedtools:**
  - a swiss-army knife of tools for a wide-range of genomics analysis tasks
  - allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files and in many different formats (.bed, .bam, .gff, …)
  - https://bedtools.readthedocs.io/en/latest/
  - function for today: **getfasta**

Tools - The MEME Suite

- **The MEME Suite** is a powerful, integrated set of web-based tools for studying sequence motifs in proteins, DNA and RNA.

- **MEME-ChIP**
  - web service designed to analyze ChIP-seq ‘peak regions’ - short genomic regions surrounding declared ChIP-seq ‘peaks’
  - works also with **CLIP-seq** ‘peak regions’
  - Given a set of genomic regions, it performs:
    - ab initio motif discovery
    - motif enrichment analysis
    - motif visualization
    - binding affinity analysis
    - motif identification
  - https://meme-suite.org/meme/tools/meme-chip

Tools - The MEME Suite

- The MEME Suite is a powerful, integrated set of web-based tools for studying sequence motifs in proteins, DNA and RNA.

- **Tomtom**
  - web service that allows the user to **compare motifs** discovered by the suite, by other tools, or taken from the literature to all of the motifs in a selected database of motifs
  - aligns each input motif with each motif in the selected database and reports the most similar pairs, along with estimates of the statistical significance of each match
  - https://meme-suite.org/meme/tools/tomtom

Real dataset

1. Download the dataset: bed file with peaks, choose isogenic replicate 1,2 https://www.encodeproject.org/experiments/ENCSR570WLM/
2. Download the chromosome 1 fasta reference
3. Unzip the files

ENCODE Data Encyclopedia Materials & Methods Help Search...

Experiments / eCLIP / *Homo sapiens* / HepG2

Experiment summary for ENCSR570WLM

doi:10.17989/ENCSR570WLM

Summary		Attribution	
Status:	released	Lab:	Gene Yeo, U
Assay:	eCLIP	Award:	U54HG0070
Target:	QKI	Project:	ENCODE
Biosample summary:	*Homo sapiens* HepG2	External resources:	RBPImage:Q GEO:GSE918
Biosample Type:	cell line	References:	PMID:32252 PMCID:PMC doi:10.1038/ doi:10.1038/
Replication type:	isogenic		
Description:	eCLIP experiment on HepG2 against QKI		
Nucleic acid type:	RNA	Aliases:	gene-yeo:47
Size range:	175-300	Date submitted:	March 22, 20
Fragmentation methods:	see document	Date released:	April 26, 201
		Tags:	

ENCODE Data Encyclopedia Materials & Methods Help Search... Sign in / Create account

1	1	*Homo sapiens* HepG2 cell line	ENCBS362ASG	ENCAB494QSS	ENCLB238JYQ
2	1	*Homo sapiens* HepG2 cell line	ENCBS308ZPA	ENCAB494QSS	ENCLB436FYS

Files

Genome browser | Association graph | File details | Include deprecated files

GRCh38 | UCSC | Visualize | Download

Displaying 10 of 10 files

Lab custom hg19 (ENCAN522PDL) processed data (5 Files) archived

Lab custom GRCh38 (ENCAN767VIB) processed data (5 Files) released

Filter files

Clear all filters

File format

- bed narrowPeak 6
- bigBed narrowPeak 4

Output type

- peaks 10
- minus strand signal of unique reads 4
- plus strand signal of unique reads 4
- alignments 4
- reads 4

