СОФИЙСКИ УНИВЕРСИТЕТ
„СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО
МАТЕМАТИКА И ИНФОРМАТИКА

SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI
FACULTY OF
MATHEMATICS AND INFORMATICS

КУРСОВ ПРОЕКТ

ДИСЦИПЛИНА: ИЗЧИСЛИТЕЛНА БИОЛОГИЯ

ТЕМА:

ДИФЕРЕНЦИАЛНА ЕКСПРЕСИЯ

Изготвил:

Виктория Ангелова Иванова

Магистърска програма: Био- и медицинска информатика

Факултетен номер: 7MI3400799

Ръководител:

гл.ас. д-р Ирена Авджиева

София

2026

Съдържание

Списък на фигурите и таблиците

1. Въведение

Експресионният анализ е метод за изследване на нивото на експресия на гени или белтъци, в дадена клетка, тъкан или организъм. Показва кои гени са „включени“ или „изключени“ и в каква степен. Анализира се количеството иРНК, което е индикатор за генна активност. Често се използва при изследване на ракови заболявания, сравнение между здрави и болни тъкани и проследяване на ефекта от лекарства.

А какво всъщност е **генна експресия**? – това е процесът, чрез който информацията, кодирана в даден ген, се използва за синтез на функционален продукт (РНК или протеин), като количеството на този продукт отразява активността на гена в дадена клетка. Накратко казано, експресията е мярка за активност и това колко усилено даден ген участва в процесите транскрипция и транслация.

Диференциална експресия означава разлика в нивото на експресия на гени (или белтъци) между две или повече състояния. Тя показва кои гени са по-активни или по-слабо активни: между здрави и болни клетки или преди и след лечение.

Експресионните данни са числови данни, които показват колко активно се „изразява“ (експресира) даден ген в определена клетка, тъкан или при дадено условие.

Целта на настоящия курсов проект е да използва едно от най-широките приложения на експресионния анализ, а именно да разгледа диференциалната експресия на гените EGFR и KRAS в туморни спрямо здрави клетки, използвайки *RNA-seq* и *microarray* изследвания от публично достъпните бази от данни Expression Atlas и Gene Expression Omnibus. Също така, проучването ще оцени наличието на връзка между експресията на двата гена в различни туморни състояния.

Гените EGFR и KRAS които ще разгледаме, са едни от най-широко изследваните онкогени, свързани с множество видове рак, като колоректален и белодробен карцином. Промените в експресията на EGFR и KRAS ще бъдат анализирани чрез сравнение между туморни и здрави проби и визуализирани с помощта на диаграми, като *boxplot*, *violin plot, scatter plot* и др.

Основните работни хипотези са, че EGFR показва повишена експресия в определени тумори, че експресията на KRAS варира в зависимост от типа рак и че между експресията на двата гена може да се наблюдава корелация, отразяваща общото им участие в определени процеси.

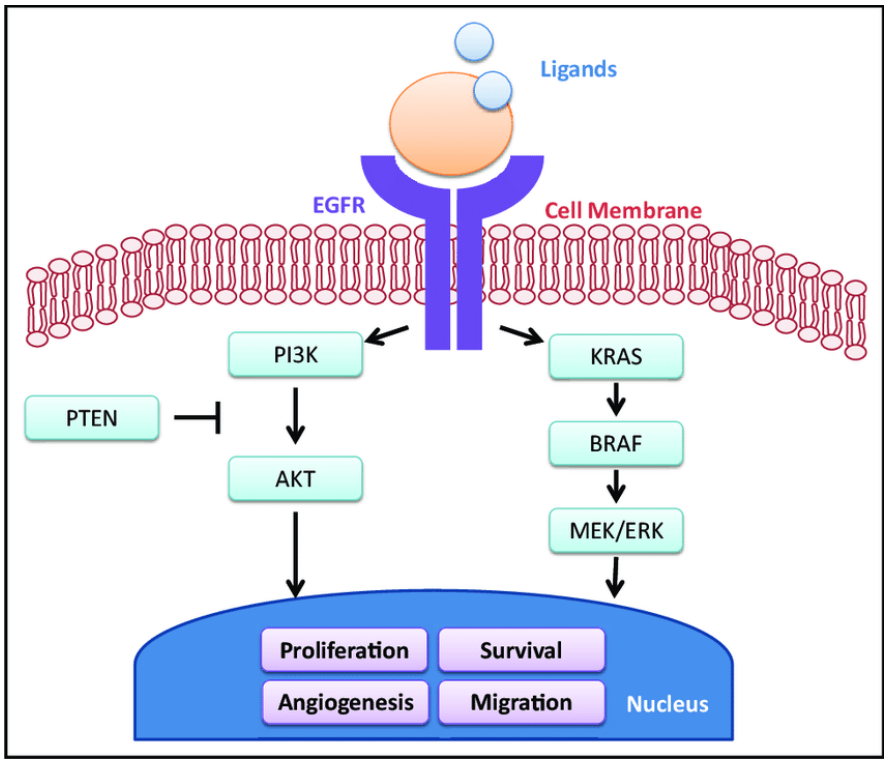
В изследванията проведени в този проект ще използваме широко познати методи за установяване на **статистическа зависимост**, като коефициент на корелация за оценка на връзката между експресията на двата гена, t-test за сравнението на експресионните нива между туморни и здрави проби, както и p-value за оценка на статистическата значимост на наблюдаваните разлики.

EGFR gene - epidermal growth factor receptor
EGFR е ген, който кодира белтък, намиращ се на повърхността на клетката. Този белтък действа като рецептор и когато към него се свърже растежен фактор (EGF), той изпраща сигнал навътре в клетката, че е време тя да расте и да се дели. Така се задействат процеси, които водят до клетъчно

делене. Когато EGFR е мутирал или прекалено активен, клетките могат да започнат да се делят неконтролирано, което е характерно за някои видове рак, особено рак на белия дроб.

KRAS gene - Kirsten rat sarcoma virus oncogene homologue
KRAS е ген, който кодира белтък (K-Ras), участващ в предаването на сигнали вътре в клетката. Този белтък действа като предавател, който включва или изключва сигналите за клетъчен растеж. При нормални условия той се активира за кратко време, когато клетката получи сигнал за делене. При определени мутации обаче KRAS остава постоянно активен, което води до непрекъснато стимулиране на клетъчното делене. Такива активиращи мутации са свързани с различни видове рак, включително рак на белия дроб, панкреаса и дебелото черво.

Тъй като продуктите на двата гена функционират в една и съща система, работата им е тясно свързана, участвайки заедно в клетъчен растеж и делене.



