



Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И
КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

ДИПЛОМСКА РАБОТА

Анализа на мрежи добиени од генетски податоци

Стефан Канан

Факултет за Информатички Науки и Компјутерско Инженерство

Св. Кирил и Методиј

Скопје, Македонија

kanan.stefan@students.finki.ukim.mk

Ментор:

д-р Љупчо Коцарев

Членови на комисија:

д-р Ана Мадевска Богданова

д-р Моника Симјанска

Абстракт—Ракот е комплицирана болест чии механизми се уште не се потполно познати. Бидејќи една клетка е составена од илјадници гени, пронаоѓањето на динамиката на системот, а со тоа и главните механизми на работа во него е тежок и обемен проблем. Еден начин на кој што можеме да го моделираме овој систем е со користењето на мрежи. Во овој труд се прават мрежи со помош на методот WGCNA користејќи податоци за генетската експресија во ткиво на карцином на белите дробови и нормално ткиво. Овие мрежи потоа се делат на повеќе подмрежи наречени модули, кои потоа можат да бидат подетално анализирани. Користејќи го овој метод во овој труд се наоѓаат 48 гени од два модули кои би можеле да имаат некаква значајна функција во клетките со рак. Освен тоа се гледаат врските помеѓу модулите, а истите се групираат во мета-модули кои ни откриваат нешто повеќе за однесувањето на овие групи од гени во ракот. Се користат сервиси за анотација со што се гледаат различните функционалности на модулите и мета-модулите во клетката и се откриваат неколку гени чија експресија може да има значаен прогностички ефект кај пациентите.

I. Вовед

Ракот е болест од која на глобално ниво заболуваат околу осумнаесет милиони луѓе годишно [2]. Оваа болест настанува кога клетките во едно ткиво ќе почнат непланирано и неконтролирано да се намножуваат во маса наречена тумор, чии клетки потоа ќе метастазираат и во други делови од телото [21]. Иако постојат многу внатрешни и надворешни механизми за спречување на овие дефекти во клетките, сепак одредени мутации може да ги спречат истите или да останат незабележани. Дел од причините поради кој за оваа болест се уште не постои лек е комплексноста на проблемот. Имено гените во една клетка се интринзично поврзани, па еден ген може

да соработува со повеќе други гени или пак да ја регулира нивната работа. Освен тоа канцерогените клетки може да се разликуваат од пациент до пациент, па дури и во еден пациент [21]. Надворешни фактори кај пациентот, како на пр. стилот на живеење, нивото на стрес, чистотата на воздухот, пушењето и др, како и огромниот број на гени во човечкиот геном додатно го отежнуваат овој проблем [21].

Еден од начините на кој што можеме да го претставиме работењето на гените во една клетка, е преку користење на мрежи [28]. Во овие мрежи секој јазол претставува еден ген, додека врските помеѓу нив го претставуваат нивниот меѓусебен однос. Вака претставени би можеле да извлечеме важни заклучоци за работата на овие гени, а со тоа можжеби и да откриеме нешто ново за механизмите на ракот. Во овој труд го користиме WGCNA методот за креирање и анализирање на мрежата на карцином на белите дробови (Squamous-cell lung carcinoma) кој се појавува на површината на бронхиите. Ракот на дишните патишта е меѓу најопасните видови на рак [21]. Глобално за овој вид на рак има околу два милиони случаи годишно [22], од кои 1.7 милиони завршуваат фатално [2]. Трудот е поделен на неколку делови: во дел II се опишуваат генетските и клиничките податоците со кои работиме, во делот III се опишува креацијата на мрежите и нејзината поделба во подмрежи наречени модули, во дел IV се опишуваат неколку интересни модули и нивните соодветни централни гени, модулите се делат според степенот на сочуваност меѓу двете ткива и истите се анотираат според експериментално добиената функција на нивните гени, за секоја од подмрежите

се пресметуваат неколку мрежни својства и истите се споредуваат меѓу двете ткива, освен тоа, со Kaplan-Meier оценувацот се наоѓаат и неколку значајни гени, во делот V се даваат неколку можни идеи за понатамошна работа. Конечно, во VI се дава заклучок.

II. Податоци

Податоците за ова истражување беа превземени од TCGA: <https://www.cancer.gov/tcga>. TCGA е платформа која им овозможува на истражувачите отворен пристап кон генетски и клинички податоци, добиени благодарение на донатори. Истите вклучуваат податоци за експресијата на гените во одреден момент во некое ткиво добиени преку мерење за бројот на PHK (RNA-Seq) во единица мерка TPM. Во нашиот случај податоците за генетската експресија за повеќе од дваесет илјади гени по претпроцесирањето се поделени на 51 нормални ткива, и 414 ткива земени од тумор. Освен тоа се служиме и со клинички податоци кои ни даваат опис за пациентите чие ткиво се истражува. Меѓу поважните карактеристики спаѓаат: полот на пациентот (116 женски ткива и 349 машки), годините, скалата на Карнофски, денови до смрт, TNM стадиумот (Т за големината и растот на примарниот тумор, N ни кажува колкав е бројот на лимнфните жлезди заразени со рак, додека дескрипторот M ни потенцира дали туморот метастазирал или не) [15], индикаторите CTLA4 (pos/neg) и PD-1 (pos/neg) (инхибициски протеини кои штитат од прекумерно активен имун систем) изразени преку нивната имунофено (ips) вредност [6], поголемото ниво на овие два протеини во системот може да укажува на појава на тумор.

