

به نام خدا

پروژه درس زیست شناسی سامانه ها



استاد درس: خانم دکتر صالحی

اعضای گروه:

مهرو حاجی مهدی - ندا اصفهانی - فائزه باهوش

علی جلالی - محمد امین کیانی

مقدمه:

GCNF گیرنده هسته ای عضو خانواده گیرنده های هورمون هسته ای را کد می کند. الگوی بیان آن نشان می دهد که ممکن است در نوروزن و رشد سلول های زایای نقش داشته باشد. این ژن سطح بیان 4Oct را در طول تمایز سلولی ES و رشد اولیه جنین موش کنترل می کند. GCNF در سلول های ES انسانی 9H با استفاده از Tet-On Inducible system (یک سیستم بیان شرطی ژن که در آن رونویسی در حضور تتراسایکلین یا داکسی سایکلین روشن یا خاموش می شود) بیش از حد بیان شد. آنالیز mRNA microarray نشان داد که بیان بیش از حد GCNF به صورت گلوبال، بیان ژن را در سلول های hES تمایز نیافته و تمایز یافته تنظیم می کند. هدف این مطالعه بررسی تنظیم بیان GCNF در سلول های ES تمایز یافته و تمایز نیافته است.

یک Dataset داده microarray از پایگاه داده GEO را به وسیله R2GEO آنالیز کنید و به سوالات زیر پاسخ دهید و به صورت گزارش کامل ارسال نمایید :

دیتاست انتخاب شده داده های بیان از سلول های ES تمایز نیافته و تمایز یافته انسانی است.

Dataset Title	Expression data from undifferentiated and differentiated human ES cells
GEO Accession	GSE76282
Platform	Affymetrix Human Genome U133 Plus 2.0 Array

1. چه دسته بندی برای این Dataset در نظر گرفته اید ؟ (حداقل دو دسته داده باید در نظر گرفته شده باشد به عنوان مثال مریض و سالم)

دیتاست شامل ۱۲ نمونه از سلول های hES با overexpression ژن GCNF است. دیتاست به دو دسته undifferentiated و differentiated گروه بندی شد (شکل ۱)

Samples		Define groups		Selected 12 out of 12 samples			
Group	Accession	Title	Source name	Cell line	Cell type	Treatment	Passages
undifferentiated	GSM1979041	hES at d0, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979042	hES at d0, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979043	hES at d0, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979044	hES+Dox at d0, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
undifferentiated	GSM1979045	hES+Dox at d0, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
undifferentiated	GSM1979046	hES+Dox at d0, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979047	hES+RA at d6, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979048	hES+RA at d6, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979049	hES+RA at d6, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979050	hES+Dox at d6, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979051	hES+Dox at d6, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979052	hES+Dox at d6, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35

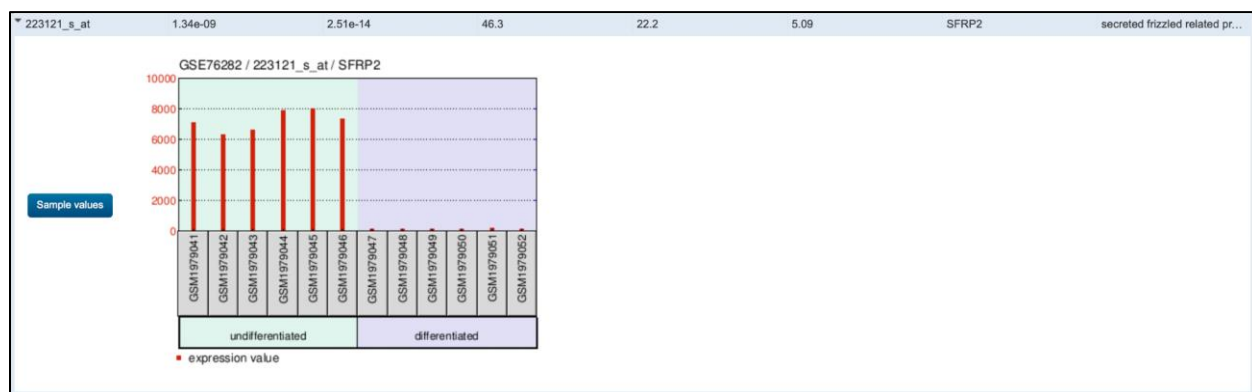
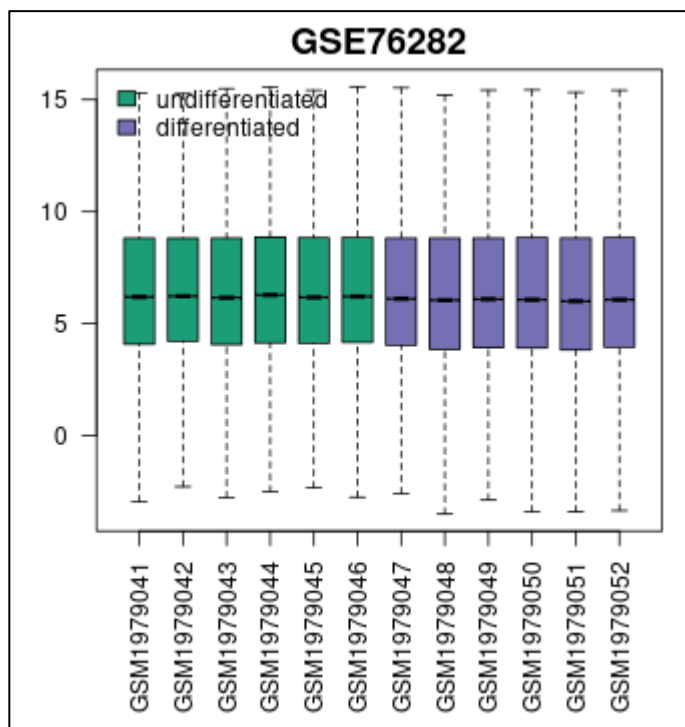
شکل 1: نمونه های گروه بندی شده دیتاست.

2. Differentially Expressed Gene (DEG) بین دسته ها را مشخص کنید. از چه cutoff هایی برای فیلتر داده استفاده کرده اید. بعد از فیلتر فایل مربوطه را ارسال نمایید.

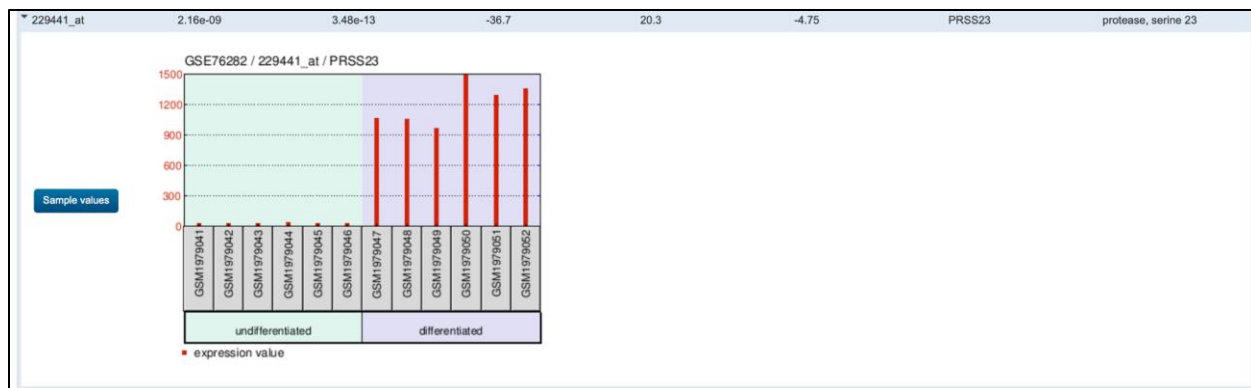
Boxplots برای تجسم توزیع مقادیر بیان ژن در هر نمونه

استفاده می‌شوند. از آنجا که نقاطی بالا و پایین خطوط ماکسیمم و مینیمم وجود ندارد بنظر میرسد داده پرت نداریم. نقاط پرت ممکن است مقادیر بیانی شدید یا غیرعادی را نشان دهند که به طور قابل توجهی از توزیع کلی منحرف می‌شوند. همچنین بنظر میرسد میانه بیان در همه گروه ها تقریباً یکسان است. میانه در مرکز نشان می‌دهد که گروه ها قابل مقایسه هستند. طول جعبه نشان دهنده گسترش یا تنوع بیان ژن در نمونه است. یک باکس وسیع‌تر، تنوع بیشتری را نشان می‌دهد. این فاکتور نیز میان گروه ها تفاوت چندانی ندارد.