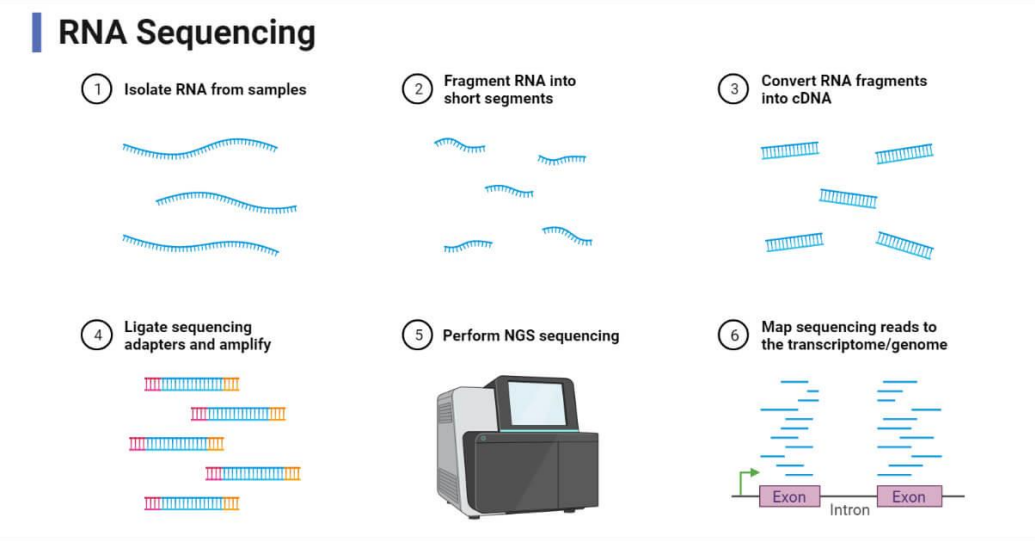
Фиг. 1 Сигнален път на EGFR и KRAS гените

2. Материали и методи

2.1. Изследвания

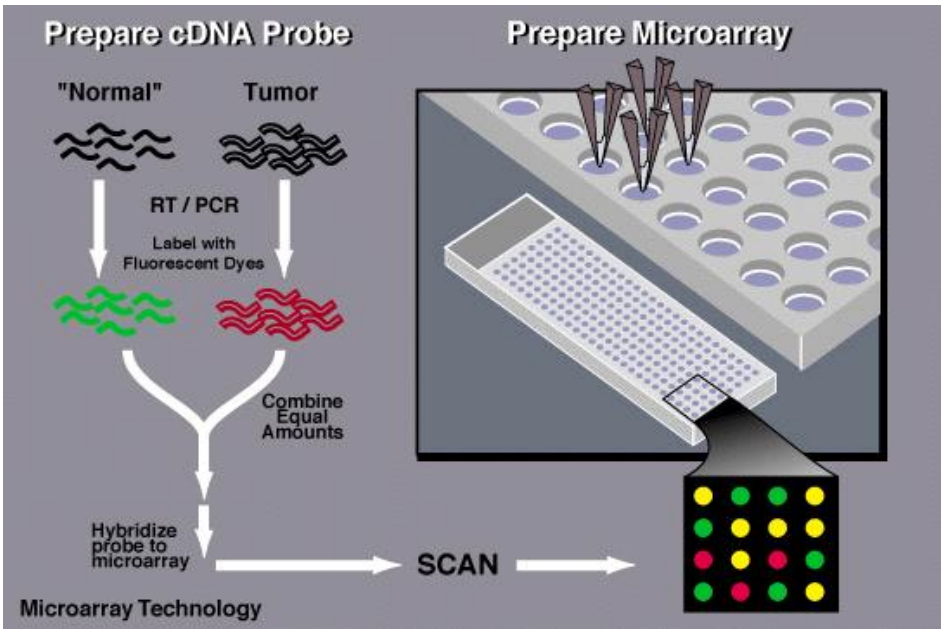
По време на набирането на подходящи изследвания, бяха разгледани два основни метода за анализ на генната експресия:

RNA-seq – работи чрез секвениране на всички РНК молекули в пробата, които първоначално се превръщат в комплементарна ДНК (cDNA). Получените последователности се подреждат спрямо референтен геном, което позволява точно измерване на експресията на всеки ген.



Фиг. 2 Метод на работа на RNA-seq

Microarray – технологията използва фиксирани върху чип сонди, специфични за отделни гени, към които се свързват маркирани РНК или cDNA молекули от пробата. Интензитетът на сигнала от всяка сонда отразява нивото на експресия на съответния ген.



Фиг. 3 Метод на работа на microarray

В настоящото изследване ще използваме microarray, тъй като за този метод са налични множество добре анотирани публични набори от данни, които позволяват gupekmna обработка и по-лесен статистически анализ.

2.2. Стратегия за търсене

Стратегията за търсене включва използване на ключови думи като *normal*, *tumor*, *lung cancer*, *colorectal cancer* и *Homo sapiens* в базата данни Gene Expression Omnibus (GEO), с цел намиране на експресионни изследвания, които сравняват туморни и здрави човешки тъкани.

Идентификаторите на гените са намерени чрез базата данни NCBI Gene, като за всеки ген е проверена официалната му номенклатура и съответните Gene ID в Affymetrix Human Genome U133 Plus 2.0 Array анотациите.

Тези ID-та са използвани за коректно разпознаване и извличане на експресионните данни от избраните изследвания.

Избраните изследвания обхващат два основни вида рак, за които е потвърдено, че EGFR и KRAS имат променена експресия. Първото изследване - **GSE19188**, съдържа данни за експресия от ранни стадии на 160 проби от рак на белия дроб**,** както на туморни, така и на нормални тъкани. Второто - **GSE41258** включва данни за колоректален карцином, които обхващат туморни проби, метастази и съответстващи нормални тъкани. Този набор от данни е подходящ, тъй като дава възможност за директно сравнение на експресията на EGFR и KRAS между туморни и здрави проби.

2.3. Данни

NCBI GEO предоставя данни получени чрез микрочипов анализ, като улеснява съпоставката им и предоставя готови нормализирани матрици за диференциален анализ, чиято експресия е логаритмувана предвид голямата разлика в експресията на различните проби.

Данните преди извличането на необходимата информация за текущото изследване, представляват матрици от десетки хиляди гени, извлечени от определен брой пациенти, като всяка клетка представлява логаритмуваната експресия на един ген при един пациент.