III. Методологија

A. Претпроцесирање

Пред да можеме да работиме со генетските податоци, истите треба да се претпроцесираат. Најпрво, бидејќи се во единица мерка TPM (едно читање на секои милион килобази), нивната дистрибуција е доста навалена па истите се трансформираат логаритамски, во нашиот случај ја користиме формулата

$$x_{new} = \log_2(x + 1) \quad (1)$$

, [19] каде бројот 1 е константа која ја користиме за да избегнеме негативни вредности како и вредности кои се близнат кон $-\infty$, во случај на немање никаква експресија кај некој ген. Ваквата трансформација ја измазнува искривената дистрибуција и прави полесно да ги воочиме разликите меѓу експресијата на гените. Следниот чекор од претпроцесирањето го извршувааме со помош на R пакетот SampleNetwork [16]. Овој пакет прави мрежна анализа каде секој јазол претставува ткиво, врската помеѓу ткивата се добива со корелација, а истата е посилна ако генетска експресија е слична кај

двата јазли и обратно ако е доста различна. Потоа за оваа мрежа се пресметуваат степенот

$$k_i = \sum_{i \neq j} a_{ij} \quad (2)$$

, каде k_i е степенот на јазолот во мрежата i [8] и коефициентот на кластерирање C , дефиниран како

$$C_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{\{(\sum_{l \neq i} a_{il})^2 - \sum_{l \neq i} (a_{il})^2\}} \quad (3)$$

[8]. Корелацијата помеѓу овие две вредности $Cor(C, K)$ е добар индикатор за хомогеноста помеѓу јазлите [16]. Се покажува дека хомогеноста помеѓу ткивата се зголемува како степенот на корелација сеближи кон -1 . Ова е слика на глобалната структура на мрежата која се дели на групи на јазли со голема густина и јазли-мостови кои ги поврзуваат овие групи и ја прават мрежата. Пакетот SampleNetwork овозможува визуелно отстранување на јазли чиј степен или коефициент на кластерирање значајно се разликува од останатите јазли се додека не се добие прифатлива хомогеност на ткивата. Иако податоците за нормалното ткиво достигна $Cor(C, K) = -1$, сепак за ткивото со тумор, корелацијата достигна само $Cor(C, K) = -0.7$. Оваа хетерогеност можеби може да се објасни со самата природа на ракот, каде гените во клетките работат непредвидливо а може да постојат повеќе поткултури на клетки во туморното ткиво [21]. Освен намалување на хетерогеноста, пакетот SampleNetwork ги отстранува сите гени со нулта варијација и прави квантилна нормализација на гените.

B. Креирање на мрежата

Мрежите беа направени користејки го R пакетот WGCNA. Се верува дека кај биолошките мрежи често пати се јавува scale-free својството [28], тоа значи дека степенот ја следи Power Law дистрибуција [18] па мрежата се дели на високо конектирани јазли наречени централни јазли и слабо конектирани јазли [27] [3]. За креирање на мрежата, најпрво се бара корелацијата помеѓу експресијата на гените, а потоа, за да го доловиме ова scale-free својство и во нашите мрежи, користиме коефициент β , кој го варираме се додека квадратниот индекс на регресија помеѓу $\log(k^\beta)$ и $\log(k)$ не се доближи доволно близку до 1 [28]. Крајниот резултат е тежинска мрежа A прикажана како матрица на соседство, чии јазли се добиени со равенката

$$a_{ij} = |Cor(x_i, x_j)|^\beta \quad (4)$$

[8]. Овој коефициент помага и при намалувањето на шумот [28]. Тежинската мрежа овозможува зачувување на некои релации и детали кои би се изгубиле со користење на нетежинска мрежа. Бидејќи ја користиме апсолутната вредност на корелацијата, и позитивните

и негативните корелации ќе се прикажат со позитивен број. Биолошки гледано негативните корелации помеѓу гените може да откријат важни работи за биолошкиот систем. Во нашиот случај користиме $\beta = 3$ за мрежата создадена од туморното ткиво и $\beta = 7$ за нормалното ткиво. WGCNA методот предвидува користење на топографски преклопувачка мрежа (TOM) пресметана со

$$w_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (5)$$

. [28] Вака пресметано, врската помеѓу два јазли во мрежата ги зема во обзор и заедничките соседи на овие јазли. Ова го олеснува барањето на групи од заемно експресирани гени и ги зајакнува врските помеѓу оние јазли кои можеби поради некоја грешка се многу мали [5] [27].

B. Барање модули

Следен чекор по креирањето на мрежата е наоѓање на подмрежи од добро-поврзани гени наречени модули. Се верува дека гените во една клетка се групираат во биолошко функционални единици [27] [8] па наоѓањето на овие групи во нашата мрежа може да ни објасни нешто повеќе за природата на овој проблем. Модулите се наоѓаат со помош на хиерархиско кластерирање [26] [12] откако најпрво ќе ја претвориме матрицата на графот од матрица на сличност во матрица на разлика.

$$dissTOM = 1 - TOM \quad (6)$$

[28]. По извршувањето на овој чекор добивме 63 модули во нормалната мрежа и 61 модули во тумор мрежата. Секој модул се означува со боја. И во двете мрежи, оние гени кои не припаѓаат на ниту еден модул се сместени во grey модулот. Топлинската мапа од овие две мрежи може да се види во (Слика 1).

Секој модул е претставен со неговиот еигенген. Еигенген-от е вектор со должина n , каде n се бројот на пациенти за кои имаме генетски податоци. Постојат повеќе начини на пресметување на овој еигенген, а ние го користиме првиот principal component од PCA [10] [4]. Овој еигенген го претставува најцентралниот ген во модулот. Вака претставени, модулите, сега можат да бидат искористени во нашите анализи, со тоа WGCNA овозможува побрзи статистички анализи, бидејќи не мора да се анализира секој ген во модулот поединечно [4] [23].

Еигенгените на овие модули може да се искористат и за меѓусебно корелирање. Модулите со соодветно висока корелација на пример би можеле да се соединат (Во овој труд тоа е $Cor > 0.75$). Освен тоа може да се видат и врските помеѓу различните модули и групите во кои тие спаѓаат (мета-модули) [10]. Во тумор метамрежата се откриваат неколку мета-модули и тоа: I (Purple, Lightyellow, Greenyellow, Black, Blue, Plum2,

Turquoise), II (Paleturquoise, Brown4, Maroon, Darkgrey, Skyblue3, Darkturquoise, Darkorange2, Darkgreen, Cyan, Brown, Green, Saddlebrown, Magenta, Red), III (Royalblue, Tan, Yellowgreen, Palevioletred3, Darkolivegreen, Grey60), IV (Pink, Darkorange, Lightgreen, Salmon, Yellow, Orangered4, Thistle2, Violet), V (Bisque4, Lavenderblush3), VI (Salmon4, Plum1, Honeydew1, Ivory), VII (Thistle1, Lightcyan1, Lightcyan, Midnightblue, Darkred, White, Darkseagreen4, Floralwhite, Darkmagenta), VIII (Darkslateblue, Coral1, Lightsteelblue1, Mediumpurple3), IX (Sienna3, Navajowhite2, Orange), X (Skyblue, Lightpink4, Steelblue). Топлинската мапа за врските помеѓу мета-модулите може да се види во (Слика 2). Меѓу поинтересните врски се издвојуваат: I - II ($Cor = -0.71$), I - III ($Cor = -0.37$), I - IV ($Cor = -0.37$)