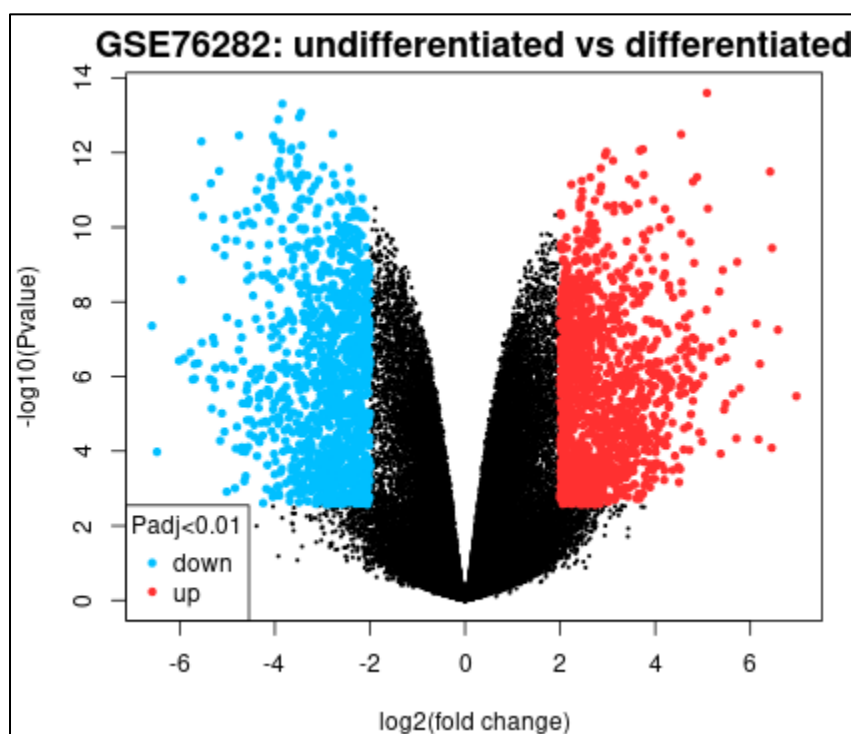
از آنجا که این نمودار برای تمام ژن های مورد نظر در نمونه ها کشیده شده است انتظار تفاوت فاحش ظاهری نداریم. اگر به دنبال تفاوت بیان ژن های مختلف در نمونه های تمایز یافته و نیافته هستیم، با کلیک روی تک تک ژن ها می‌توان این اطلاعات را بدست آورد (شکل ۲ و ۳)



شکل 2: ژن 2SFRP مثالی از یک ژن که در گروه سلول های تمایز نیافته بیان بیشتری داشته است



شکل 3: ژن 23PRSS مثالی از یک ژن که در گروه سلول های تمایز یافته بیان بیشتری داشته است.



شکل 4: volcano plot را برای ژن ها نشان می دهد. (برای بررسی هر یک از ژن ها به [این لینک](#) مراجعه کنید)

p-value نشان دهنده احتمالی است که تفاوت های مشاهده شده در بیان ژن بین دو گروه شانسی باشد. مقدار p کوچکتر نشان دهنده سطح بالاتری از معناداری آماری است. برای مثال اگر $p - value = 0.001$ آنگاه:

$$-\log_{10} 0.001 = 3$$

ولی اگر اگر $p - value = 0.1$ آنگاه:

$$-\log_{10} 0.1 = 1$$

پس هرچه significance بیشتر باشد ژن در بخش های بالایی محور y است. در volcano plot، مقدار LogFC بالاتر نشان دهنده تفاوت بیشتر در بیان ژن بین دو گروه مورد مقایسه است.

LogFC به عنوان نسبت $2\log$ سطح بیان ژن در یک شرایط (به عنوان مثال، سلول های بنیادی تمایز نیافته) به سطح بیان ژن در شرایط دیگر (به عنوان مثال، سلول های بنیادی تمایز یافته) محاسبه می شود. LogFC مثبت نشان می دهد که ژن در گروه اول نسبت به گروه دوم تنظیم مثبت شده است، در حالی که یک LogFC منفی نشان دهنده کاهش تنظیم است.

به عنوان مثال، اگر یک ژن دارای LogFC برابر با ۲ باشد، به این معنی است که بیان آن در گروه اول تقریباً چهار برابر بیشتر از گروه دوم است. به طور مشابه، مقدار LogFC برابر با -۱ نشان می دهد که بیان ژن در گروه اول تقریباً نصف گروه دوم است.

در بحث شناسایی ژن های بیان شده متفاوت، LogFC بالاتر نشان دهنده تغییر بیشتری در بیان ژن است. ژن های با مقادیر LogFC بالاتر معمولاً برای تجزیه و تحلیل و بررسی بیشتر انتخاب می شوند. با این حال، مهم است که معناداری آماری (p -value) را در کنار LogFC در نظر بگیریم تا اطمینان حاصل شود که تفاوت های مشاهده شده به دلیل تغییرات رندوم نیستند.

برای بررسی ژن ها و محدود کردن تعداد آنها به علاوه بالا بردن اطمینان در بررسی ها cutoff های مورد استفاده $2 < \text{LogFC}$ (ژن هایی با LogFC بالای ۲ و زیر -۲ حائز اهمیت هستند) و $0.01 > p\text{-adj}$ (ژن هایی با معناداری زیر ۰.۰۱ مهم هستند) در نظر گرفته شده اند. همانطور که در نمودار هم مشخص است، مقادیر x که بین ۲ و -۲ است و همچنین

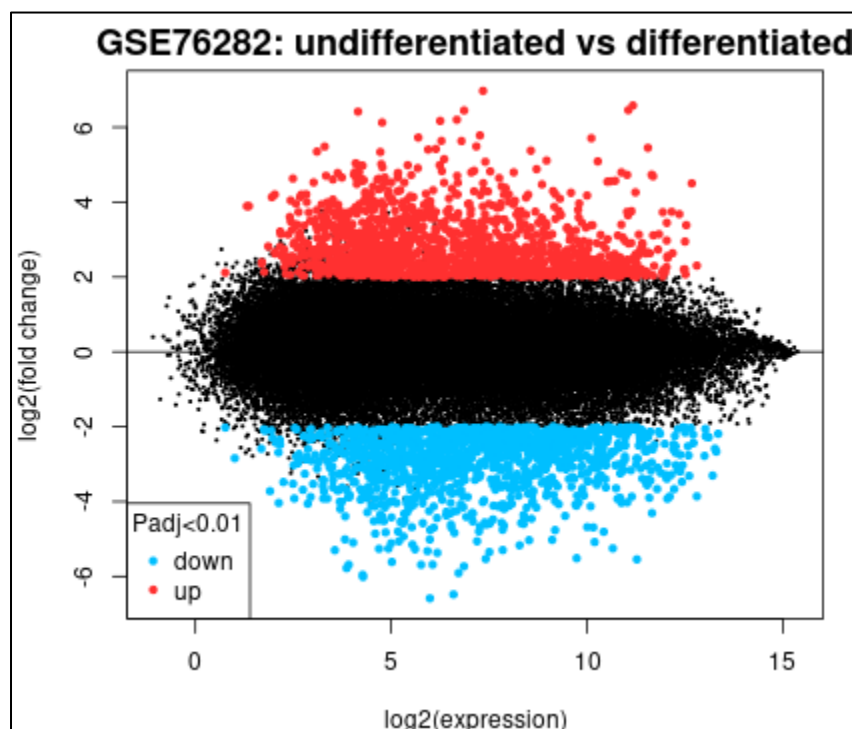
$$P\text{-adj} < 0.01$$

$$\log_2(0.01) < -2$$

$$-\log_2(0.01) > 2$$

$$Y > 2$$

به عنوان ژن هایی که از نظر بیان تفاوتی در دو گروه نداشته اند در نظر گرفته شده اند.