	"ID_REF"	"GSM475656"	"GSM475657"	"GSM475658"	"GSM475659"	"GSM475660"	"GSM475661"	"GSM475662"	"GSM475663"	"GSM475664"	"GSM475665"	"GSM475666"	"GSM475667"	"GSM475668"
70	"1007_s_at"	0.278536376	-0.005101812	-0.619296868	-0.400919703	-0.025467569	-1.277200911	0.034773135	-0.657267127	1.192051644	-0.303082114	-0.74629		
71	"1053_at"	0.460966633	-0.800087059	-0.233028863	-0.474353185	-0.57598891	0.839695813	0.938249303	-0.32025206	0.103448537	-0.420840454	-0.374284021	-0.4	
72	"117_at"	-0.272634196	-1.059394824	-0.12179781	-0.721780636	-0.320456921	-1.113349763	-0.24029401	-0.185089094	0.633424856	-0.300901277	-0.59698		
73	"121_at"	0.293016021	0.053426368	-0.604698307	-0.140630771	0.223255159	0.245966118	-0.154346181	-0.049041566	-0.634417436	0.179776162	0.528778891	-0.2	
74	"1255_g_at"	1.536637393	-0.141373795	-0.401689218	-0.166218067	-0.386273554	-0.204877773	-0.153164885	-0.456434392	0.105857448	-0.365263232	-0.1		
75	"1294_at"	-0.366586266	0.140577897	0.366296528	0.630699329	0.421735957	-1.533058124	-0.547019744	0.503521299	-0.490678221	0.294598056	0.332057225	0.378645	
76	"1316_at"	0.473544943	0.439606748	-0.526704993	-0.084876941	0.099755047	0.078781991	-0.269630021	-0.374086568	0.212797924	-0.36976087	-0.149348893	-0.3	
77	"1320_at"	-0.139187405	0.106126537	0.212234528	0.116829168	-0.130689531	-0.36702353	-0.297475176	-0.035722534	0.174848468	0.391292678	-0.081919184	-0.1	
78	"1405_i_at"	0.887567471	0.979195547	0.285527379	0.729472508	-0.180695375	-4.471468824	-1.134520817	-0.112860517	-0.23868128	0.171110845	1.287664825	-0.09168	
79	"1431_at"	-0.087691099	-0.154109411	0.005556152	0.087508571	-0.08822113	0.06158284	-0.079375038	-0.068212541	-0.362926084	-0.073953786	0.018130765		
80	"1438_at"	0.72586152	-0.470827568	-0.741927808	-0.713321132	-0.537121537	-0.427834953	0.28600451	-0.762674801	1.333339622	-0.287869297	0.188498		
81	"1487_at"	-0.258574169	-0.305918156	-0.294888597	-0.119031555	0.123155491	0.295551576	-0.049476805	-0.482791703	-0.062683221	-0.347813317	0.25		
82	"1494_f_at"	0.382999874	0.167227148	-0.192475904	-0.181821661	0.056528378	-0.208202357	-0.300759494	-0.351374208	0.251137156	0.132484638	0.496635762	0.00	
83	"1552256_a_at"	1.51032466	-0.363014224	-0.255769462	-0.051572068	0.114575709	1.543210689	-1.351082885	-0.293280435	0.287297256	0.122989302	0.105854211		
84	"1552257_a_at"	0.295527162	-0.399725679	-0.422779496	-0.548885674	-0.273946573	1.123508451	-0.077855384	-0.569936436	1.692318983	-0.646747833	-0.3		
85	"1552258_at"	-0.0432266	-0.108063885	0.210893062	-0.20304578	-0.083412272	-0.349683345	-0.371037362	-0.029716636	-0.321643199	-0.227940609	-0.0		
86	"1552261_at"	0.912457942	-0.198521851	-0.226199776	-0.061331108	0.134513769	-0.220220116	-0.163241993	0.015070947	-0.072114725	0.27181316	0.451435		
87	"1552263_at"	-0.667870271	0.310488505	0.303034456	0.305928678	-0.438244674	-1.337607963	0.19436631	0.650084224	0.424121837	0.204307973	0.38614548	0.026700	
88	"1552264_a_at"	-1.161214134	-0.446460588	-0.074873059	0.108946111	-0.22083194	-0.365465478	0.032961577	0.046730568	-0.253776572	0.646621183	0.030160023		
89	"1552266_at"	-0.087506907	-0.155270168	-0.076889428	-0.15719945	-0.203344366	0.208318588	-0.147044811	-0.408018845	0.360410979	-0.128351488	-0.0		
90	"1552269_at"	-0.075820357	0.730724038	-0.118638846	-0.288418983	0.132820736	0.262764709	0.175647052	0.343987643	0.354382895	-0.381158753	-0.033449436		
91	"1552271_at"	0.12431423	-0.272252398	-0.368332124	-0.171621744	-0.017267701	0.03603483	-0.177790418	-0.488280483	0.578024761	0.020169599	0.194057		
92	"1552272_a_at"	-0.013780876	-0.074565112	-0.20111468	-0.217542674	0.157181025	-0.23051897	-0.078043399	-0.479709903	0.452680667	0.193379021	0.161892819		
93	"1552274_at"	-1.138924661	-0.449307949	0.767928351	-0.529689667	-0.270872631	2.248801886	-0.042931793	0.227170927	-0.401711315	0.090743209	-0.15656		
94	"1552275_s_at"	-1.423838841	-0.518601288	0.817213451	0.459875006	-0.13408496	2.453533737	0.332793597	0.391950186	-0.973432159	0.92547283	0.82405645	0.673276	
95	"1552276_a_at"	0.539265001	-0.174576768	0.088095085	0.192544963	0.441223197	-0.043489024	-0.573131917	-0.191081019	0.22887765	0.15687254	0.482045655	-0.3	

Фиг. 4 Данни, извлечени от GEO

Информацията е извлечена чрез предоставената от GEO библиотека GEOparse, за да може резултатите да бъдат лесно обработени и изобразени на програмният език Python.

2.4. Статистически подходи

t-test е метод от статистиката, който се използва за сравняване на средните стойности на две групи данни, за да се установи дали разликата между тях е статистически значима. В този проект

ще използваме t-test, за да сравним нивата на експресия на гените EGFR и KRAS между туморни и здрави проби и да проверим дали наблюдаваните разлики са реални, а не резултат от случайни вариации.

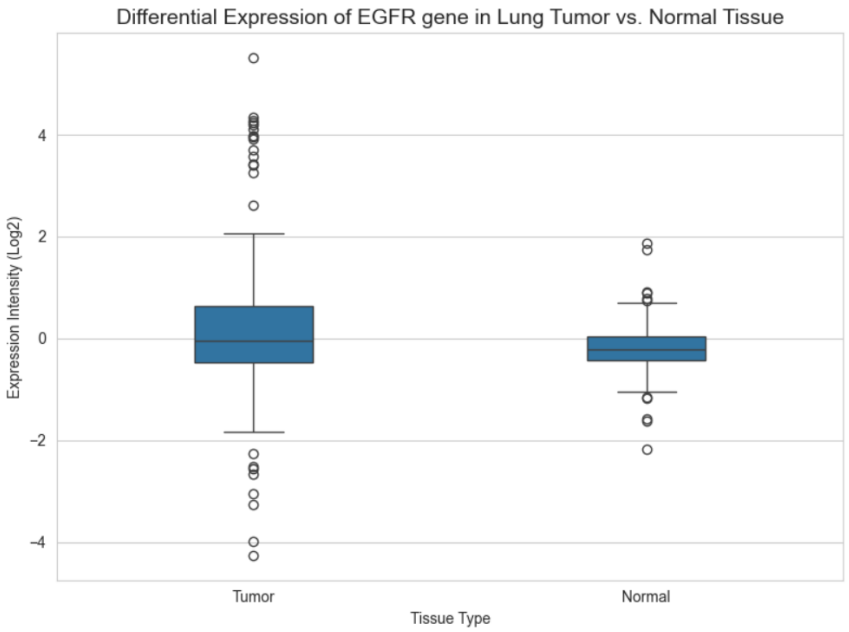
p-value - е методът чрез който установяваме значимостта на разликата. Стойността показва вероятността да се наблюдава такава разлика между двете групи, ако всъщност няма реална разлика между тях. Малка p-стойност ($p < 0.05$) означава, че е малко вероятно резултатът да е случаен и затова приемаме, че разликата в експресията е статистически значима.