IV. Резултати

A. Барање на интересни модули

Еден начин како модулите го скратуваат обемот и времето на пресметки е при анализирање на врските помеѓу клиничките податоци и гените. Наместо пресметување на корелацијата на секој ген поединечно, корелацијата се врши помеѓу секој модул и клиничкиот податок [8] [4] [3]. Во нашиот случај се откриваат неколку вакви интересни врски во тумор мрежата, меѓу поинтересните се: yellow со скалата на карнофски (0.31), plum2 со ips_ctla4_neg_pd1_pos (0.51), ips_ctla4_pos_pd1_neg (0.40), ips_ctla4_pos_pd1_pos (0.53), turquoise со ips_ctla4_neg_pd1_pos (0.53), ips_ctla4_pos_pd1_neg (0.32) и ips_ctla4_pos_pd1_pos (0.6), grey со пол (0.75). Откако ги наоѓаме интересните модули можеме да продолжиме на самите гени во овие модули.

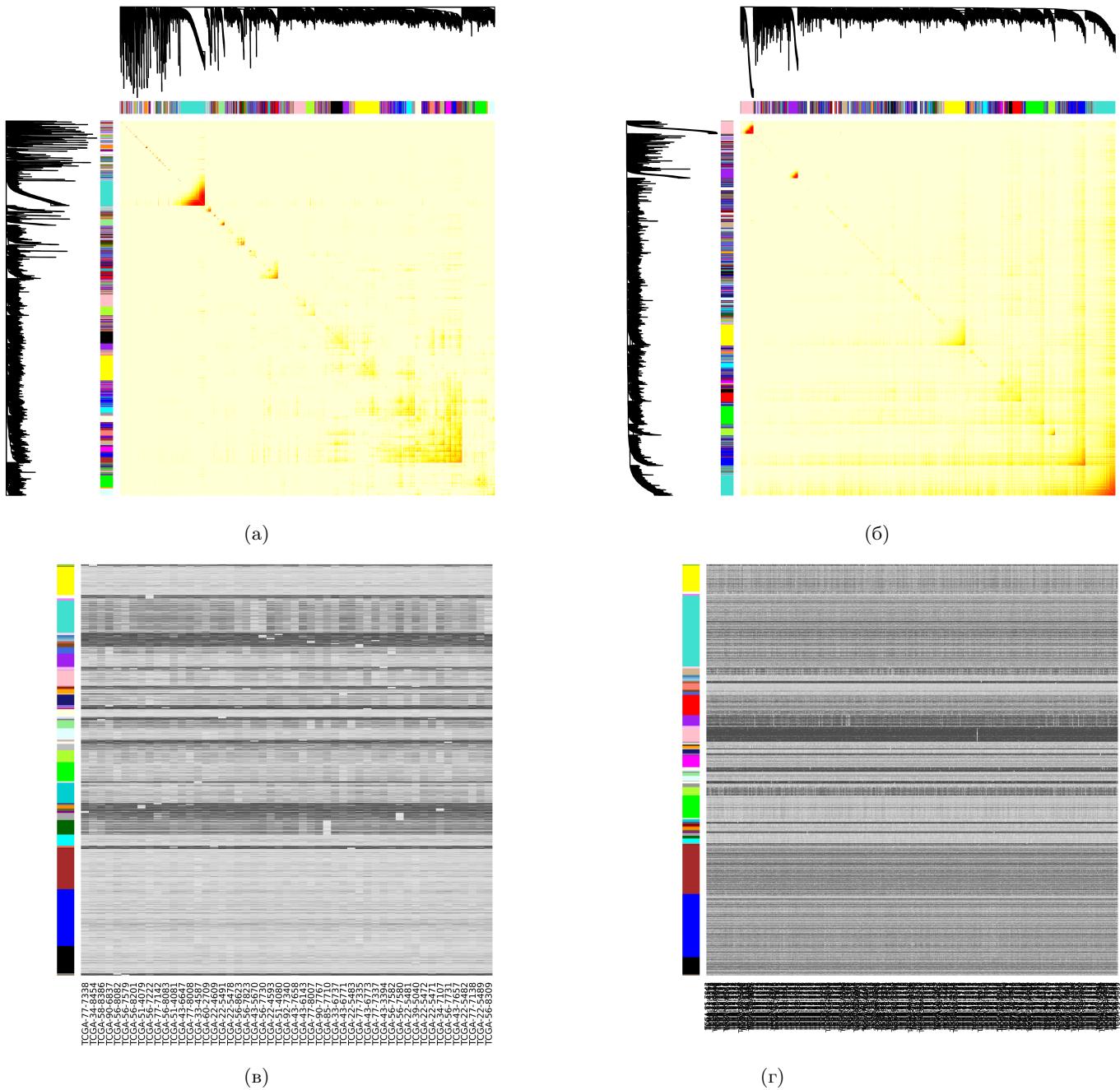
B. Централни гени

Централните јазли во една биолошка мрежа се нарекуваат централни гени и истите би можеле да играат клучна улога во клетките, овие гени се добри кандидати за понатамошно истражување како можни биомаркери [14]. Во WGCNA, централните гени се јазлите кои имаат висок интра-модуларен степен и висока значајност [14] [26] [28]. Значајноста на гените е некој екстерен фактор кои ние им го придаваме, пример: веројатноста дека генот се јавува во туморни клетки или пак корелацијата на генот со некој клинички податок како во нашиот случај. Значајноста на гените се пресметува со

$$GS_x = |Cor(x_i, t_i)|^\beta \quad (7)$$

, каде x ја претставува генетската експресија, t е некој клинички податок [8] [3]. Наместо степен, може да се користи и мерката Module Membership, која е интризнично поврзана со степенот,

