MED5018: Introduction to biomedical Python programming

Final Project

Final project

Requirements:

- Choose a topic (from suggested topics or think of your own), analyze the data, visualize your results and describe your findings. **Use packages learned from this course in your analysis**
- **Include >= 3 different types of plots**
- Upload your code (.py or .ipyn), results and other related files onto **Github**. Describe your project in the README file
- Report:
  - Due **Week 18** (2025.6.20)
  - Includes Introduction, Methods (includes link to your github page), Figures, Results and Discussion
  - 3-4 pages in Chinese/English

Final project

Evaluation:

Report	60	
Formatting		10
Analysis & visualization		30
Results & discussion		20
Code	40	
github page		10
code modularization		20
code readibility		10

Grading

- Attendance: 20%
- Assignments: 40%
- Final report: 40%

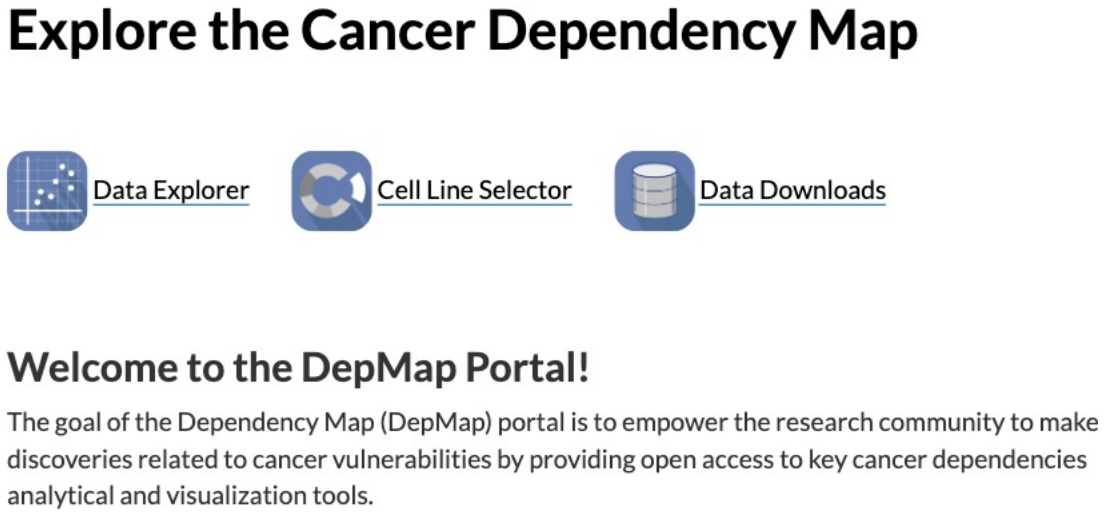
Topic 1: epidemiology

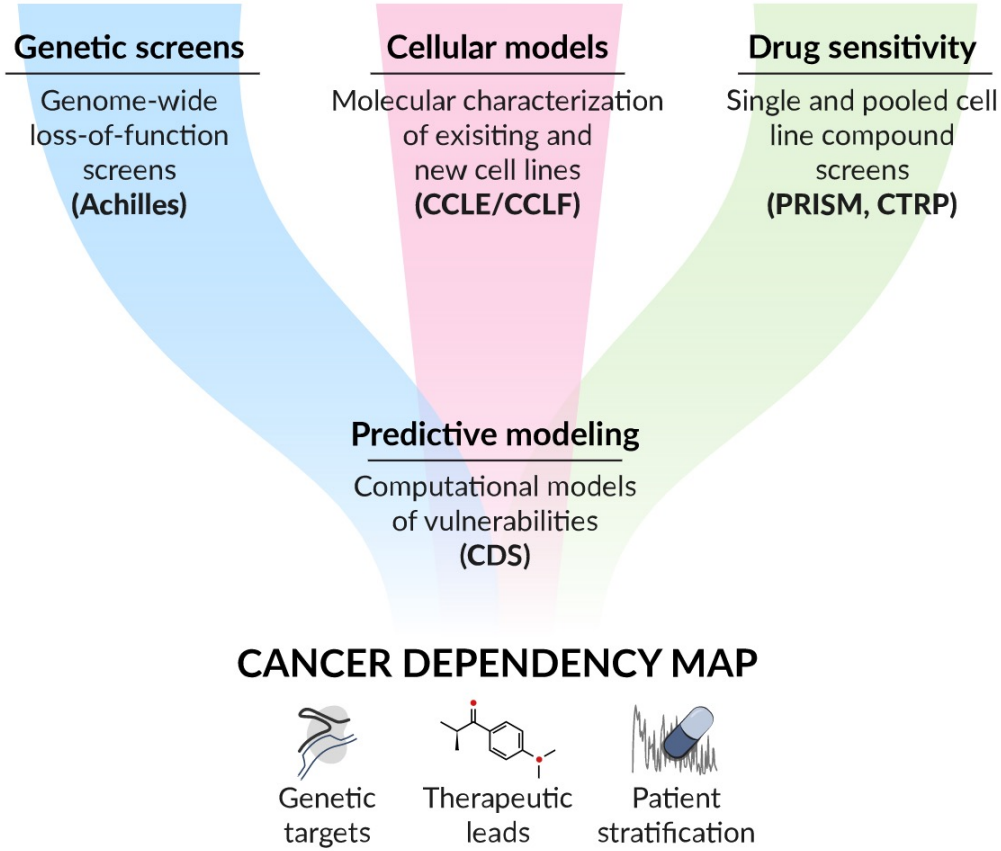
- Datasets:
- https://www.kaggle.com/datasets/belayethossainds/cancer-and-deaths-dataset-19902019-globally/data
- https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/data
- https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
- https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world
- https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset

Topic 2: Cancer

- Dataset: Depmap

https://depmap.org/portal/





https://depmap.org/portal/download/all/

Topic 2: Cancer

- Dataset: Depmap