3. Резултати и дискусия

Резултатите от анализа са извлечени чрез различни библиотеки на езика Python, като за целта са използвани графики и статистически разултати, описани изцяло в приложените Jupyter Notebook файлове *expression_lung_cancer.ipynb и expression_colorectal_cancer.ipynb.* В тази секция всяка графика е включена като фигура с описание на резултатите.

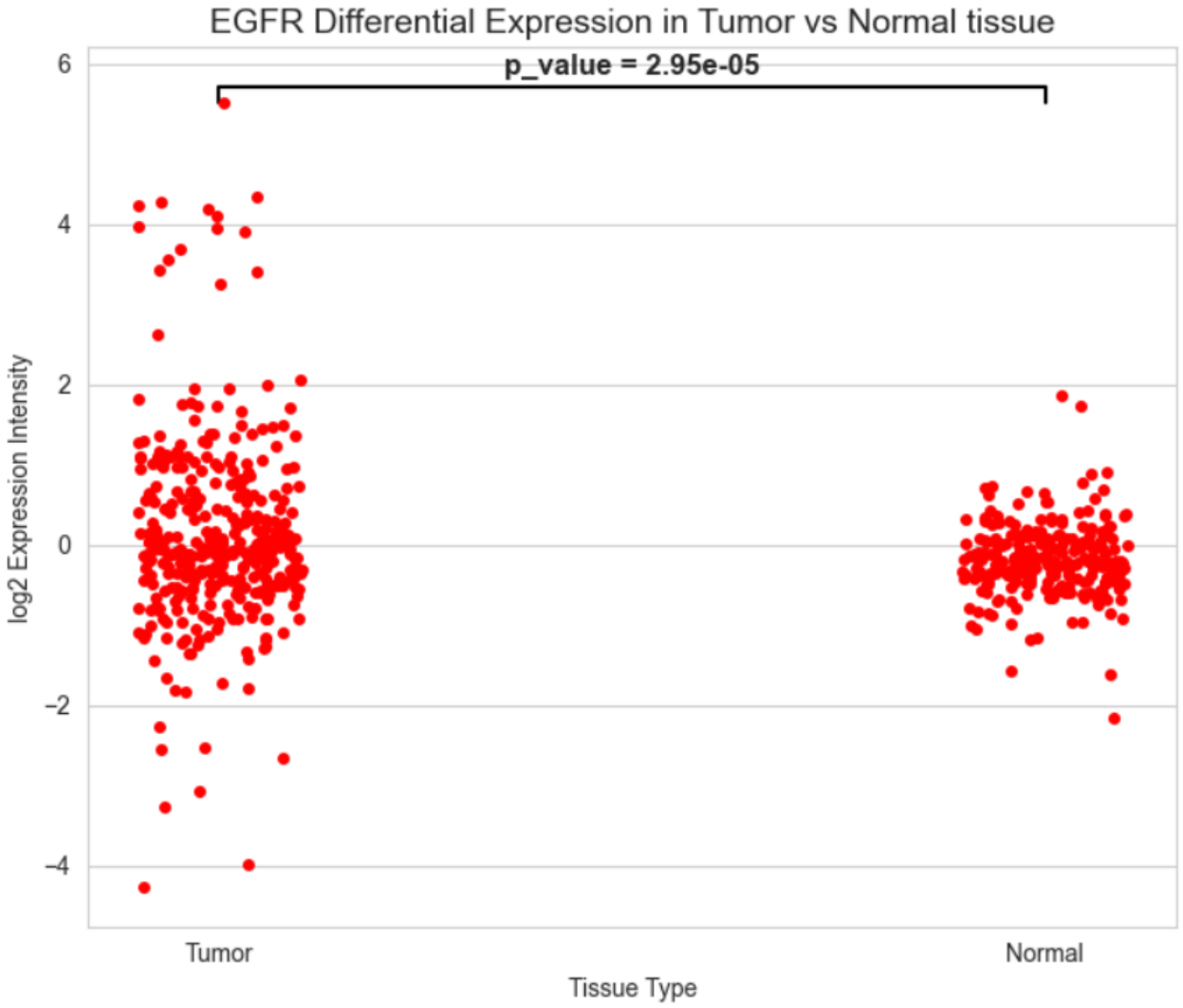
3.1. Рак на белия дроб

Графиката показва разпределението на нивата на експресия на гена EGFR в здрави и туморни проби от бял дроб. Наблюдава се по-висока експресия на EGFR в туморните проби в сравнение със здравите, което показва повишена експресия на гена при белодробен рак.