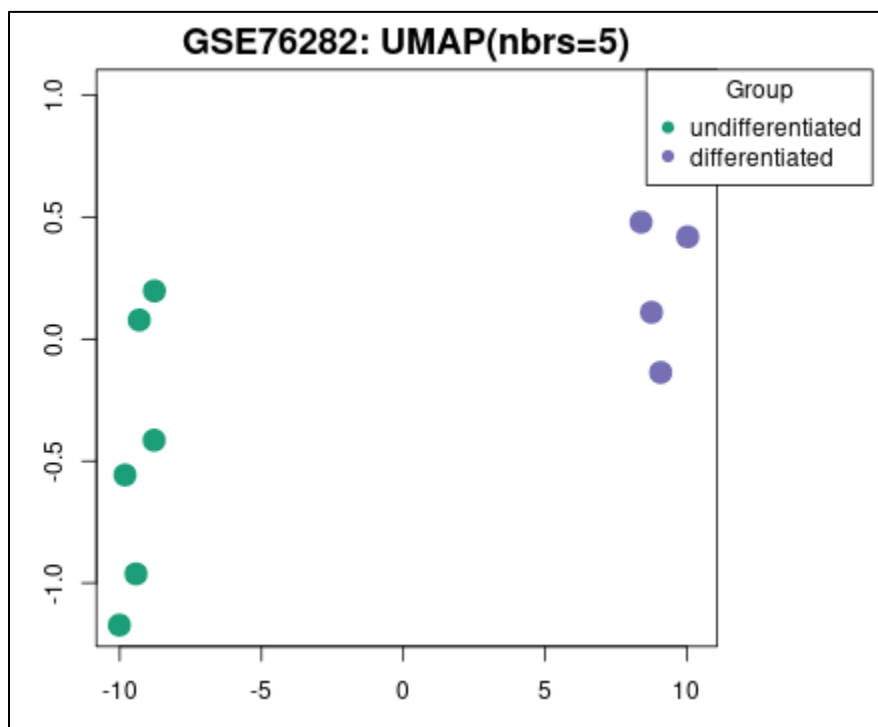
شکل 5: mean-difference plot را برای ژن ها نشان می دهد. (برای بررسی هر یک از ژن ها به [این لینک](#) مراجعه کنید)

ژن‌هایی که در اطراف خط مرکزی قرار گرفته‌اند ($\text{LogFC} = 0$) تفاوت‌های کمتری در بیان بین گروه‌ها دارند، در حالی که ژن‌هایی که دورتر از خط مرکزی قرار دارند، تغییرات بزرگ‌تری دارند. عرض توزیع می‌تواند میزان تغییرات کلی بیان ژن مشاهده شده در مجموعه داده را نشان دهد؛ در نمودار این میزان بین -6 و 6 است.

محور X میانگین سطح بیان یک ژن را نشان می‌دهد که می‌تواند به شناسایی ژن‌هایی با سطوح بیان کم یا زیاد در شرایط مورد مطالعه کمک کند. ژن‌هایی با سطوح بیانی خیلی بالا یا خیلی پایین در انتها و ابتدای محور X یافت می‌شوند.

اگر به ژن‌های مشکمی نگاه کنید، تمایل نقاط به سمت راست محور دیده می‌شود؛ این موضوع بیانگر این نکته است که ژن‌هایی با بیان کم، تنوع بیشتر در LogFC دارند در حالی که ژن‌هایی که بیان بیشتری دارند معمولاً به سمت ($\text{LogFC} = 0$) متمایل می‌شوند.

نمودار MA همچنین می‌تواند برای ارزیابی کیفیت داده‌ها و اثربخشی روش‌های نرمال‌سازی مفید باشد. توزیع یکنواخت نقاط داده در اطراف خط مرکزی نشان‌دهنده نرمال‌سازی خوب و حداقل سوگیری است. به نظر می‌رسد داده‌ها به صورت یکنواخت توزیع شده باشند بنابراین نیازی به Force normalization نیست.



شکل 6: UMAP plot را برای ژن‌ها نشان می‌دهد.

UMAP (Uniform Manifold Approximation and Projection) یک تکنیک کاهش ابعاد است که معمولاً برای تجسم و تجزیه و تحلیل داده‌های بیان ژن با ابعاد بالا، از DEGs استفاده می‌شود. UMAP می‌تواند اطلاعاتی درباره روابط، کلاسترها و الگوهای درون DEG ارائه دهد.

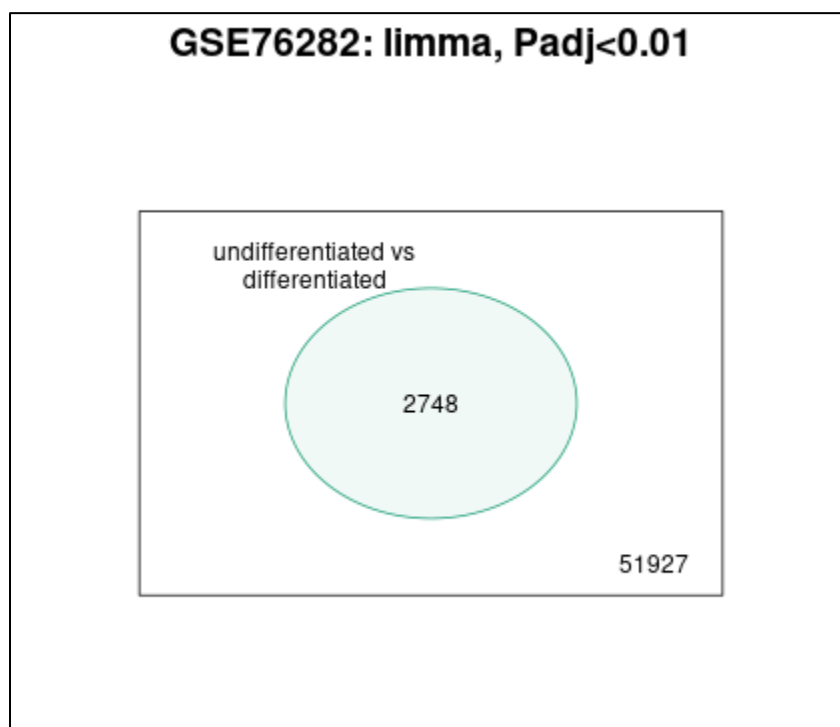
UMAP می‌تواند کلاسترها یا گروه‌هایی از DEG‌ها را با الگوهای بیان مشابه آشکار کند. با نمایش داده‌های بیان ژن با ابعاد بالا در فضایی با ابعاد پایین‌تر، UMAP می‌تواند مناطقی را شناسایی کند که DEG‌هایی با پروفایل‌های بیانی مشابه بیشتر در

آنها یافت می‌شوند. خوشه‌ها در نمودارهای بالا می‌توانند مازول‌ها یا مسیرهای عملکردی را نشان دهند که تحت تأثیر شرایط مورد مطالعه قرار می‌گیرند.

می‌توان دید که DEG‌ها به خوبی در کلاسترهایی جداگانه مرتبط با گروه بندی مد نظر قرار گرفته‌اند. هر یک از نقاط نمودار ژن‌هایی با بیان مشابه هستند که با در نظر گرفتن پارامتر همسایه‌ها در یک گروه قرار گرفته‌اند. $n_neighbours$ تعادل UMAP را در ساختار لوکال و گلوبال در داده‌ها کنترل می‌کند.

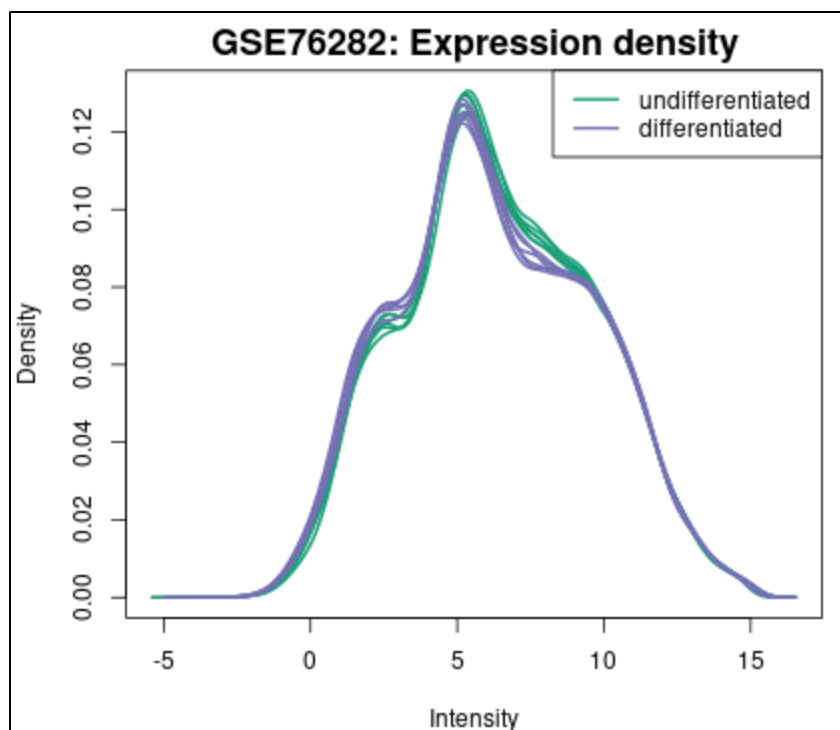
در نمودار بالا نقاط پرت وجود ندارد. نقاط پرت در نمودارهای UMAP می‌توانند ژن‌هایی را با الگوهای بیان منحصر به فرد یا نابجا در مقایسه با سایر DEG‌ها نشان دهند. این نقاط پرت ممکن است نشان دهنده ژن‌هایی با عملکردهای تخصصی را نشان دهد.

توجه داریم که UMAP یک تکنیک کاهش ابعاد غیر خطی است که هدف آن حفظ ساختارهای لوکال و گلوبال داده‌ها است. این یک نمایش با ابعاد پایین ایجاد می‌کند که در آن نقاط داده نزدیک در فضای با ابعاد بالا به موقعیت‌های نزدیک در نمودار UMAP پیش‌بینی می‌شوند. بنابراین مقادیر محور x و y در نمودارهای UMAP تفسیر بیولوژیکی مستقیمی ندارند.