Replicates

- 1 4
- 2 4
- 1, 2 2

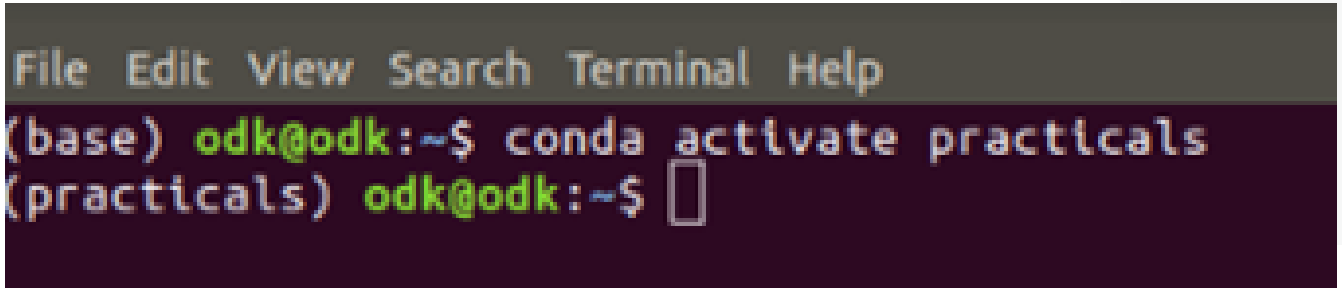
Accession	Default	File type	Output type	Isogenic replicate	Genome assembly	Date added	File size	File status
ENCFF815XNW	★	bed narrowPeak	peaks	2	GRCh38	2016-11-30	2.05 MB	released
ENCFF594IKL	★	bigBed narrowPeak	peaks	2	GRCh38	2016-12-03	3.29 MB	released
ENCFF704OCI		bed narrowPeak	peaks	1, 2	GRCh38	2018-12-03	214 kB	released
ENCFF551IJQ		bed narrowPeak	peaks	1	GRCh38	2016-11-30	2.76 MB	released
ENCFF984WOV		bigBed narrowPeak	peaks	1	GRCh38	2016-12-03	5.32 MB	released

Real dataset

4. Create and activate conda environment for today's practicals
   - Open the **terminal**

     ```
     conda create --name practicals
     conda activate practicals
     ```
5. Installation of necessary packages:
   - if it turns out you're missing a channel for installing some of the tool, you can add them by following cmd:

     ```
     conda config --add channels NAME
     conda install bedops
     conda install -c bioconda bedtools
     ```



Real dataset

6. Sort intervals in downloaded file and then extract chromosome 1 positions
   - `sort-bed PATH/TO/peaks.bed > PATH/TO/OUTPUT/sorted_peaks.bed`
7. Unify intervals length to 100 nt
   - `awk -F '\t' '{X=50; mid=(int($2)+int($3))/2;printf("%s\t%d\t%d\t%s\n",$1,(mid-X<0?0:mid-X),mid+X, $4);}' PATH/TO/chr1_peaks.bed > PATH/TO/OUTPUT/chr1_peaks_extended.bed`

chr5	132827787	132827811	QKI_HepG2_IDR
chr5	131548752	131548805	QKI_HepG2_IDR
chr2	241250904	241250940	QKI_HepG2_IDR
chr4	99202668	99202713	QKI_HepG2_IDR
chr4	99202713	99202762	QKI_HepG2_IDR
chr11	18505526	18505674	QKI_HepG2_IDR
chr8	118027137	118027182	QKI_HepG2_IDR
chr2	158492773	158492841	QKI_HepG2_IDR
chr2	64644932	64645037	QKI_HepG2_IDR
chr11	96158990	96159068	QKI_HepG2_IDR
chr20	8753903 8754001	QKI_HepG2_IDR	1000 +
chr6	2115465 2115530	QKI_HepG2_IDR	1000 -
chr13	108220327	108220437	QKI_HepG2_IDR
chr3	60693996	60694106	QKI_HepG2_IDR
chr3	149966520	149966623	QKI_HepG2_IDR

chr1	632859	632909	+
chr1	634491	634541	+
chr1	1047070	1047120	+
chr1	1047217	1047267	+
chr1	1338918	1338968	-
chr1	1613960	1614010	-
chr1	2404761	2404811	-
chr1	2405613	2405663	-
chr1	5890334	5890384	-
chr1	6212772	6212822	+
chr1	6212838	6212888	+
chr1	6457586	6457636	+
chr1	6790878	6790928	+
chr1	7708938	7708988	+
chr1	7752607	7752657	+
chr1	7755819	7755869	+
chr1	7755905	7755955	+
chr1	8016504	8016554	-

chr1	632834	632934	+
chr1	634466	634566	+
chr1	1047045	1047145	+
chr1	1047192	1047292	+
chr1	1338893	1338993	-
chr1	1613935	1614035	-
chr1	2404736	2404836	-
chr1	2405588	2405688	-
chr1	5890309	5890409	-
chr1	6212747	6212847	+
chr1	6212813	6212913	+
chr1	6457561	6457661	+
chr1	6790853	6790953	+
chr1	7708913	7709013	+
chr1	7752582	7752682	+
chr1	7755794	7755894	+
chr1	7755880	7755980	+
chr1	8016479	8016579	-
chr1	8016533	8016633	-