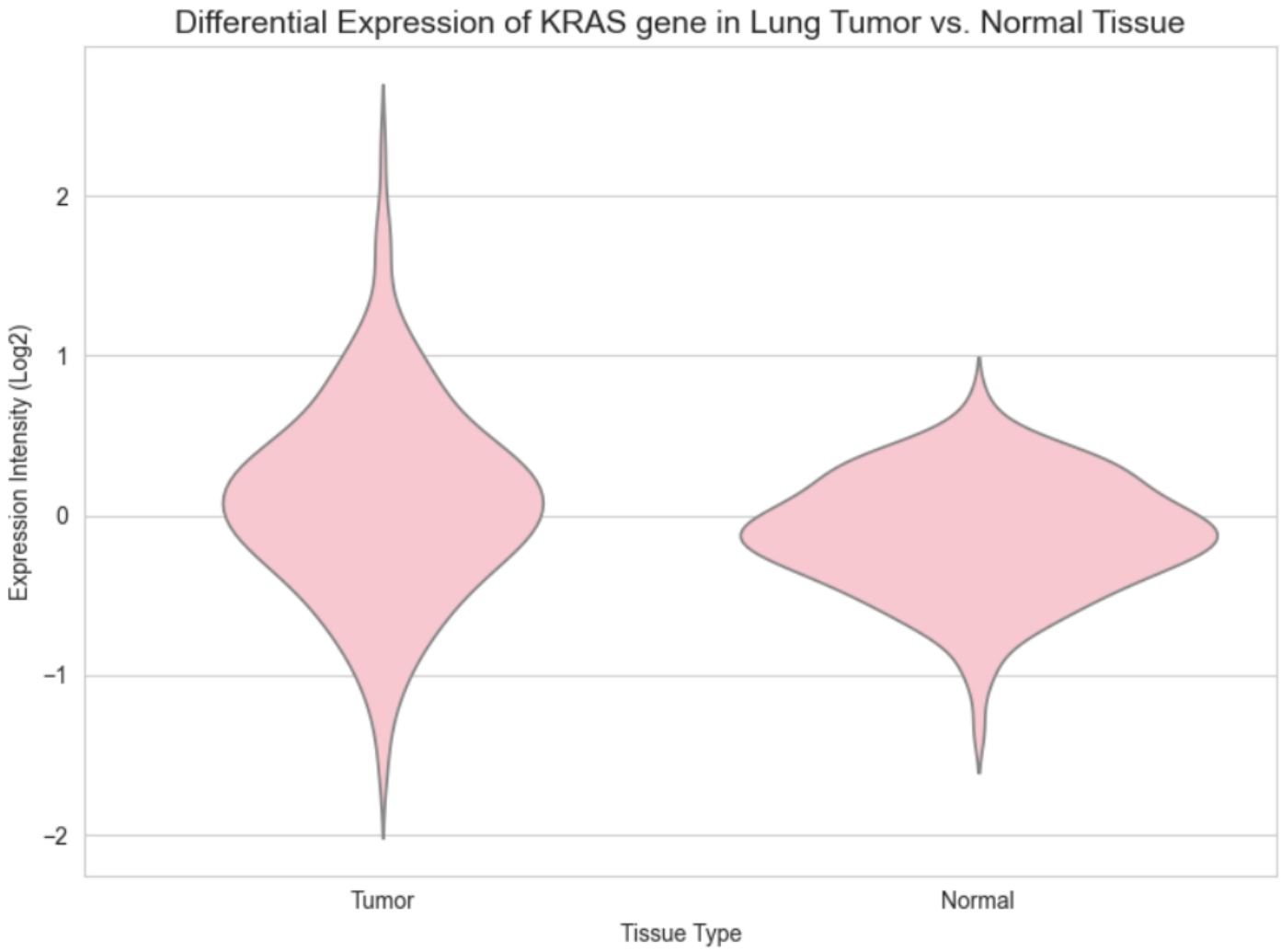
Фиг. 5 Boxplot на експресията на EGFR генът

След проведен t-test за разликата между двете групи, установяваме че тя е статистически значима ($p < 0.05$), което предполага, че увеличената експресия на EGFR е свързана с туморното състояние и вероятно има роля в развитието на заболяването.



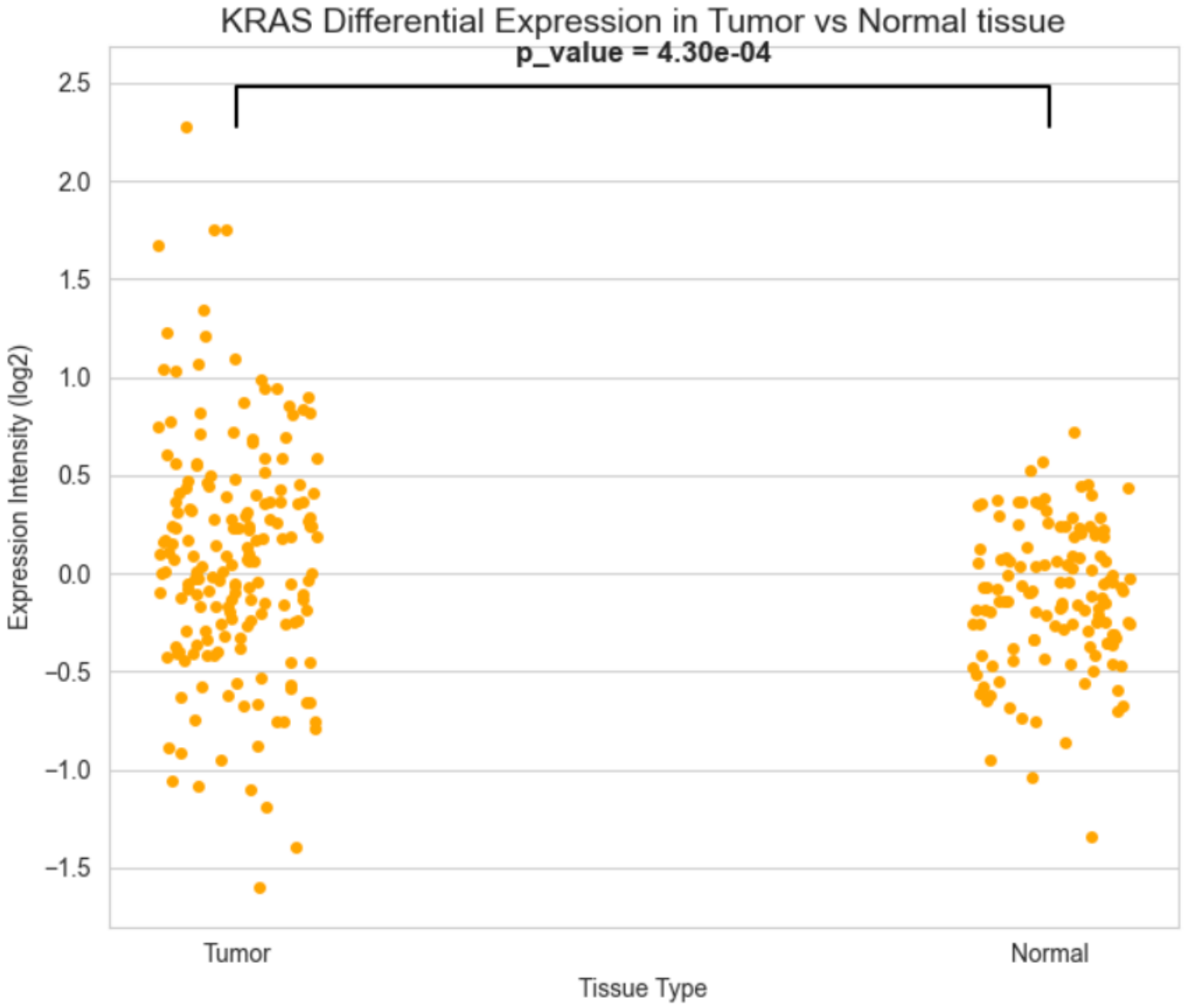
Фиг. 6 Статистически анализ на EGFR генът

Violin plot-ът илюстрира разпределението и плътността на експресионните нива на гена KRAS в здрави и туморни белодробни проби. Разпределенията на двете групи показват по-голяма експресия в туморните проби.