sample_info.csv

Metadata for all of DepMap's cancer models/cell lines.

DepMap_ID	cell_line_nar	stripped_cel	CCLE_Name	alias	COSMICID	sex	source	RRID	WTSI_Maste	sample_coll
ACH-00001	SLR 21	SLR21	SLR21_KIDNEY				Academic la	CVCL_V607		kidney
ACH-00003	MHH-CALL-	MHHCALL3	MHHCALL3_HAEMATOPOIETIC_AND_			Female	DSMZ	CVCL_0089		bone_marr
ACH-00003	NCI-H1819	NCIH1819	NCIH1819_LUNG			Female	Academic la	CVCL_1497		lymph_nod
ACH-00004	Hs 895.T	HS895T	HS895T_FIBROBLAST			Female	ATCC	CVCL_0993		fibroblast
ACH-00004	HEK TE	HEKTE	HEKTE_KIDNEY				Academic la	CVCL_WS59		kidney
ACH-00005	TE 617.T	TE617T	TE617T_SOFT_TISSUE			Female	ATCC	CVCL_1755		soft_tissue
ACH-00006	SALE	SALE	SALE_LUNG			Male	Academic la	CVCL_WS60		lung
ACH-00006	REC-1	REC1	REC1_HAEMATOPOIETIC_AND_LYMP			Male	DSMZ	CVCL_1884		lymph_nod
ACH-000071		HS706T	HS706T_BONE			Female	ATCC	CVCL_0863		fibroblast
ACH-00007	NCO2	NCO2	NCO2_HAEMATOPOIETIC_AND_LYMI			Female	HSRRB	CVCL_3043		haematopo
ACH-00007	MJ	MJ	MJ_HAEMATOPOIETIC_AND_LYMPH			Male	ATCC	CVCL_1414		haematopo
ACH-00007	TE 125.T	TE125T	TE125T_FIBROBLAST			Female	ATCC	CVCL_1740		fibroblast
ACH-000084		MUTZ3	MUTZ3_HAE	MUTZ3		Male	DSMZ	CVCL_1433		haematopo
ACH-00008	NCI-H684	NCIH684	NCIH684_LARGE_INTESTINE			Male	KCLB	CVCL_9980		liver
ACH-00009	Panc 05.04	PANC0504	PANC0504_PANCREAS			Female	ATCC	CVCL_1637		pancreas
ACH-00011	NCC-StC-K	NCCSTCK14	NCCSTCK140_STOMACH			Female	RIKEN	CVCL_3055		stomach
ACH-00011	Hs 863.T	HS863T	HS863T_FIBROBLAST			Female	ATCC	CVCL_0959		fibroblast