Real dataset

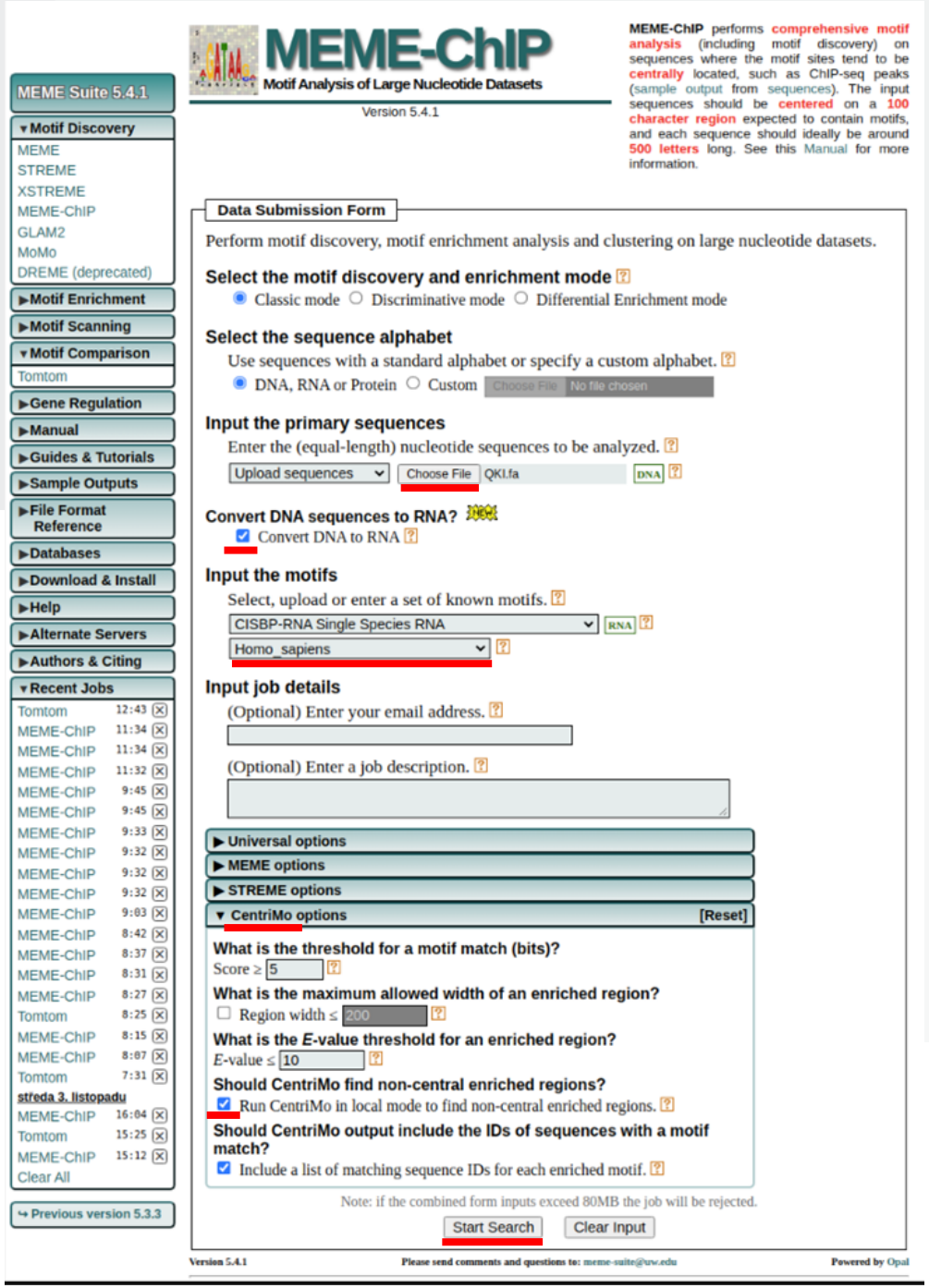
8. Extract sequences from a reference FASTA file for each of the intervals

```
bedtools getfasta -s -fi PATH/TO/chr1.fasta -bed PATH/TO/chr1_peaks_extended.bed -fo
PATH/TO/QKI_chr1.fa
```

```
>chr1:632834-632934()
GCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTGTACGCCCTTTTCCTAACACTCACAACAAAACTAACTAATACTAACATCTCAGACGCTCAGG
>chr1:634466-634566()
TAGCCATGTGATTTCACTTCCACTCCACAACCCTCCTCATACTAGGCCTACTAACCAACACACTAACCATATACCAATGATGGCGCGATGTAACACGAGA
>chr1:1047045-1047145()
GGGGGTTATGGTCTTGGGACTCGGCCCCCTCAAACATGTGCGTGCCGGGGACCCCACGCCTAACCCGTCTCTCTCGTTGCAAGCCGGTGTGGCACACTGC
>chr1:1047192-1047292()
CCACTAACCTCATGACCATCTGACTAACATCCACCTTCCCTTGCACCCTTGTGGCTTGCTGCTGGGGCCTGTGCCTGGGCCAGCCTGGATGCCAGGCAGA
>chr1:1338893-1338993()
ACTGGGCTGACACCCCACCCTGCAGACCAGGAAGTAATGAGAACAGGGCAGGCCCCTTCCCCTCCCCGCATGCCCCACCCGAGAGCGCAGGCTGTTAGTC
>chr1:1613935-1614035()
TTTGAGCCTTTGGAAAACGGTATCGTTAGGCATGTGGCGAAAACGTTGGGGTACTTGAAAAAAAGGCTGGCCATGGGTTAGTAAAAAGCTAGATATGTGA
>chr1:2404736-2404836()
ATGTGGCACACGCCCTCGAGGCATTTTAACACTGCGCTTCAGGAAATCTCAAGTTCCATCTTGTGTTAGTAACGTACCCACATTTTGCTGGAGTTAGTTT
>chr1:2405588-2405688()
AAAGCGCAGCCAGGGACAGCTTTCTGTTCTCTCCCAGGGTGGCTAGGTTAGTATCTTACATGACAAAAAACTGAGAGTGTTCTAACTTCTGTGCAAGCAA
>chr1:5890309-5890409()
CCCTTCATACAATGGAGAAGGCTTGGGAAGAATTCCAGGGAAGACGAGTGAAAGAATCCATGGATTTAGGTTTTAGTATACAAGGAGAATGGAAAAGGAC
>chr1:6212747-6212847()
GCTGCCGAGTGAACCCTCTGTCCCTGAGCTAACCCACATACTAGCAGAGGAGGAAGTCAGAGTCGGCCACTAACCAGATGCAAATCCCCACACTCTTCCC