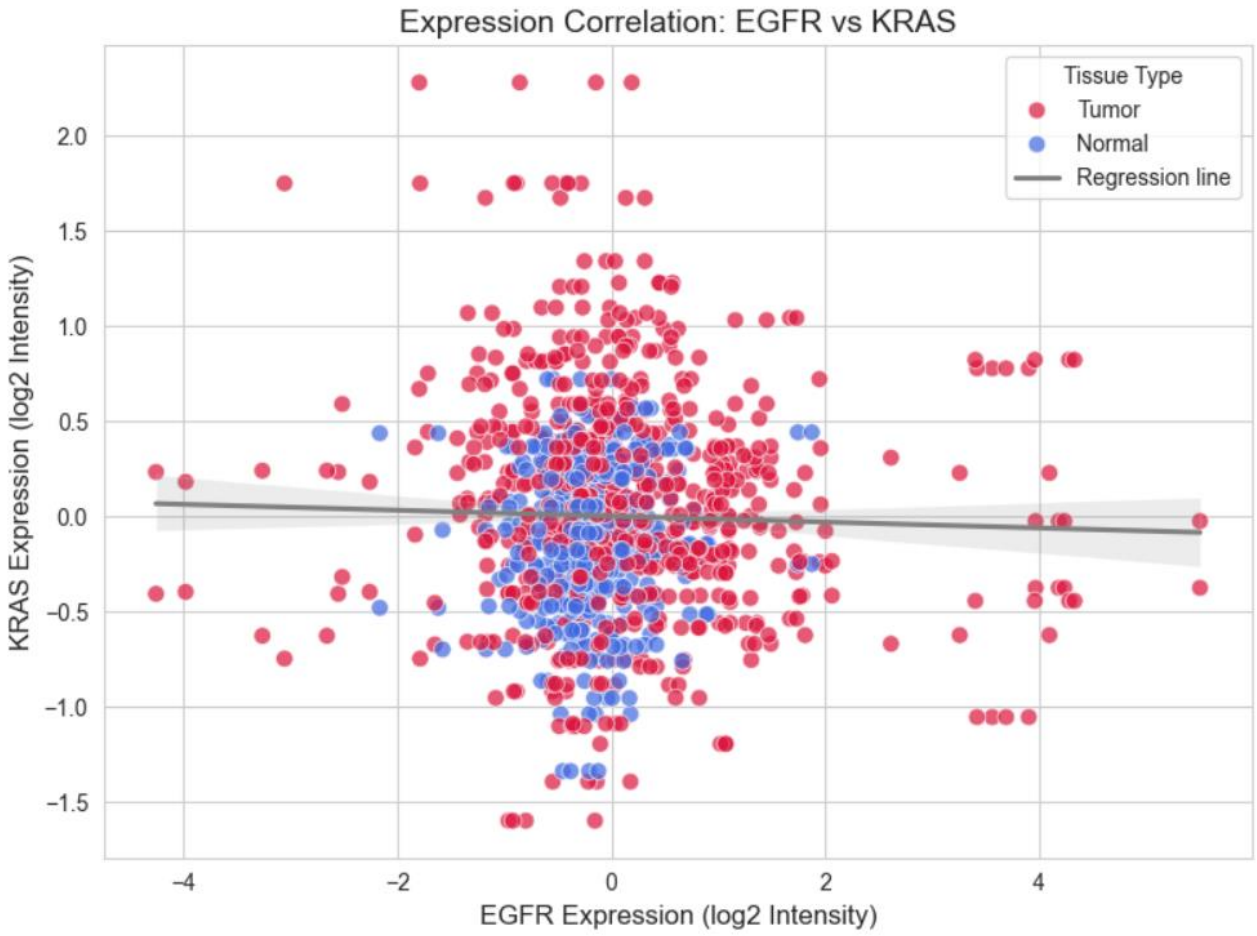
Фиг. 7 Violin plot за експресията на KRAS генът

Статистическият анализ, извършен чрез t-test показва разлика между групите ($p < 0.05$), което показва, че при белодробния рак ролята на KRAS, въпреки, че по-леко изразена заради по-високата стойност на p-value е свързана с промяна в нивото на експресия и с мутационна разлика.



Фиг. 8 Статистически анализ на KRAS генът

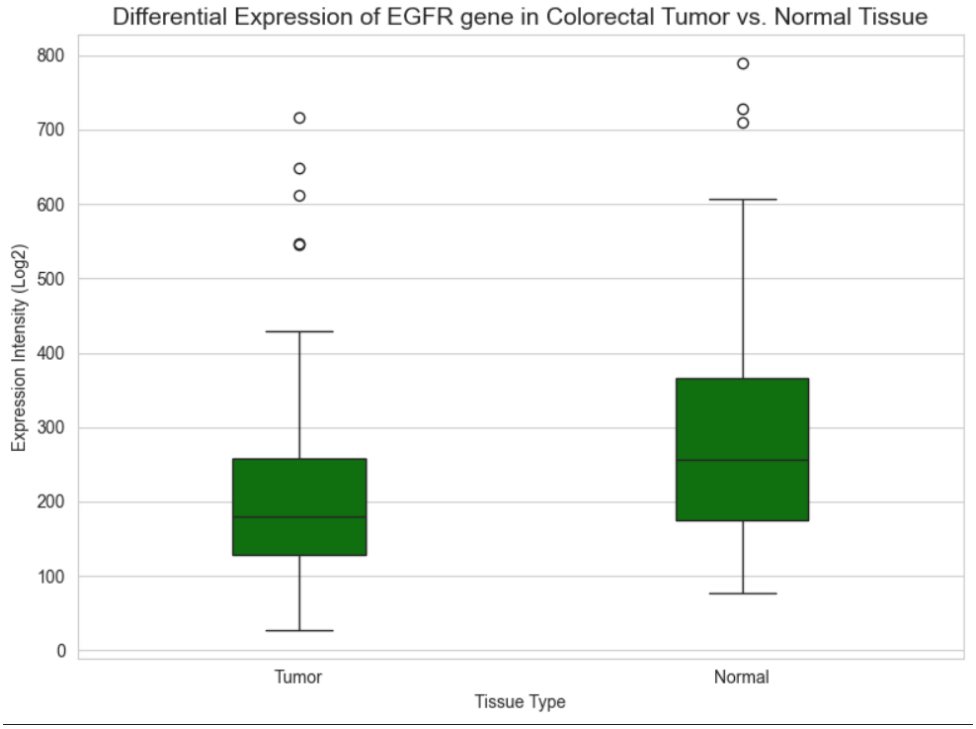
Анализът на корелацията между експресията на гените EGFR и KRAS в туморните проби от бял дроб показва слаба зависимост между двата гена. Изчислената корелация не показва силна линейна връзка, което предполага, че експресията на EGFR и KRAS се регулира относително независимо. Това е в съответствие с биологичните данни, според които активността на KRAS при белодробен рак често се определя от мутации, а не от промени в експресионното ниво, въпреки че и двата гена участват в един и същ сигнален път.