$$MM_x = |Cor(x_i, M)| \quad (8)$$

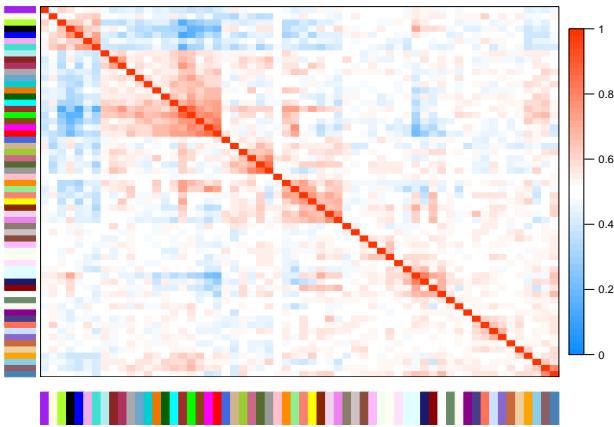


Слика 1: Топлински мапи за (а) нормалната мрежа и (б) тумор мрежата. Експресија по модули за (в) нормалното и (г) туморното ткиво, колоните се пациенти а редовите гените кои го сочинуваат модулот.

, М е еигенгенот на модулот [11]. Освен значајноста на гените, можеме да ја најдеме значајноста на еден модул (Module Significance) и значајноста на централните гени (Hub Gene Significance) која го претставува наклонот на регресијата меѓу значајноста на гените и нивниот степен [8]. Значајноста на централните гени може да ни открие важни биолошки информации, но не мора да значи дека гените со ваква мала вредност не можат да бидат важни за системот [8].

Еден ген во некој модул го сметавме за централен

ако има $GS > 0.2$ и $MM > 0.8$. Нашата анализа откри 48 различни централни гени од два различни модули, turquoise и од grey; и тоа: APOL3, CD247, CD2, CD3D, CD3E, CXCR3, CXCR6, IL12RB1, ITK, NKG7, SH2D1A, SIRPG, SLA2, ACAP1, ARHGAP9, CCL4, CCR5, CD5, CD6, CD96, CORO1A, CST7, GIMAP2, GIMAP5, GIMAP7, GPR174, GZMK, ICOS, IL21R, IL2RB, ITGAL, ITGB7, LCK, P2RY10, PTPN22, PTPRCAP, PYHIN1, RHOH, SIT1, SLAMF1, SLAMF6, TBC1D10C, TRAF3IP3, UBASH3A, CYorf15A, EIF1AY, RPS4Y1,



Слика 2: Топлинска мапа, која ги истакнува метамодулите и нивните врски

USP9Y. Значајноста на модулот и централниот ген беа прилично мали во Turquoise модулот, но кај Grey имаа вредности $HGS = 0.24$ и $MS = 0.16$.

Понатамошна статистичка анализа открива значајни разлики во експресијата на овие централни гени во зависност од корелираната клиничка состојба на пациентот. Во (Слика 3) се прикажани експресиите на два гени, генот APOL3 и неговата врска со `ips_ctla4_pos_pd1_pos` и генот USP9Y и неговата врска со полот на пациентот.

В. Анализа на модули

Мрежната структура може значително да се разликува од едната до другата мрежа. Користејќи ја мерката $Z_{summary}$ можеме да одредиме дали еден модул е целосно или делумно зачуван помеѓу мрежата која ја тестираме и референцната мрежа, или пак е ендемичен за мрежата која ја анализираме [11]. Оваа мерка е дефинирана со

$$Z_{summary} = \frac{Z_{density} + Z_{connectivity}}{2} \quad (9)$$

. $Z_{density}$ е медијаната од Z статистиките за: средната вредност на степенот, средната вредност на приврзаноста на гените до модулот (kME) и пропорцијата на варијација описана преку евгенгенот на модулот, додека $Z_{connectivity}$ е медијана од Z статистиките од корелацијата на интрамодуларниот степен, корелација и kME помеѓу тест и референцната мрежа. Овие статистики се наоѓаат со тоа што јазлите за секој модул во тест мрежата се пермутираат n пати (во нашиот случај 200 пати), а за секоја пермутација на модулот се пресметуваат степенот, средната вредност и другите вредности во референцната мрежа. Потоа ја пресметуваме истата вредност за оригиналниот модул во тест мрежата и ја добиваме Z статистиката

$$Z_a = \frac{obs_a - \mu_a}{\sigma_a} \quad (10)$$

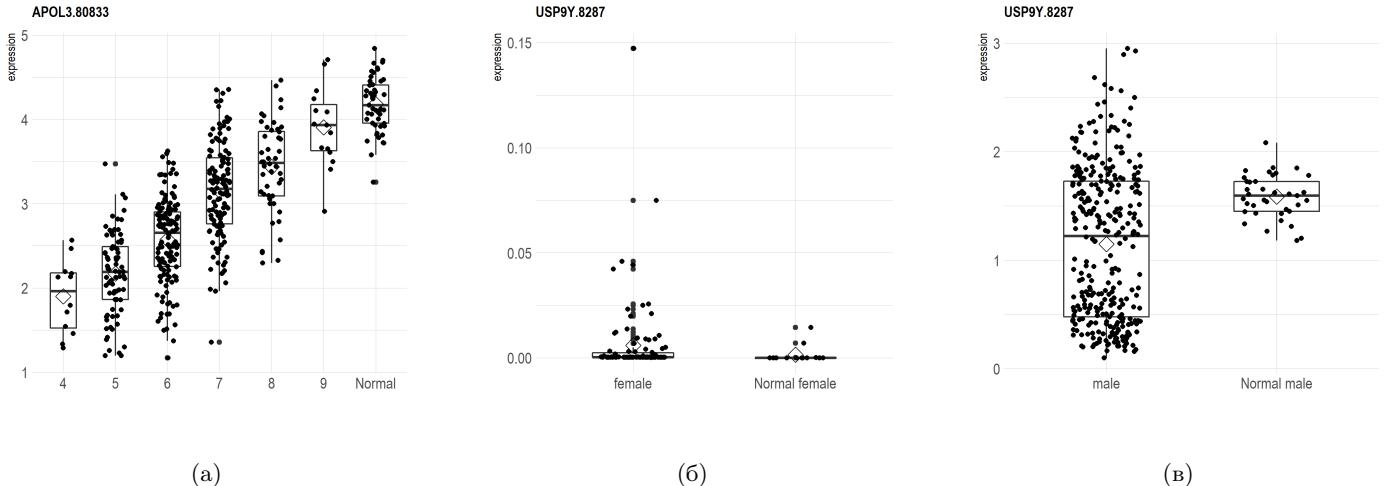
каде μ_a и σ_a се средната вредност и варијацијата добиена од вредностите во референцната мрежа [11]. Се покажува дека повеќето сочувани модули имаат вредност $Z_{summary} > 10$, оние кои се послабо сочувани имаат вредност меѓу $2 \geq Z_{summary} \leq 10$, додека оние кои не се сочувани имаат вредност помала од 2 [11]. Нашата анализа покажа дека се зачувани 20 модули (меѓу кои најважен е Turquoise со $Z_{summary}$ 21 и највеќе сочуван Purple со $Z_{summary}$ 140), 24 се помалку презервирали, додека 16 воопшто не се презервирали.