شکل 7: Venn diagram را برای ژن‌ها نشان می‌دهد. (برای بررسی هر یک از ژن‌ها به [این لینک](#) مراجعه کنید)

نمودار ون یک نمایش گرافیکی است که معمولاً برای نشان دادن همپوشانی یا اشتراک بین مجموعه‌های مختلف ژن‌های بیان شده متفاوت (DEG) در گروه بندی‌های مختلف استفاده می‌شود. از این نمودار می‌توان DEG‌هایی که به طور خاص در یک گروه بیان بیشتر یا کمتری داشته‌اند را مشخص کرد. یا می‌توان نسبت DEG‌ها را در گروه‌های مختلف با همدیگر مقایسه کرد.



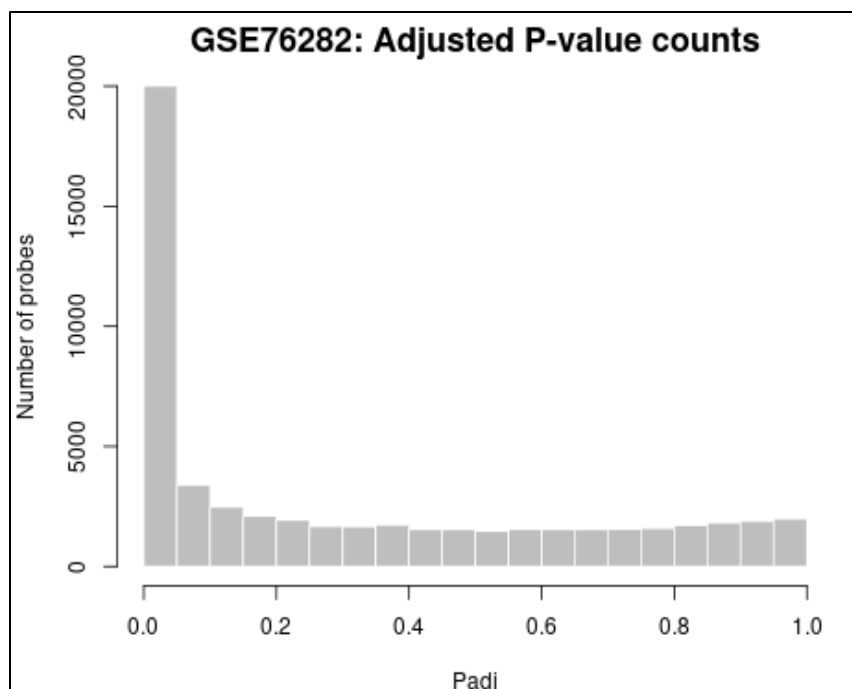
شکل 8: Expression density plot

در نمودار چگالی بیان، محور X مقادیر بیان ژن و محور Y نشان دهنده چگالی یا فراوانی وقوع آن مقادیر بیانی است. محور X مربوط به محدوده سطوح بیان ژن مشاهده شده در مجموعه ژنهای بیان شده متفاوت (DEGs) است. محور Y نشان دهنده چگالی یا فراوانی وقوع مقادیر بیان ژن است. این احتمال یافتن یک مقدار عبارت خاص را در DEG ها منعکس می کند. مقادیر محور Y غیر منفی هستند و چگالی نسبی یا فراوانی مقادیر بیان را در نقاط مختلف در امتداد محور X نشان می دهند.

با مشاهده این نمودار به نظر می رسد که بین گروه ها تقریباً پیک و الگوهای یکسان دارند. می توان گفت الگوهای بیان ژن بین گروه های مورد مقایسه مشابه است.

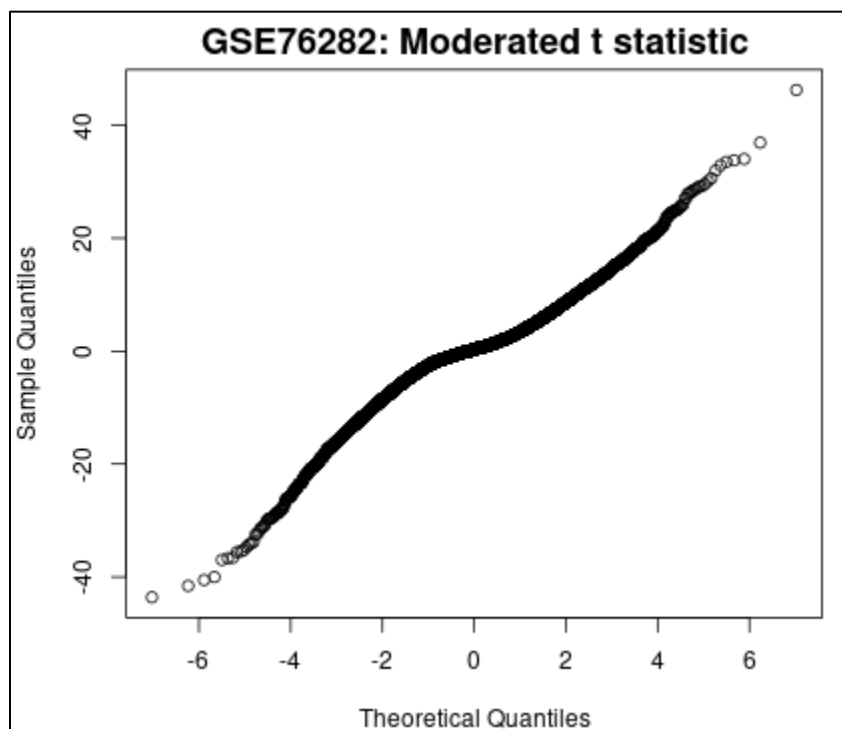
پیک ها و الگوهای ثابت نشان می دهد که ژن ها بین گروه ها مکانیسم های تنظیمی یکسان یا مشابه دارند. این می تواند نشان دهنده وجود فاکتورهای رونویسی مشترک، مسیرهای سیگنالینگ یا سایر عوامل تنظیمی باشد که بیان ژن را به شیوه ای هماهنگ در بین گروه ها تحت تاثیر قرار می دهد. همچنین می توان نتیجه گرفت که احتمالاً فرآیندها یا مسیرهای بیولوژیکی زیربنایی در هر دو گروه مشترک هستند یا دارای سطوح فعالیت مشابه دارند. بنابراین گروه ها مقایسه شده ممکن است ویژگی های عملکردی یا پاسخ های بیولوژیکی مشابهی داشته باشند. این فرایندهای بیولوژیکی مشابه ممکن است به علت وجود مولکولی های conserve و حفاظت شده باشند.

انتظار داریم نقاط پرت به صورت قله ها پهای مجزا ظاهر می شوند که از الگوی کلی منحنی چگالی منحرف می شوند. با توجه به نمودار به نظر می رسد بیان پرت وجود ندارد.



شکل 9: Adjusted p-value count plot

Null hypothesis است. به عبارتی مقادیر p زیادی وجود دارد که نتایج آماری معنی‌داری را نشان می‌دهد، که نشان‌دهنده تفاوت قوی بین بیان در گروه‌های سلول‌های تمایز یافته و تمایز نیافته است. بنابراین انتظار می‌رود توزیع یکنواخت مقادیر p در

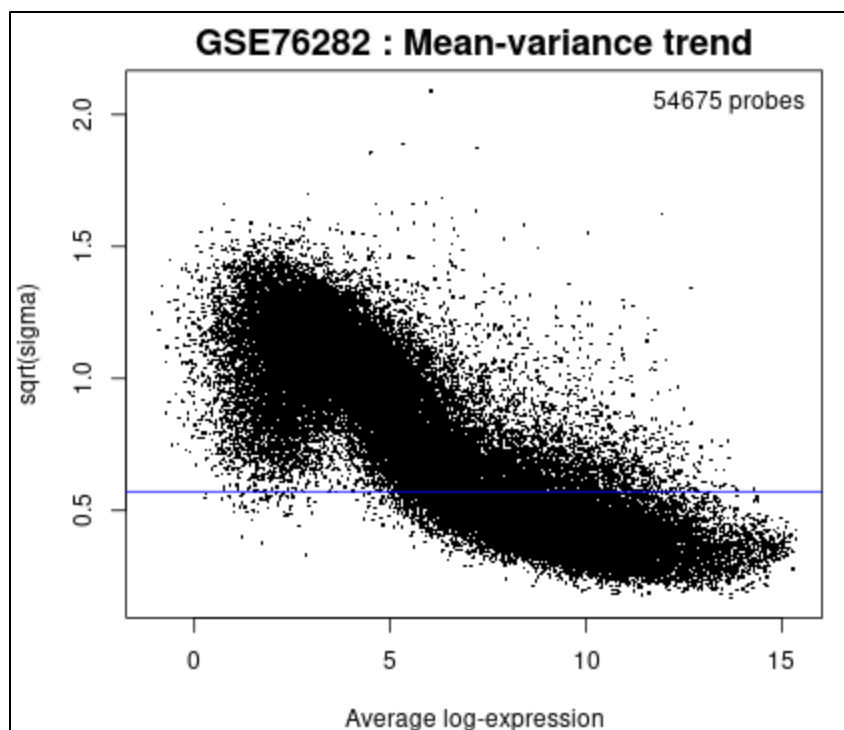


شکل 10: Moderated t statistic plot

نشان‌دهنده توزیع نرمال است. بنابراین می‌توان بیان کرد که دیتاست ما و بیان ژن‌ها تقریباً از حالت نرمال پیروی می‌کند.