Фиг. 9 Корелационен анализ на EGFR и KRAS гените

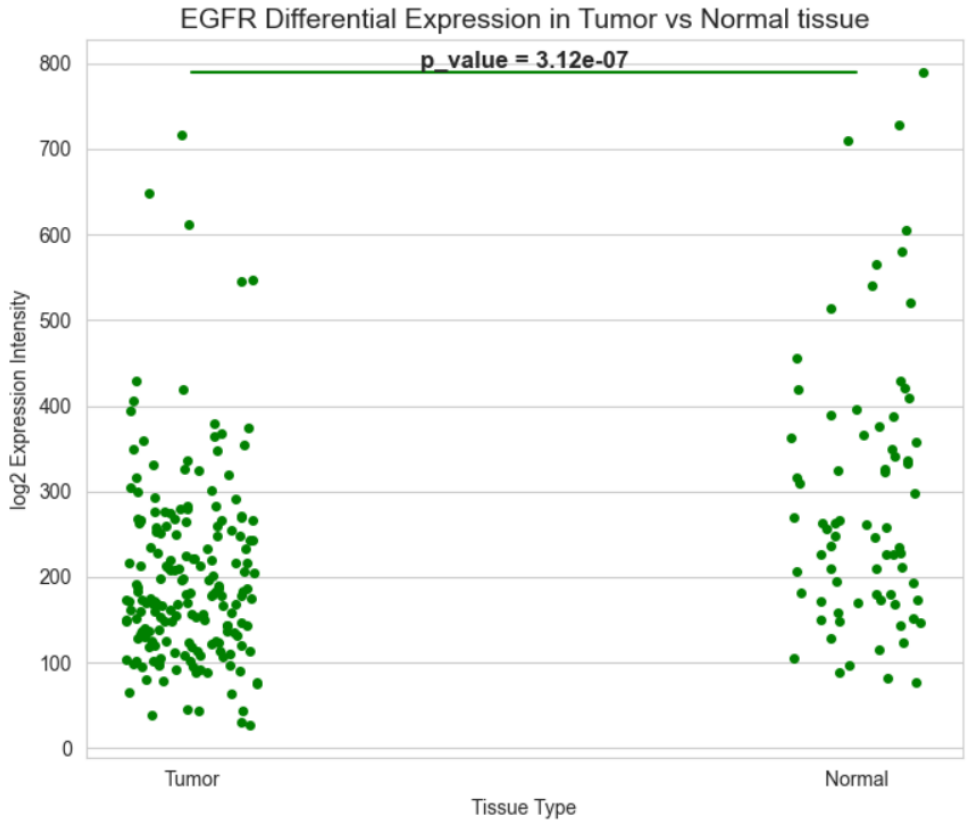
3.2. Колоректален рак

Графиката представя разпределението на експресионните нива на гена EGFR в здрави и туморни проби от дебело черво. Наблюдава се повишена експресия на EGFR в нормалните проби спрямо туморните, и въпреки непредсказуемият резултат, доказва предполагаемото участие на гена в туморното развитие при колоректален рак.



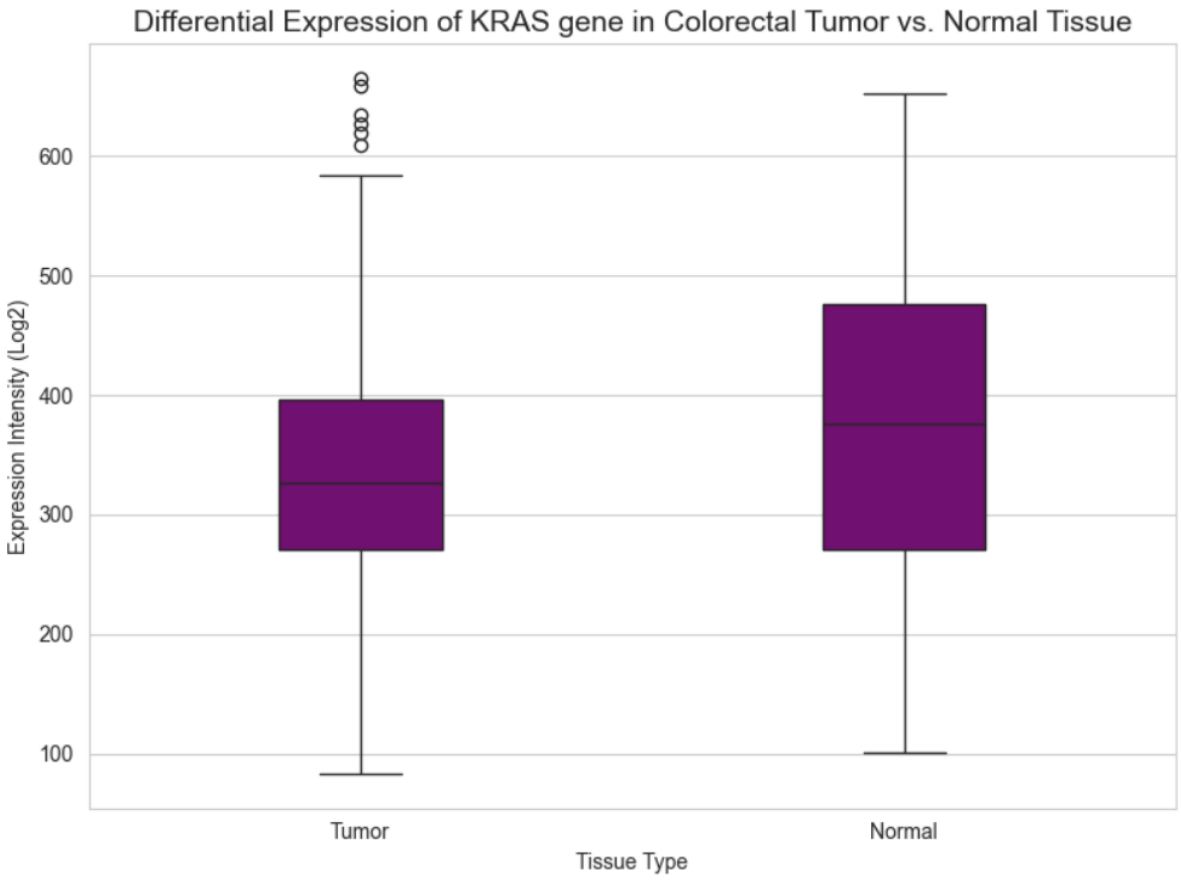
Фиг. 10 Експресия на EGFR генът при колоректален рак

Разликата между двете групи е статистически значима ($p < 0.05$), което показва, че промените в експресията на EGFR са свързани с туморното състояние.