•DepMap_ID: Static primary key assigned by DepMap to each cell line
•cell_line_name: Original cell line name, including punctuation
•sample_collection_site: Tissue collection site
•primary_or_metastasis: Indicates whether tissue sample is from primary or metastatic site
•primary_disease: General cancer lineage category
•…

Topic 2: Cancer

- Dataset: Depmap

CCLE_expression.csv

Gene expression TPM values of the protein coding genes for DepMap cell lines. Values are inferred from RNA-seq data using the RSEM tool and are reported after log2 transformation, using a pseudo-count of 1; log2(TPM+1).

Genes

Cell lines

	TSPAN6 (71	TNMD (641(	DPM1 (8813	SCYL3 (5714	C1orf112 (5	FGR (2268)	CFH (3075)	F
ACH-001113	4.33199178	0	7.36439734	2.79285535	4.47053687	0.02856915	1.22650853	3
ACH-001289	4.56681515	0.5849625	7.10653677	2.54349588	3.50462039	0	0.18903382	3
ACH-001339	3.15055968	0	7.37903173	2.33342373	4.22727899	0.05658353	1.31034012	6
ACH-001538	5.08533967	0	7.15410924	2.54596837	3.08406426	0	5.86814348	6
ACH-000242	6.72914486	0	6.53760669	2.45680615	3.86789646	0.79908731	7.20838069	5
ACH-000708	4.27202319	0.18903382	7.02292259	2.55581616	3.84197312	0	0.0976108	4
ACH-000327	3.33771109	0	5.92718536	1.94485845	2.67807191	0.01435529	3.08915913	6
ACH-000233	0.05658353	0	6.09360243	3.97085365	3.73118324	0.02856915	6.09296851	3
ACH-000461	4.0161397	0	6.53387478	2.22650853	3.02147973	0.02856915	0.08406426	5
ACH-000705	4.41142625	0	6.41244282	2.36457243	4.27500705	0.04264434	0.20163386	2
ACH-001794	3.87184365	0.05658353	6.76394266	1.93734439	3.15218342	0.62293035	6.94462422	7
ACH-002023	5.26491169	0	6.75688986	2.28392177	3.81352469	0	4.0591822	6
ACH-000528	4.51222689	0	7.09982118	2.84398384	4.67242534	0.01435529	0.81557543	6
ACH-001655	3.592158	0	6.74711882	0.92599942	1.83995959	0.02856915	0.05658353	5
ACH-000167	0.04264434	0	6.71011763	2.35332329	3.98550043	5.85972113	0.27500705	3
ACH-000792	3.2794713	0	6.39025496	1.74846123	3.43562859	0.08406426	0.422233	6
ACH-001098	5.90929309	0	6.78398041	3.29278175	3.08746284	0	4.89820835	5
ACH-000570	5.26491169	0	6.83048352	2.83592407	4.17791779	0.17632277	5.84046323	6
ACH-000351	3.97269265	0	8.16133347	2.35049725	4.0976108	1.60880924	0.3448285	
ACH-000769	4.7131459	0	6.19377174	2.4409522	4.17632277	0	0.11103131	5