>chr1:6212813-6212913()
CCACTAACCAGATGCAAATCCCCACACTCTTCCCCTTAGCGCTTGACCGTGCCTCCCAGCTGCTAACTGGCCTCAAATGATGCATGTGAGGTCAGGATTC
>chr1:6457561-6457661()
CCCTGCCTCCTATTAACCTGGCCTTTTCTACCCTTCAGTTAACCTAACCCCACTATCAATCACCTTGATTGTCTGGCCCTCAGAATGTACTTTCTGCCCC
>chr1:6790853-6790953()
CAATTTGAAATACCCCTTTTCTTTTTTCCTCTATTAAATTAGATTTACCATCTCCACAACGTATATAGAAACCAATTCTGCTACTATTTCACTCTTGTGA
>chr1:7708913-7709013()
TATCAACTACTAAAAATTAATCATTCTCTCCATTTTTTCAGCTTTCGTGTTTCACCTGACTTTCACCACCCCATACATCATGTTTCACTCTCCAGCTGGC
```

Real dataset

9. Open the **MEME Suite** web
10. Open the **MEME-ChIP** tool
11. Pick appropriate setup
12. **Run** the analysis



Homework

- **Re-do the motif analysis on the artificial dataset**
- 4 different datasets (1 dataset per student) + 1 bonus dataset
  - will be sent by email
- **Task:**
  - download the data
  - extend the intervals to 100 nt
  - extract sequences for the intervals
  - use MEME-ChIP to analyse motifs in dataset
  - try to identify domain/protein/protein family

  (look also at the CISBP database and pfam database - by clicking through the results)
- **Bonus task 1:**
  - Download the Motifs in MEME Text Format, upload the file to Tomtom tool, choose the CISBP-RNA Single Species RNA (Homo Sapiens) motif database and look at the results of the motif comparison tool
- **Bonus task 2:**
  - Repeat the analysis on the bonus (voluntary) dataset
- **We'll discuss the results on the practicals 3. 12. 2021**