За сите од пронајдените модули пресметуваме интрамодуларна централност, средната вредност на коефициентот на класерирање, степен (средна вредност и варијанса), густина, големина и хетерогеност [8] [5]. Модулите беа групирани по степенот на презервираност и споредени со нивните соодветни модули од нормалното ткиво. Оваа споредба може да се види во (Слика 4). За презервираните модули беше искористен Kruskal-Wallis тестот кој покажува разлика во варијацијата ($p = 0.002$), централноста ($p < 0.0001$), коефициент на класерирање ($p = 0.0001$) и ANOVA тестот кој покажа разлика во хетерогеноста ($p < 0.001$). Во послабо презервираните модули разлика има меѓу: средната вредност ($p < 0.0001$), густината ($p < 0.005$) и коефициентот на класерирање ($p < 0.001$) со Welch тестот; варијацијата ($p < 0.0001$) и централноста ($p < 0.0001$) со тестот на Kruskal-Wallis, и хетерогеноста ($p < 0.001$) со ANOVA тестот. За непрезервираните постои разлика меѓу средната вредност ($p < 0.0001$), хетерогеноста ($p < 0.001$) и централноста ($p < 0.0001$) пресметани со Welch тестот и варијацијата ($p < 0.0001$), густината ($p < 0.0001$) и коефициентот на класерирање ($p < 0.0001$) со Kruskal-Wallis тестот.

Централноста во модулите од тумор мрежата е релативно пониска од модулите во нормалната мрежа без разлика на нивото на сочуваност, ова би можело да укажува на губењето на повеќето од централните јазли од нормалната мрежа. Степенот и варијацијата се со пониски вредности од нормалната мрежа, а истото е случај и со хетерогеноста. Хетерогеноста е уште еден намален индикатор кај модулите од тумор мрежата кој укажува на губењето на scale-free својството во истата [8]. Истото е случај и кај коефициентот на класерирање и густината (во помалку сочуваните и несочуваните) што би можело да значи дека модулите во тумор мрежата се помалку сврзани за разлика од модулите во нормалната мрежа.

Г. Анотација на гени

Анотацијата на гени е метод со кој можеме да откриеме некои од функциите на гените во модулот користејќи претходно добиени експериментални резултати. За потребите на овој проект беше



Слика 3: (а) Експресија на APOL3 според имунофено (CTL4 Pos, PD1 Pos) оценката наспроти нормалната експресија и експресија на USP9Y кај (б) жени и (в) мажи во нормално и туморно ткиво

искористена алатката gProfiler [20]. gProfiler користи сервиси за генетска онтологија (GO [1] [25]), наоѓање на молекуларни и биолошки патишта (KEGG и др), протеини и поврзување на гени со различни болести. Во (табела I) може да се видат резултатите од анотацијата на модулот Turquoise, Purple, како и анотацијата на модулите според нивната сочуваноста и на некои мета-модули.

Мета-модулот I највеќе се поврзува со сигнализирањето, придвижувањето, регулација на имуниот систем и прилепувањето на клетките. Интересно е тоа што овие функции се механизми користени од оваа болест. Мета-модулот II е з bogатена со гени кои имаат функција во внатрешниот раз соок, анатомската структура, метабослките процеси, фазите во клетката, поправање на грешки во ДНК и др. Меѓу поинтересните анотирани модули се и Turquoise, кој е з bogатен со гени поврзани со имуниот систем, активација на Т клетките, туберколоза, сигнализација и др. Како и Purple модулот кој составен од гени поврзани со движењето на клетката, создавањето на органели во клетката, како и абнормалности во респираторниот систем, болеста хунтингтон и др.

Пребарувањето низ литератур и анотацијата на централните гени откри повеќе гени поврзани со сигналната трансдукција, Т-клетки и други имуни функции, како и цели за таргетирање и гени за кои е познато дека се поврзани со рак на белите дробови и нефертилноста.

Д. Оценувањот на Kaplan-Meier

Овој непараметарски оценувач има за цел да го пресмета процентот на преживеани пациенти дефинирани според некој фактор (клинички стадиум, генетска експресија) со текот на времето. За Kaplan-

Meier се потребни податоци кои го следат времето до смрт на пациентите или времето до цензура (за цензура се смета времето се додека пациентот не се: откажал од студијата, бил излечен и др.) [7]