Фиг. 11 Статистически анализ на експресията на EGFR генът

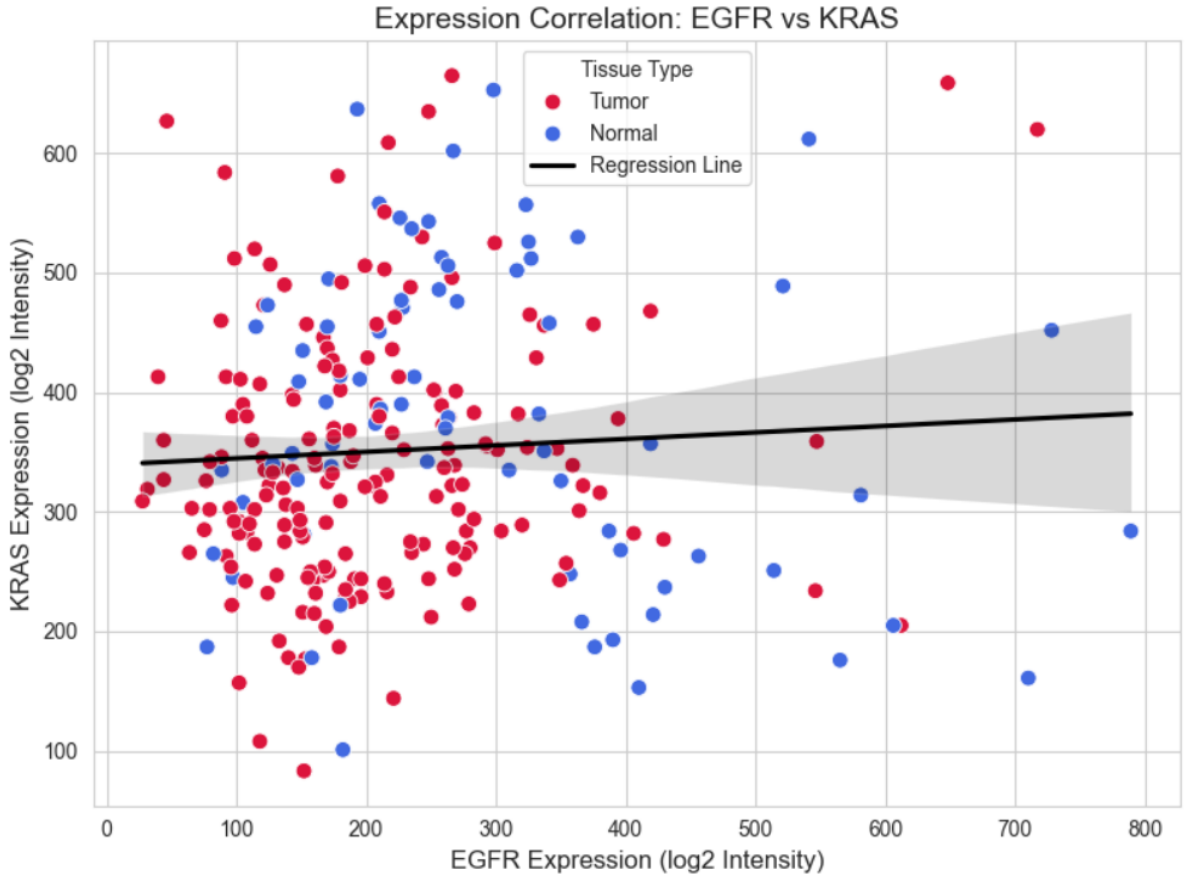
Разпределението на експресията на гена KRAS в здрави и туморни колоректални проби, отново показва повишена експресия в нормалните.



Фиг. 12 Експресия на KRAS генът при колоректален рак

Корелационният анализ между експресията на EGFR и KRAS в туморните проби от колоректален рак показва слаба зависимост между двата гена. Не се вижда силна линейна връзка, което предполага, че експресията на двата гена се регулира независимо.

Това наблюдение е в съответствие с биологичните данни, според които EGFR и KRAS участват в един и същ сигнален път, но активността на KRAS често е определяна от мутационния му статус.



Фиг. 13 Корелационен анализ между експресията на двата гена

4. Заключение и възможно бъдещо развитие

В настоящия курсов проект беше анализирана диференциалната експресия на гените EGFR и KRAS в туморни спрямо здрави тъкани, използвайки публично достъпни данни от бази като NCBI Gene, Gene Expression Omnibus и Expression Atlas. Резултатите показаха, че експресията на EGFR е повишена в някои видове тумори, и понижена в други, което е в съответствие с публикувани данни за неговата роля в стимулирането на клетъчното делене. За разлика от него, експресията на KRAS варира в зависимост от типа тумор, което предполага, че неговата роля е по-тясно свързана с мутация, отколкото с нивото на експресия. Анализът на корелацията между експресията на двата гена показва слаба зависимост в определени тумори, което отразява общото им участие в сигнални пътища, регулиращи клетъчния растеж, но не и зависимостта им. Получените резултати потвърждават значението на EGFR и KRAS като ключови гени в туморните изследвания.

В бъдещи изследвания анализът може да бъде разширен чрез включване на по-голям брой туморни видове и допълнителни експресионни набори от данни, което би позволило по-верни статистически изводи. Освен това, комбинирането на експресионни данни с информация за мутационния статус на EGFR и KRAS би дало по-пълна представа за тяхната роля в туморните изследвания.

Като следваща стъпка може да се приложи и анализ на други гени от същия сигнален път, като BRAF или ERK, за да се изследват по-широко регулаторните механизми. Като заключение, участието на клинични данни, като преживяемост и отговор на терапия, би позволило резултатите да имат по-голямо практическо значение в медицината.

5. Списък на съкращенията

EA – Expression Atlas

GEO – Gene Expression Omnibus

EGF – Epidermal Growth Factor

EGFR - epidermal growth factor receptor

KRAS - Kirsten rat sarcoma virus oncogene homologue

6. Речник на термините

EGF – протеин, който се свързва към EGFR

GPL570 platform - Affymetrix Human Genome U133 Plus 2.0 Array анотации

"201983_s_at", "201984_s_at", "211607_x_at", "210984_x_at" - специфични идентификатори на EGFR гена, използвани в микрочиповият анализ

"208926_at", "214352_s_at" - специфични идентификатори на KRAS гена, използвани в микрочиповият анализ

7. Използвани източници

EGFR and KRAS сигнален път- https://www.researchgate.net/figure/Overview-of-the-EGFR-pathway-and-downstream-signaling-pathways-including-KRAS-Adapted_fig1_237071364

EGFR ген – https://www.ncbi.nlm.nih.gov/gene/1956

KRAS ген – https://en.wikipedia.org/wiki/KRAS

NCBI база от данни - https://www.ncbi.nlm.nih.gov/

NCBI GEO база от данни - https://www.ncbi.nlm.nih.gov/geo/

Expression Atlas база от данни - https://www.ebi.ac.uk/gxa/home

RNA-seq експресионен анализ - https://microbenotes.com/rna-sequencing-principle-steps-types-uses/

Microarray експресионен анализ - https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology

Изследване рак на бял дроб - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19188

Изследване колоректален рак https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41258

Affymetrix анотации на гените: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

8. Приложения

expression_lung_cancer.ipynb

expression_colorectal_cancer.ipynb