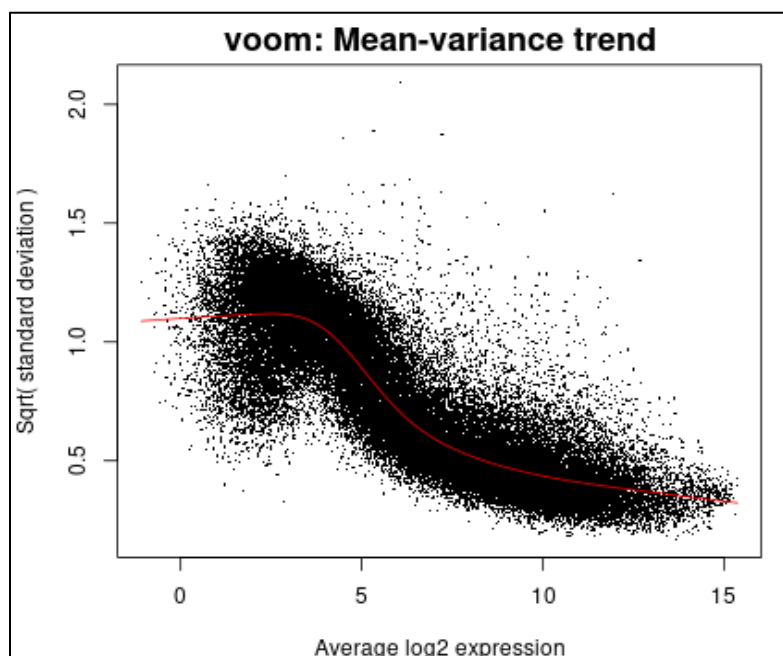
محور X مقدار adjusted p-value است. این مقدار کوچکترین سطح معناداری و false discovery rate است که در تست‌ها مقایسه‌ای چندگانه وجود دارد. به عبارتی در بررسی تفاوت بین ژن‌های مختلف در گروه‌های مختلف، میزان p-value تنظیم می‌شود. تجمع مقادیر p کوچکتر را در سمت چپ محور X دیده می‌شود. این نشان‌دهنده تعداد بیشتری از یافته‌ها یا شواهد معنا دار علیه

کل محدوده محور X فرض صفر را تایید می‌کند. آزمون t فرض می‌کند که سطح بیان ژن تقریباً از توزیع نرمال پیروی می‌کند. با فرض نرمال بودن امکان استفاده از تئوری آماری تثبیت شده، محاسبه p-value، فواصل اطمینان و آزمون فرضیه را ممکن می‌سازد. این معیارهای آماری به تعیین معناداری DEG بین گروه‌ها کمک می‌کند. در نمودار مقادیر t مشاهده شده را با مقادیر مورد انتظار با فرض توزیع نرمال قابل مشاهده است. اگر آماره t دقیقاً از یک خط مورب در نمودار پیروی کند،



Mean-variance trend plot : 11 شکل

آپشن precision weights در تحلیل DEG با هدف در نظر گرفتن روند میانگین واریانس مشاهده شده در نمودار M-A است. در این روش به ژن های خاص بر اساس سطوح بیان و تنوع آنها وزن های خاصی داده می شود. در شرایطی که روند واریانس و



Mean-variance trend plot : 12 شکل

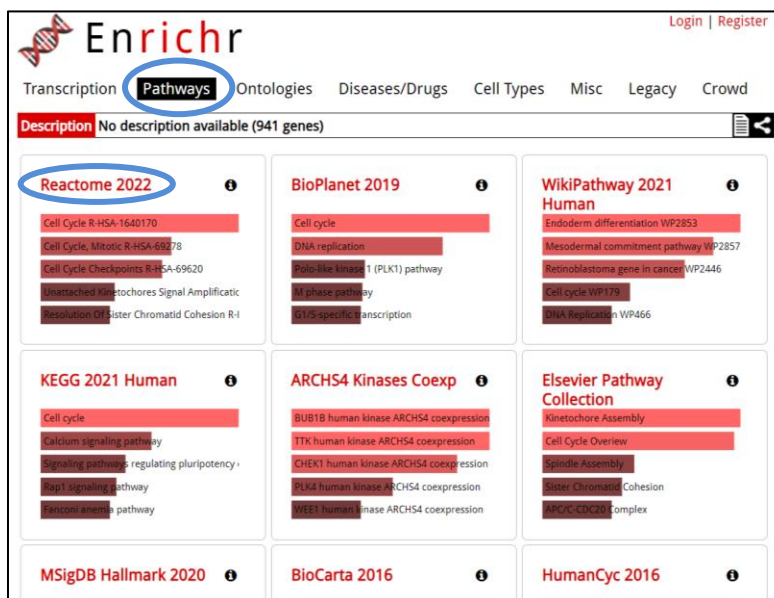
مقادیری که به طور قابل توجهی از خط مورب در نمودار QQ منحرف می شوند ممکن است نشان دهنده وجود مقادیر پرت باشد. به نظر می رسد مقادیر پرت وجود ندارد. در نمودار Mean-variance هر نقطه نشان دهنده یک ژن است. برای بررسی رابطه میانگین واریانس داده های بیانی، پس از فیت کردن یک مدل خطی استفاده می شود بنابراین نیازی به انتخاب گروه ندارد و نشان دهنده تنوع در داده هاست. این نمودار می تواند در ارزیابی استفاده یا عدم استفاده از گزینه وزن های دقیق (precision weights) در تحلیل داده ها کمک کند.

میانگین به هم وابسته باشند و رابطه واضحی داشته باشند (با افزایش میانگین واریانس به طور پیوسته زیاد یا کم شود)، این روش وزن بیشتری به ژن ها با الگوهای بیان متفاوت قابل توجه می دهد و می تواند دقت تشخیص DEG را بهبود بخشد. این کار کمک می کند تا در آنالیز به ژن هایی با بیان متفاوت و معنادار اهمیت بیشتری داده شود. بنظر میرسد روند واضح و قابل اعتمادی بین واریانس و میانگین دیده نمی شود بنابراین دادن وزن به ژن ها و انتخاب گزینه precision weights کمک چندانی به تشخیص DEGs نمی کند.

3. برای DEGs ها آنالیز enrichment انجام شود و نشان داده شود که در چه processes Biological، function Molecular و Pathway هایی نقش دارند.

شکل 13: ورودی سایت Enrichr

در این قسمت ژن های فیلتر شده بر اساس مقدار adjusted p-value (مقادیر کمتر از 0.01) و LogFC (مقادیر بالاتر از 2) را به ورودی سایت Enrichr میدهم (تعداد 1180 ژن). همچنین ژن هایی که اسم آنها در فایل دانلود شده از GEO ذکر نشده بود و یا چندین ژن یا miRNA نوشته شده بود حذف شدند. پس از سابمیت داده ها نتایج زیر حاصل می شود:



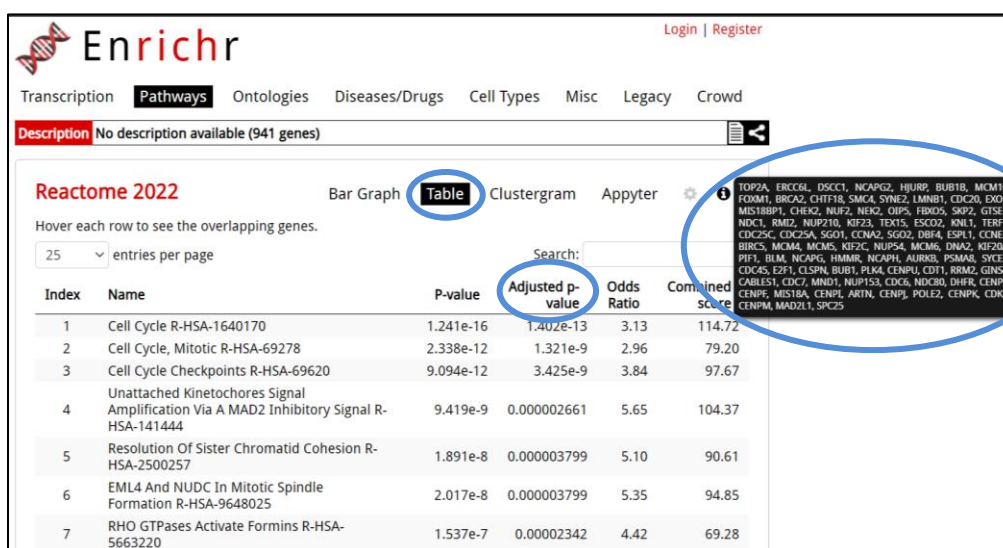
شکل 14: در بخش Pathway ها تعدادی مسیرهای بیولوژیک که این ژن ها در آن دخیل هستند نشان داده میشود (جهت دسترسی به این صفحه اینجا کلیک کنید)

اگر روی نتیجه اول که برای دیتابیس Reactome از سال 2022 است کلیک کنیم انواع Pathway هایی که ژن های ورودی در آنها ایفای نقش میکنند، مرتب شده بر اساس p-value نمایش داده میشوند.