Процентот на преживеани пациенти до момент t , се пресметува со

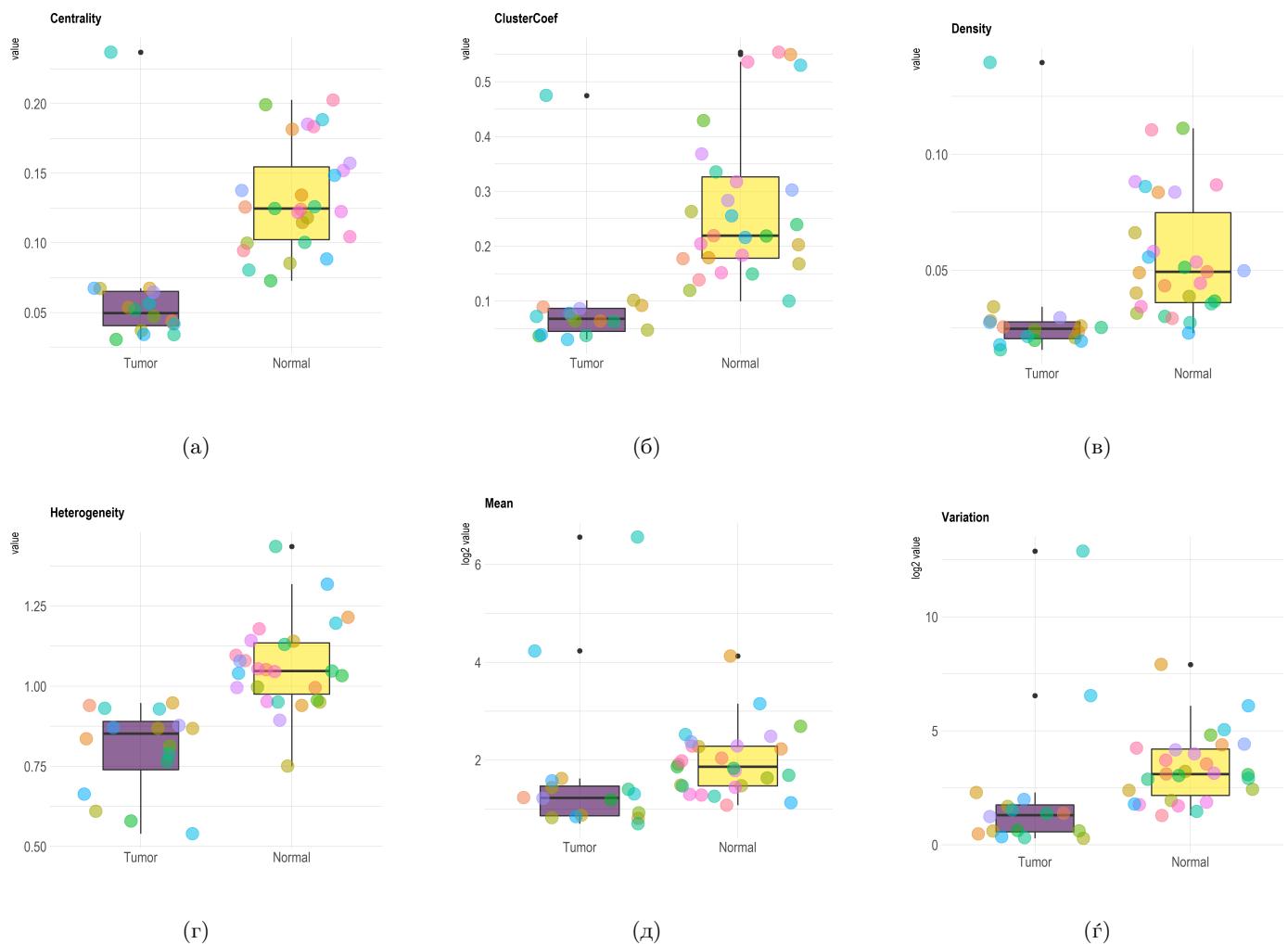
$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (11)$$

[7], каде d_i се пациентите кои умреле во време t , а n_i се сите пациенти кои преживеале или не биле цензорирани до момент t .

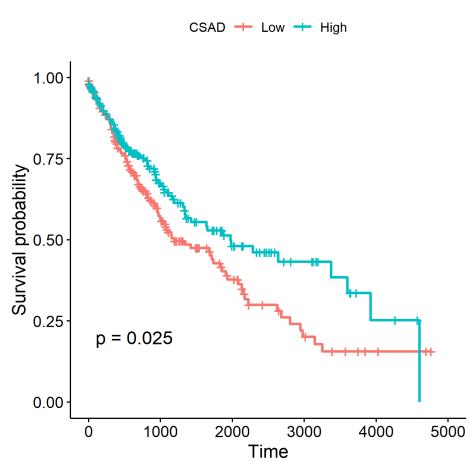
Нашата анализа не најде значаен резултат во централните гени (поделени според нивото на експресија во ткивото во високо и ниско експресирани случаи), ниту пак користејќи ги клиничките податоци, најзначајниот резултат меѓу овие гени беше генот TRAF3IP3 со p вредност од 0.076. Но потребни се повеќе податоци за да се види дали овој резултат би станал значаен. Со олабавување на ограничувањето за централните гени (поставување на $\beta = 1$) и нивна анализа со Kaplan-Meier оценувањето се добиваат неколку позначајни гени: AIF1 ($p = 0.047$), C1orf38 ($p = 0.0076$), CCR1 ($p = 0.048$), CD14 ($p = 0.0012$), CSAD ($p = 0.025$), ENG ($p = 0.027$), FGG ($p = 0.0055$), LRRC25 ($p = 0.012$), NAA30 ($p = 0.0068$), ROS1 ($p = 0.0072$), SFTA2 ($p = 0.0047$), SFTPD ($p = 0.029$), SMEK1 ($p = 0.026$), ZNF692 ($p = 0.0041$), SIGLEC9 ($p = 0.027$). Во (Слика 5) може да се види Kaplan-Meier оценувањето за генот CSAD.

V. Понатамошна Работа

Нашата анализа откри 48 различни централни гени чија експресија се менува во зависност од клиничката состојба на пациентот. Постојат докази



Слика 4: (а) Централноста, (б) коефициент на кластерирање, (в) густина, (г) хетерогеноста, (д) средна вредност на степенот, (ѓ) варијација на степенот во несочуваните модули во тумор мрежата споредени со нормалната мрежа



Слика 5: Kaplan-Meier оценувањот за генот CSAD.

дека централните гени добиени со анализа на генетска експресија не секогаш се повторливи при користење на други податоци [9], понатамошната анализа би можела да се фокусира на експерименталните резултати за овие централни гени, нивната прогностичка моќ и нивната можна улога како биомаркери во клетката [14].

Интересен правец на истражување е споредувањето на повеќе методи на кластерирање на гените во модули, и нивната биолошка значајност. на пример споредба на WGCNA методот, со Independent Component Analysis и др. Иако во овој труд се анализираат некои мрежни метрики како централизацијата, густината, коефициентот на кластерирање и др, понатамошна анализа би можела да се фокусира подетално на структурата на мрежата со помош на графлети [17] и ентропијата во мрежите.