Topic 2: Cancer

- Dataset: Depmap

CRISPR_gene_effect.csv

Gene KO

Cell lines

DepMap_ID	A1BG (1)	A1CF (2997	A2M (2)	A2ML1 (144	A3GALT2 (1	A4GALT (53	A4GNT (511
ACH-00000	-0.1348083	0.05976414	-0.0086653	-0.0035722	-0.1062113	-0.0082569	0.01871112
ACH-00000	0.08185267	-0.0564005	-0.1067377	-0.0144985	0.07820912	-0.1375616	0.16865681
ACH-00000	-0.094196	-0.0145984	0.10042603	0.16910279	0.03236276	-0.1480495	0.16893121
ACH-00000	-0.011544	-0.1231889	0.08069221	0.06104554	-0.0134537	-0.0169221	-0.0294744
ACH-00000	-0.0507823	-0.0374662	0.06888547	0.090375	0.01263396	-0.079339	-0.0178085
ACH-00001	0.09176169	-0.0246847	0.03825127	0.20230519	-0.0895683	-0.2382661	-0.0422404
ACH-00001	-0.1467414	0.02739757	0.17800132	0.23979694	0.08683216	-0.2281534	0.18666851
ACH-00001	-0.0592493	-0.0900571	0.03989619	0.1120302	-0.1006963	-0.0871284	-0.0291913
ACH-00001	-0.0347549	-0.0984152	0.05022526	0.04611864	-0.0390687	-0.1265427	-0.3360773
ACH-00001	-0.2037093	-0.0047761	-0.0662448	0.01511034	-0.0078323	-0.0217672	0.09482096
ACH-00001	0.02261886	-0.0446184	0.02289812	0.0156794	-0.2951142	-0.1182438	0.06740902
ACH-00001	-0.1408936	-0.0565507	-0.1755851	0.07845473	0.04436803	0.04486578	0.06807933
ACH-00001	0.01936719	-0.0090009	0.12312889	0.07834327	-0.0485334	0.06822092	-0.0300908
ACH-00002	-0.1282732	-0.0617642	-0.0235949	-0.030921	0.01273444	-0.1392151	-0.0311323
ACH-00002	-0.0616031	0.03907231	0.11447683	-0.0772903	-0.1159471	-0.035359	-0.0070858
ACH-00002	-0.0840977	-0.1266262	-0.0234172	-0.0602367	-0.071407	0.01246419	0.0518981
ACH-00002	-0.0416814	0.01019081	0.01113129	-0.0016377	0.08894099	-0.1256668	-0.0146221
ACH-00002	-0.1369808	-0.1449458	-0.0124462	0.11648105	0.01680668	-0.068883	-0.1323596
ACH-00002	-0.1892002	-0.0789077	0.10356003	0.12831619	-0.0773874	-0.0108697	-0.0928028
ACH-00002	-0.0290174	0.00135555	0.06048758	0.10720116	-0.1944451	-0.067306	-0.0322075

Gene Effect scores derived from CRISPR knockout screens published by Broad's Achilles and Sanger's SCORE projects.

Negative scores imply cell growth inhibition and/or death following gene knockout. Scores are normalized such that nonessential genes have a median score of 0 and independently identified common essentials have a median score of -1.

Topic 2: Cancer

- Dataset: Depmap

Drug_sensitivity.csv

Drug sensitivity; logfold change values relative to DMSO. The more negative the value, the more sensitive.