Cell Cycle R-HSA-1640170
Cell Cycle, Mitotic R-HSA-69278
Cell Cycle Checkpoints R-HSA-69620
Unattached Kinetochores Signal Amplification Via A MAD2 Inhibitory Signal R-HSA-141444
Resolution Of Sister Chromatid Cohesion R-HSA-2500257
EML4 And NUDC In Mitotic Spindle Formation R-HSA-9648025
RHO GTPases Activate Formins R-HSA-5663220
Mitotic Spindle Checkpoint R-HSA-69618
Mitotic Prometaphase R-HSA-68877
POU5F1 (OCT4), SOX2, NANOG Activate Genes Related To Proliferation R-HSA-2892247

شکل 15: گراف مسیرهای بیولوژیک بر اساس ژن های ورودی در سایت Enrichr

همچنین می توان این نتایج را به صورت جدول هم مشاهده کرد.



Enrichr

Transcription Pathways Ontologies Diseases/Drugs Cell Types Misc Legacy Crowd

Description: No description available (941 genes)

Reactome 2022

Hover each row to see the overlapping genes.

25 entries per page

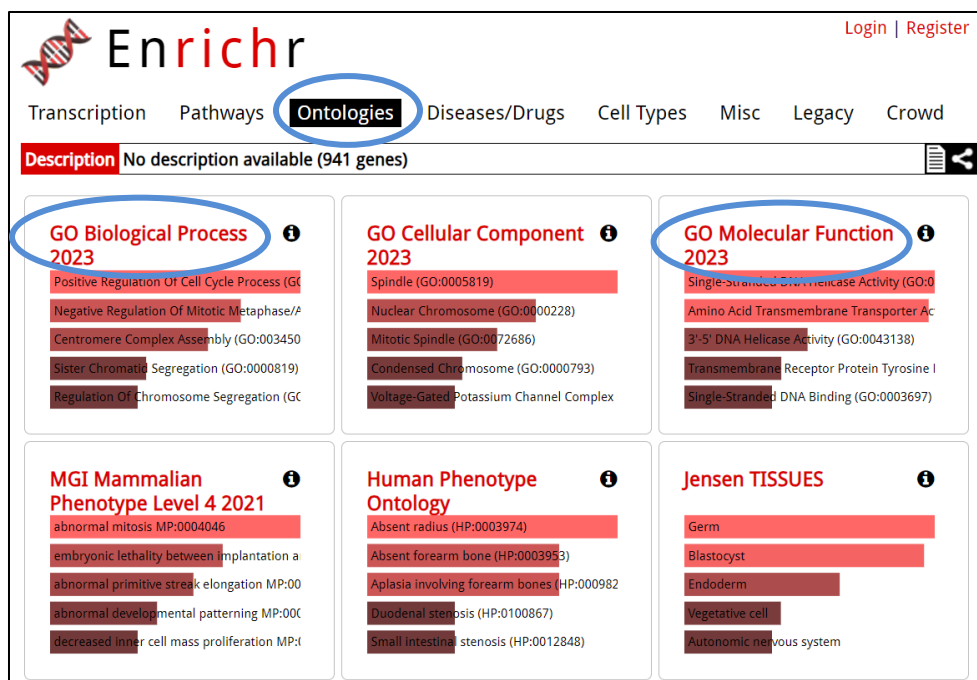
Search:

Index	Name	p-value	Adjusted p-value	Odds Ratio	Combined Score
1	Cell Cycle R-HSA-1640170	1.241e-16	1.402e-13	3.13	114.72
2	Cell Cycle, Mitotic R-HSA-69278	2.338e-12	1.321e-9	2.96	79.20
3	Cell Cycle Checkpoints R-HSA-69620	9.094e-12	3.425e-9	3.84	97.67
4	Unattached Kinetochores Signal Amplification Via A MAD2 Inhibitory Signal R-HSA-141444	9.419e-9	0.000002661	5.65	104.37
5	Resolution Of Sister Chromatid Cohesion R-HSA-2500257	1.891e-8	0.000003799	5.10	90.61
6	EML4 And NUDC In Mitotic Spindle Formation R-HSA-9648025	2.017e-8	0.000003799	5.35	94.85
7	RHO GTPases Activate Formins R-HSA-5663220	1.537e-7	0.00002342	4.42	69.28

Genes: TOP2A, ERCC1, DSCC1, NCAPG2, HUWR, BUB1B, MCM10, FOXM1, BRCA2, CHTF18, SMCA, SYNE2, LMNB1, CDC20, EXO1, MIST1BP1, CHEK2, NUF2, NEK2, OIP5, FBXO5, SKP2, GISE1, NDC1, RRM2, NUP210, KIF23, TEX15, ESCO2, KNL1, TERF1, CDC25C, CDC25A, SGO1, CCNA2, SGO2, DBF4, ESPL1, CENPE, BIRC5, MCM4, MCM5, KIF2C, NUP54, MCM6, DNA2, KIF20A, PIK1, BLM, NCAPG, HIMM1, NCAPH, AURKB, PSMA5, SYCE2, CDK5, E2F1, CLSPN, BUB1, PLMA, CENPA, CDT1, RRM2, GINS4, CABLES1, CDK7, MIND1, NUP153, CDCA, NDC80, DHFR, CENPE, CENPF, MIST1B, CENPL, ARTN, CENPL, POLE2, CENPK, CDK1, CENPM, MAD2L1, SPC25

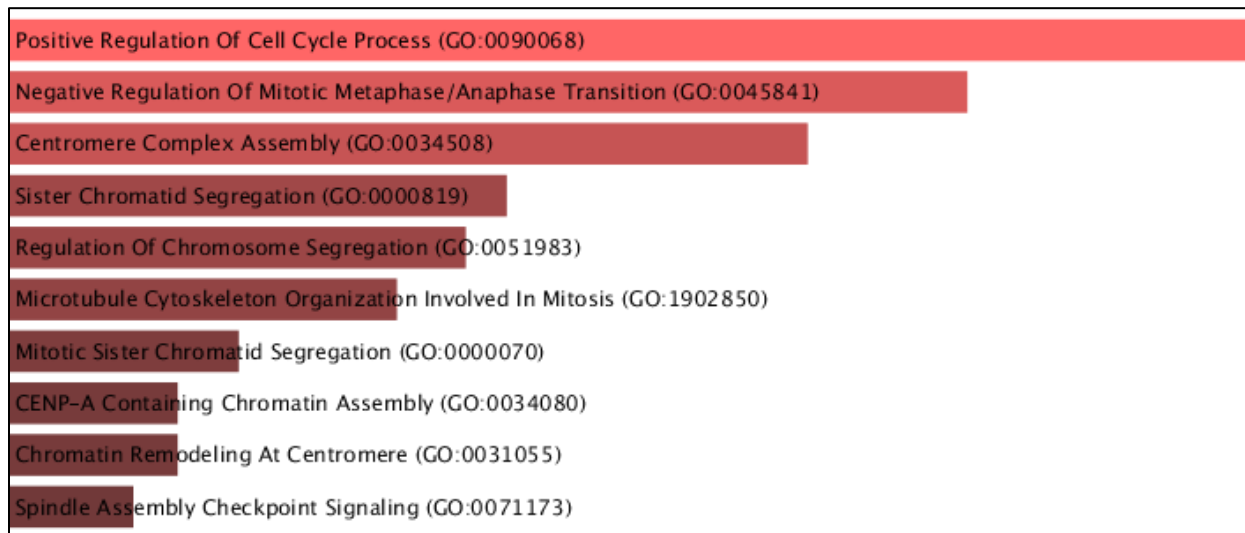
شکل 16: جدول مربوط به مسیر های بیولوژیک ژن های ورودی (جهت دسترسی به داده های کامل جدول اینجا کلیک کنید)

همانطور که در تصویر پیدا است، مسیرهای مختلف سلولی بر اساس مقدار p-value مرتب شده اند. هرچه مسیر p-value کمتری داشته باشد به این معنا است که داده های significant تری داریم بنابراین ژن های بیشتری از لیست ورودی به سایت در این مسیرها دخیل هستند. با نکه داشتن بر روی هر کدام از مسیر ها نیز ژن های آن را نمایش میدهد.