Иако овој тип на карцином често пати се поврзува

Табела I: Анотација на важни модули, мета-модули, како и анотации според сочуваноста на модулите. Секоја анотација има значајност $p < 0.05$.

	Анотации
Meta-Module I	GO BIOLOGICAL PROCESS » anatomical structure development, biological adhesion, cell surface receptor signaling pathway, movement of cell or subcellular component, regulation of biological quality, regulation of cell communication, regulation of signaling, cellular response to chemical stimulus, cellular response to organic substance, tube development, tube morphogenesis, regulation of defense response, regulation of immune response, regulation of localization, transport ◇ GO CELLULAR COPMONENT » cell periphery, plasma membrane, cytoplasm, extracellular space ◇ HP » Abnormality of the nasopharynx, Abnormality of the paranasal sinuses, Recurrent respiratory infections, Recurrent upper respiratory tract infections, Respiratory tract infection, Sinusitis ◇ KEGG » Epstein-Barr virus infection, NOD-like receptor signaling pathway, Pertussis, Staphylococcus aureus infection ◇ REAC » Cytokine Signaling in Immune system, Immune System ◇ WP » Human Complement System
Meta-Module II	GO BIOLOGICAL PROCESS » metabolic process, anatomical structure morphogenesis, cell development, cellular biosynthetic process, cellular component biogenesis, cellular component organization or biogenesis, cellular macromolecule biosynthetic process, cellular metabolic process ◇ GO CELLULAR COMPONENTS » intracellular organelle part, cytoplasm, cytosol, envelope, nucleoplasm ◇ KEGG » Cell cycle, Mismatch repair ◇ REAC » Metabolism of proteins
Сочувани модули	GO BIOLOGICAL PROCESS » metabolic process, biosynthetic process, gene expression, mRNA metabolic process, anatomical structure development, RNA metabolic process, cell differentiation ◇ GO CELLULAR COMPONENTS » intracellular, organelle, cell-cell junction, cell projection ◇ GO MOLECULAR FUNCTION » nucleic acid binding, DNA-binding transcription factor activity ◇ KEGG » Spliceosome ◇ REAC » Metabolism of RNA, mRNA Splicing ◇ WP » Cytoplasmic Ribosomal Proteins, mRNA Processing
Полусочувани модули	GO BIOLOGICAL PROCESS » cellular aromatic compound metabolic process, cellular component biogenesis, gene expression ◇ GO CELLULAR COMPONENT » intracellular, nuclear lumen, intracellular organelle part, nuclear part ◇ GO MOLECULAR FUNCTION » DNA binding, nucleic acid binding ◇ KEGG » Herpes simplex virus 1 infection ◇ REAC » Gene expression (Transcription), Generic Transcription Pathway, Metabolism of proteins, RNA Polymerase II Transcription
Несочувани модули	GO BIOLOGICAL PROCESS » cell communication, signaling ◇ GO CELLULAR COMPONENTS axon, cell periphery, integral component of membrane, plasma membrane ◇ GO MOLECULAR FUNCTION » olfactory receptor activity ◇ KEGG » Chemical carcinogenesis, Olfactory transduction ◇ REAC » Biological oxidations, Olfactory Signaling Pathway ◇ WP » Metapathway biotransformation Phase I and II
Turquoise модул	GO BIOLOGICAL PROCESS » immune response, cell activation, defense response, response to other organism, T cell activation, response to stimulus, cell communication, signal transduction ◇ GO CELLULAR COMPONENTS » plasma membrane, cell periphery, membrane part, membrane ◇ GO MOLECULAR FUNCTION » cytokine receptor activity, signaling receptor activity, MHC protein binding, lipid binding ◇ HP » Abnormality of the lymph nodes, Splenomegaly, Autoimmunity, Abnormal cellular immune system morphology, Decrease in T cell count, Recurrent infections ◇ KEGG » Natural killer cell mediated cytotoxicity, Tuberculosis, T cell receptor signaling pathway, B cell receptor signaling pathway, Pathways in cancer, Apoptosis, Pertussis ◇ REAC » Immune System, PD-1 signaling, Hemostasis, Platelet activation, signaling and aggregation ◇ WP » T-Cell antigen Receptor (TCR) Signaling Pathway, Apoptosis, Chemokine signaling pathway, B Cell Receptor Signaling Pathway
Purple модул	GO BIOLOGICAL PROCESS » cilium movement, axoneme assembly, microtubule-based movement, microtubule-based process, cell projection assembly, organelle assembly, specification of symmetry, inner dynein arm assembly ◇ GO CELLULAR COMPONENTS » cilium, axoneme, cell projection part, cytoplasmic region ◇ GO MOLECULAR FUNCTION » dynein light chain binding, microtubule motor activity, motor activity, ATP-dependent microtubule motor activity, plus-end-directed ◇ HP » Abnormal ciliary motility, Abnormal respiratory motile cilium physiology, Bronchiectasis, Respiratory insufficiency due to defective ciliary clearance, Sinusitis, Recurrent sinusitis, Nasal polypsis, Abnormality of the nasal mucosa ◇ KEGG » Huntington disease, Salivary secretion ◇ REAC » Carboxyterminal post-translational modifications of tubulin, Diseases of glycosylation ◇ WP » Metapathway biotransformation Phase I and II

со пушењето, сепак не постоеа доволно адекватни податоци за пациентите-пушачи. Во иднина би можело да се разгледа овој ефект во контекстот на нашата мрежа. Конечно нашата анализа откри дека USP9Y е поврзан со нефертилноста кај мажите [13], а за EIF1AY, RPS4Y1 дека истите се добри индикатори за полот на пациентот [24], како и псевдо-генот CYorf15A. Сите овие гени беа дел од grey модулот, во кој се внесуваат сите гени кои не припаѓаат во ни еден друг модул. Една хипотеза која во иднина би можела да се тестира е дали креирањето на оделни мрежки врз база

на пол би ги внело овие гени во некој друг модул без притоа истите да се појават како централни гени.