Drug

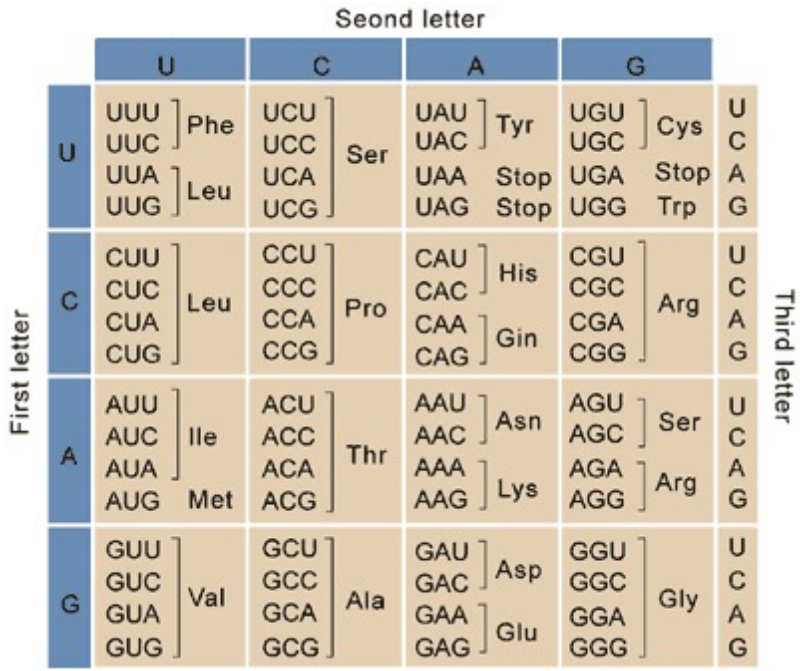
Cell lines

	RS-0481 (BF	oleuropein (	isoleucine (E	gepefrine (B	cloranolol (E
ACH-00132	-0.442835	-0.143624	-0.075496	0.1906309	-0.011447
ACH-00131	-0.039683	0.2882696	-0.208114	0.045799	0.5106338
ACH-00130	0.2723067	0.2741152	-0.209725	-0.149599	-0.126682
ACH-00130	-0.163176	-0.174646	-0.307695	-0.117653	0.1360773
ACH-00123	-0.433834	-0.134896	0.0503954	0.0247503	-0.000615
ACH-00121	0.0817448	-0.486275	-0.118985	0.8314202	-0.51631
ACH-00121	0.3527695	0.4903075	-0.107439	-0.075238	0.3715598
ACH-00121	0.3791271	-0.744636	0.6659031	-0.181121	0.3089592
ACH-00120	0.423955	-0.012615	-0.0538	-0.354739	-0.187207
ACH-00120	0.5354427	0.0293117	0.181144	-0.186969	0.2006027
ACH-001193		-0.159399			0.0335186
ACH-00119	0.3582464	0.3445328	0.0361066	0.1978186	0.159737
ACH-00119	-0.416401	0.6704201	-0.580173	0.2308938	-0.125217

Topic 3: Codon usage bias

- **Background:**

- synonymous codons: Different codons that encode the same amino acid



Topic 3: Codon usage bias

- **Background:**

- synonymous codons: Different codons that encode the same amino acid

- Codon usage bias: preferential or non-random use of synonymous codons
  - Different species have consistent and characteristic codon biases.

  - A critical factor determining gene expression and cellular function by influencing diverse processes such as RNA processing, protein translation and protein folding.

Topic 3: Codon usage bias

- Goal:

To determine codon usage bias of different species (human, mouse, yeast, Drosophila, C. elegans, Arabidopsis, …); identify patterns (similarities vs. differences, …)

Escherichia coli: (Genetic code: Standard)

Triplet	Amino acid	Fraction	Frequency/ Thousand	Number	Triplet	Amino acid	Fraction	Frequency/ Thousand	Number
TTT	F	0.58	22.1	80995	TCT	S	0.17	10.4	38027
TTC	F	0.42	16.0	58774	TCC	S	0.15	9.1	33430
TTA	L	0.14	14.3	52382	TCA	S	0.14	8.9	32715
TTG	L	0.13	13.0	47500	TCG	S	0.14	8.5	31146

Human: (Genetic code: Standard)

Triplet	Amino acid	Fraction	Frequency/ Thousand	Number	Triplet	Amino acid	Fraction	Frequency/ Thousand	Number
TTT	F	0.45	16.9	336562	TCT	S	0.18	14.6	291040
TTC	F	0.55	20.4	406571	TCC	S	0.22	17.4	346943
TTA	L	0.07	7.2	143715	TCA	S	0.15	11.7	233110
TTG	L	0.13	12.6	249879	TCG	S	0.06	4.5	89429

Topic 4: Ramachandran plot

- https://zmjo02e9v0.feishu.cn/wiki/VtXIw9oI3iaQAikjBuncMojsnig?from=from_copylink

Topic 5: Deep mutational scanning

- https://zmjo02e9v0.feishu.cn/wiki/VGQYwA80NiNcItkYVSUc11WSnac?from=from_copylink