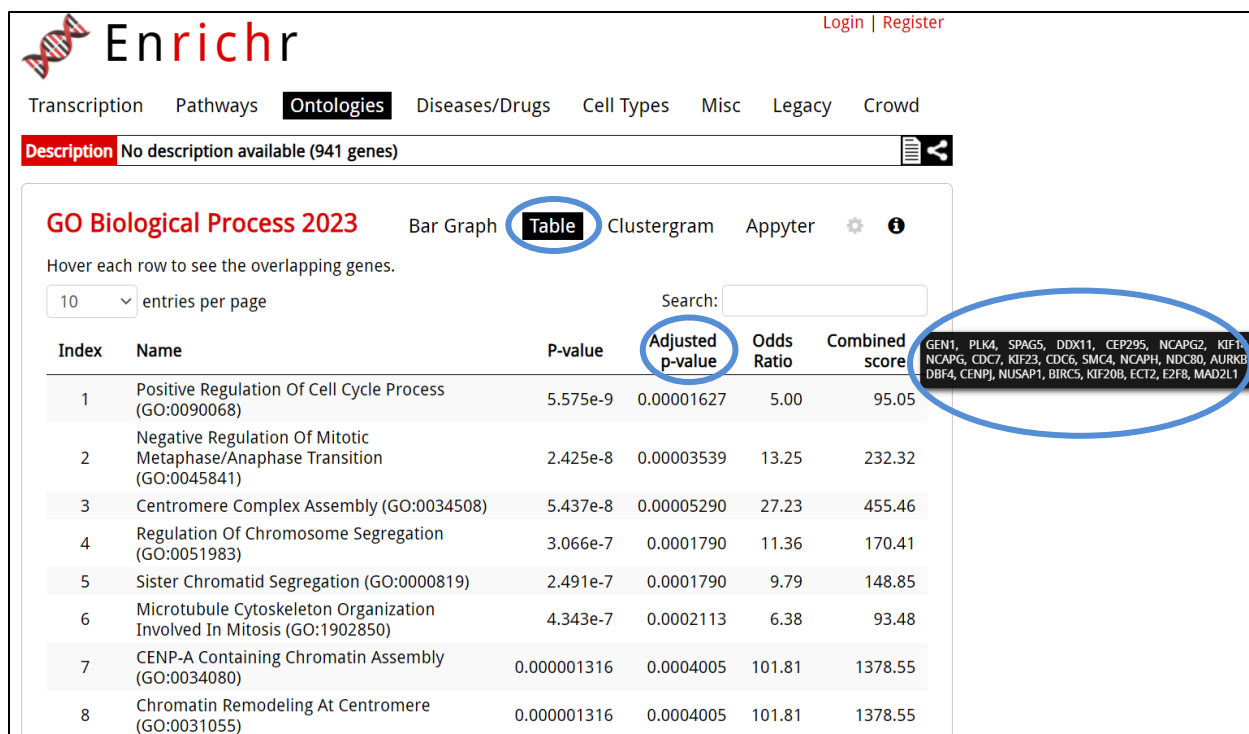
شکل 17: انواع داده های مربوط به بخش Ontologies

برای یافتن Biological Process و Molecular Function ژن های ورودی به بخش Ontology مراجعه میکنیم. در این بخش همانطور که در تصویر بالا نیز پیداست، داده های مربوط به Biological Process و Molecular Functions برای سال 2023 از دیتابیس Gene Ontology طبقه بندی شده اند که با کلیک بر روی هر کدام می توان جزئیات بیشتری از آنها بدست آورد.

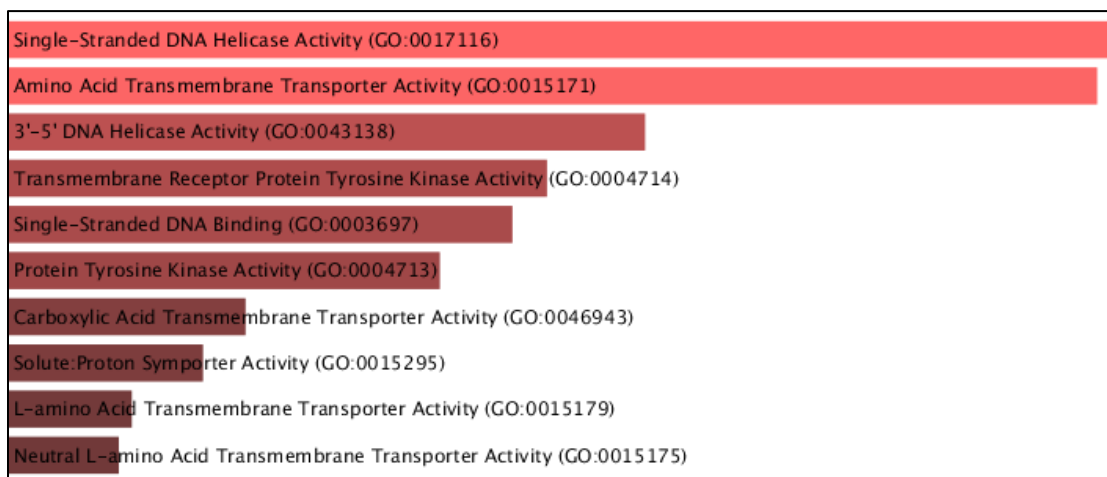


شکل 18: بار گراف Biological Process برای سال 2023 از دیتابیس GO

در این نمودار هر کدام از مسیر های بیولوژیکی بر حسب مقدار p-value مرتب شده اند به طوری که مسیر تنظیم مثبت چرخه سلولی Significant ترین داده و شامل بیشترین تعداد ژن ها است.



شکل 19: جدول مربوط به Biological Process ژن های ورودی (جهت دسترسی به داده های کامل جدول اینجا کلیک کنید) مشابه قسمت قبل، میتوان داده ها را در قالب یک جدول هم مشاهده کرد. اگر به هر سطر از جدول اشاره کنیم، ژن های دخیل در این مسیر زیستی را نمایش میدهد. در این جدول هم داده ها بر اساس p-value مرتب شده اند. برای بخش Molecular Functions هم مشابه Biological Process عمل میکنیم.



شکل 20: بار گراف Molecular Function برای سال 2023 از دیتابیس GO

شکل 16: جدول مربوط به مسیر های بیولوژیک ژن های ورودی (جهت دسترسی به داده های کامل جدول اینجا کلیک کنید)

Transcription Pathways **Ontologies** Diseases/Drugs Cell Types Misc Legacy Crowd

Description No description available (941 genes)

GO Biological Process 2023

GO Cellular Component 2023

GO Molecular Function 2023 Bar Graph **Table** Clustergram Appyter

Hover each row to see the overlapping genes.

10 entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Single-Stranded DNA Helicase Activity (GO:0017116)	0.00004569	0.01521	9.52	95.09
2	3'-5' DNA Helicase Activity (GO:0043138)	0.0003199	0.06462	11.31	90.99
3	Endodeoxyribonuclease Activity, Producing 5'-Phosphomonoesters (GO:0016888)	0.004866	0.2106	12.19	64.91
4	Succinate Transmembrane Transporter Activity (GO:0015141)	0.02011	0.2578	13.53	52.85
5	5'-Flap Endonuclease Activity (GO:0017108)	0.02011	0.2578	13.53	52.85
6	Thiamine Transmembrane Transporter Activity (GO:0015234)	0.02011	0.2578	13.53	52.85
7	Chondroitin Sulfotransferase Activity (GO:0004404)	0.02011	0.2578	13.53	52.85