VI. Заклучок

Во овој труд, со користење на методот WGCNA се изврши мрежна анализа на нормално и туморно ткиво земено од пациенти со рак на белите дробови (Squamous Cell Carcinoma). Мрежите потоа беа додатно поделени на помали подмрежи наречени модули, кои имаат за цел да ги најдат вистинските функционални генетски единици. Потоа секој модул се корелира со некој клинички податок за пациентот. За оние модули

кои имаат поголема корелација се бараат централните гени (гени кои имаат висок степен и генетска значајност), овие гени би можеле да ни објаснат нешто повеќе за овој вид на рак, во нашиот случај, поголемиот број на гени се разликуваат во експресијата во зависност од имунофено вредноста за протеините CTLA4 и PD1, се наоѓаат и гени чија експресија е можен индикатор за полот на пациентот, но нивната значајност во овој вид на рак не е позната. Освен тоа се бара сочуваноста на модулите, при што тие се делат на сочувани, помалку сочувани и несочувани. Потоа беше направена анотација на гени (според модули, метамодули и сочуваност) користејќи ја алатката gProfiler која ни откри некои од функционалностите на овие групи од гени во мрежата. Освен тоа за овие категории беа пресметани неколку вредности (централизација, густина, коефициент на кластерирање, хетерогеноста и др) при што се покажа разлика во структурата на мрежите помеѓу двете мрежи. Конечно беше пресметан процентот на преживеани пациенти со текот на времето според нивната клиничка состојба и нивото на генетска експресија на одредени гени (високо и ниско експресирани), иако оваа анализа не покажа значајни резултати меѓу централните гени, сепак со олабавување на ограничувањето за централни гени се добиваат неколку гени чија експресија, иако не покажува радикална разлика во Kaplan-Meier оценувачот сепак дава статистички значајни резултати.

Литература

- [1] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium*. *Nature Genetics* 25, 1 (May 2000), 25–29.
- [2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68, 6 (2018), 394–424.
- [3] Chen, L., Yuan, L., Qian, K., Qian, G., Zhu, Y., Wu, C.-L., Dan, H. C., Xiao, Y., and Wang, X. Identification of Biomarkers Associated With Pathological Stage and Prognosis of Clear Cell Renal Cell Carcinoma by Co-expression Network Analysis. *Frontiers in Physiology* 9 (2018), 399.
- [4] DiLeo, M. V., Strahan, G. D., Bakker, M. d., and Hoekenga, O. A. Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLOS ONE* 6, 10 (Oct. 2011), e26683.
- [5] Dong, J., and Horvath, S. Understanding network concepts in modules. *BMC Systems Biology* 1 (June 2007), 24.
- [6] Givechian, K. B., Wnuk, K., Garner, C., Benz, S., Garban, H., Rabizadeh, S., Niazi, K., and Soon-Shiong, P. Identification of an immune gene expression signature associated with favorable clinical features in Treg-enriched patient tumor samples. *NPJ Genomic Medicine* 3 (June 2018).
- [7] Goel, M. K., Khanna, P., and Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research* 1, 4 (2010), 274–278.
- [8] Horvath, S., and Dong, J. Geometric Interpretation of Gene Coexpression Network Analysis. *PLoS Computational Biology* 4, 8 (Aug. 2008).
- [9] Jahid, M. J., and Ruan, J. Identification of biomarkers in breast cancer metastasis by integrating protein-protein interaction network and gene expression data. In 2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GEN-SIPS) (Dec. 2011), pp. 60–63.
- [10] Langfelder, P., and Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* 1, 1 (Nov. 2007), 54.
- [11] Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. Is My Network Module Preserved and Reproducible? *PLoS Computational Biology* 7, 1 (Jan. 2011).
- [12] Langfelder, P., Zhang, B., and Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 5 (Mar. 2008), 719–720.
- [13] Lee, K. H., Song, G. J., Kang, I. S., Kim, S. W., Paick, J.-S., Chung, C. H., and Rhee, K. Ubiquitin-specific protease activity of USP9y, a male infertility gene on the Y chromosome. *Reproduction, Fertility, and Development* 15, 1-2 (2003), 129–133.
- [14] Lesterhuis, W. J., Rinaldi, C., Jones, A., Rozali, E. N., Dick, I. M., Khong, A., Boon, L., Robinson, B. W., Nowak, A. K., Bosco, A., and Lake, R. A. Network analysis of immunotherapy-induced regressing tumours identifies novel synergistic drug combinations. *Scientific Reports* 5 (July 2015).
- [15] Li, S., Liu, X., Liu, T., Meng, X., Yin, X., Fang, C., Huang, D., Cao, Y., Weng, H., Zeng, X., and Wang, X. Identification of Biomarkers Correlated with the TNM Staging and Overall Survival of Patients with Bladder Cancer. *Frontiers in Physiology* 8 (2017), 947.
- [16] Oldham, M. C., Langfelder, P., and Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Systems Biology* 6 (June 2012), 63.
- [17] Przulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics (Oxford, England)* 23, 2 (Jan. 2007), e177–183.
- [18] Rai, A., Pradhan, P., Nagraj, J., Lohitesh, K., Chowdhury, R., and Jalan, S. Understanding cancer complexome using networks, spectral graph theory and multilayer framework. *Scientific Reports* 7 (Feb. 2017).
- [19] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Soccia, N. D., and Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* 14, 9 (Sept. 2013), 3158.
- [20] Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* 47, W1 (July 2019), W191–W198.
- [21] Ray, S. The Cell: A Molecular Approach. *The Yale Journal of Biology and Medicine* 87, 4 (Dec. 2014), 603–604.
- [22] Siegel, R. L., Miller, K. D., and Jemal, A. Cancer statistics, 2019. *CA: a cancer journal for clinicians* 69, 1 (Jan. 2019), 7–34.
- [23] Song, L., Langfelder, P., and Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13 (Dec. 2012), 328.
- [24] Staedtler, F., Hartmann, N., Letzkus, M., Bongiovanni, S., Scherer, A., Marc, P., Johnson, K. J., and Schumacher, M. M. Robust and tissue-independent gender-specific transcript biomarkers. *Biomarkers: Biochemical Indicators of Exposure, Response, and Susceptibility to Chemicals* 18, 5 (Aug. 2013), 436–445.
- [25] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* 47, D1 (Jan. 2019), D330–D338.
- [26] Yin, L., Cai, Z., Zhu, B., and Xu, C. Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. *Genes* 9, 2 (Feb. 2018).
- [27] Yip, A. M., and Horvath, S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8 (Jan. 2007), 22.
- [28] Zhang, B., and Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4 (2005), Article17.