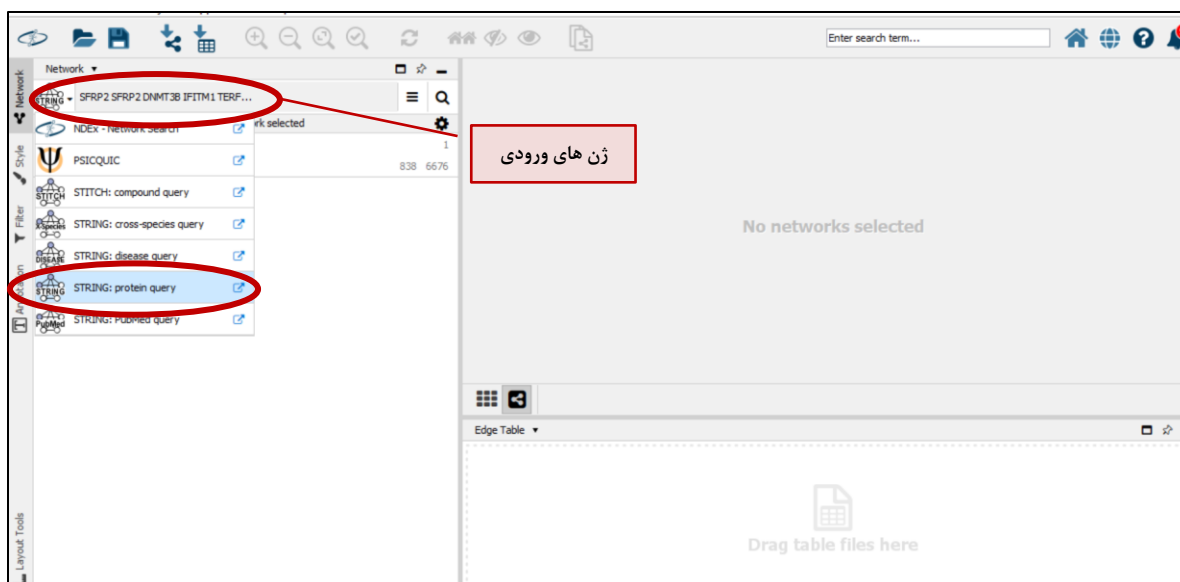
PIF1, POLQ, DSCC1, MCM4, MCM5, DNA2, MCM6

شکل 21: جدول مربوط به Molecular Function برای سال 2023 از دیتابیس GO

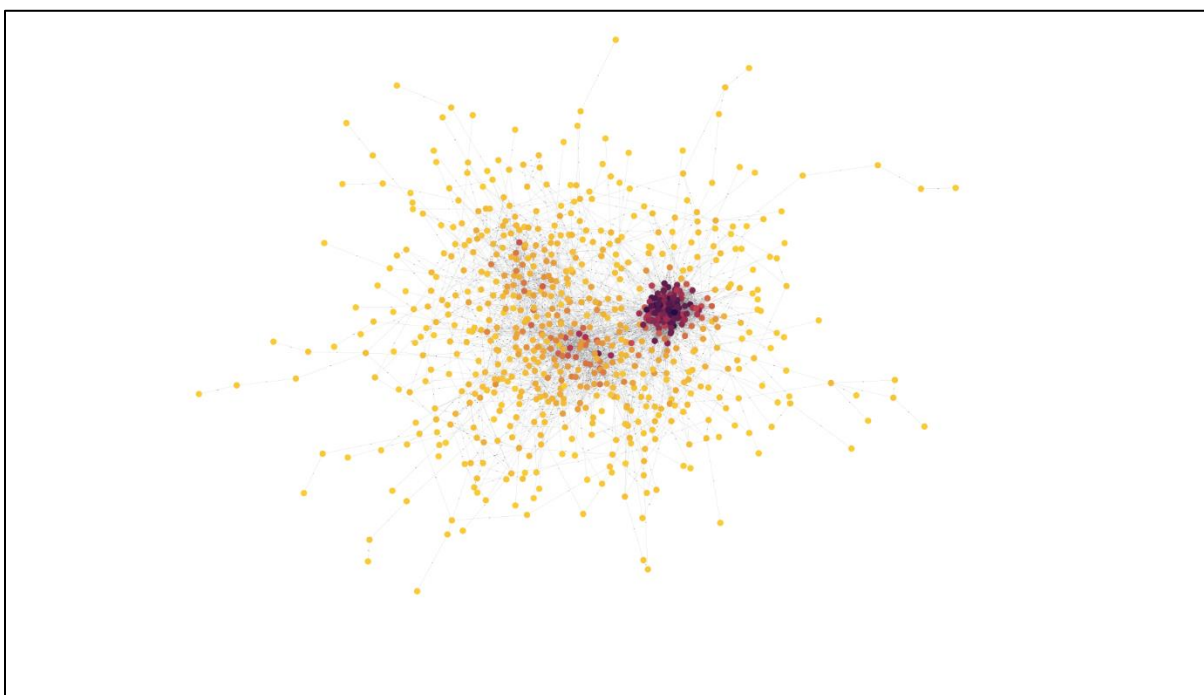
همانطور که در جدول بالا هم پیداست، مهمترین عملکرد این ژن ها، اثر بر فعالیت هلیکاز های DNA تک رشته ای است. جدول مربوط به داده های هرکدام از بخش های این سوال نیز در [این لینک](#) قرار داده شده است.

4. برای ژن های upregulated در یکی از گروه ها، شبکه protein-protein interaction و gene network regulatory را از پایگاه داده های مربوطه دانلود کرده و به وسیله Cytoscape رسم نمایید. لطفا گره ها را بر حسب درجه آمیزی نمایید و در مورد ژن هایی با بالاترین مقادیر درجه و Betweenness centrality در شبکه ها مقداری بحث کنید.

ژن هایی با LogFC مثبت نشان دهنده بیان بیشتر این ژن ها در دسته undifferentiated است. بنابراین با اعمال فیلتر در جدول مربوط به این فیلتر ها میتوانیم ژن های upregulate شده در این دسته را پیدا کنیم. به علت تعدد ژن ها و بررسی راحت تر داده ها، ژن های $\text{LogFC} > 2$ فیلتر شدند. (لیست ژن های فیلتر شده با استفاده از [این لینک](#) قابل دسترسی است) لیست ژن های بدست آمده را در بخش Search سایتواسکیپ وارد کرده و با استفاده از STRING: protein query شبکه Protein-Protein Interaction را رسم میکنیم.

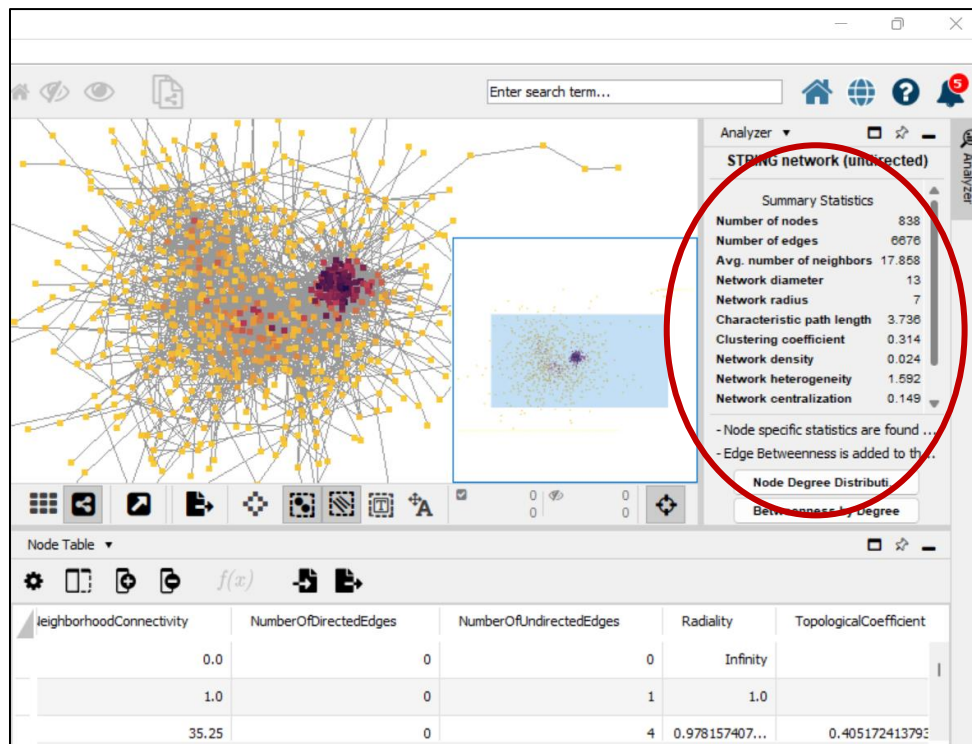


شکل 22: ورودی ژن های فیلتر شده به سایتواسکیپ با استفاده از Protein query دیتابیس STRING پس از دریافت شبکه مربوط به این ژن ها با استفاده از پنجره Style گره ها را بر حسب درجه آمیزی میکنیم. به این ترتیب گره های مربوط به ژنها با درجه بالاتر پررنگ تر از باقی ژن ها نمایش داده میشوند. جهت بررسی Betweenness نیاز به یک شبکه همبند داریم. بنا به دلایل مختلف از جمله نبود اطلاعات کافی درباره بعضی از ژن ها در دیتابیس STRING، برخی از گره ها در شبکه بدست آمده درجه صفر داشتند. با حذف این گره ها در نهایت شبکه همبندی به دست آمد که برای بررسی Betweenness مناسب است.



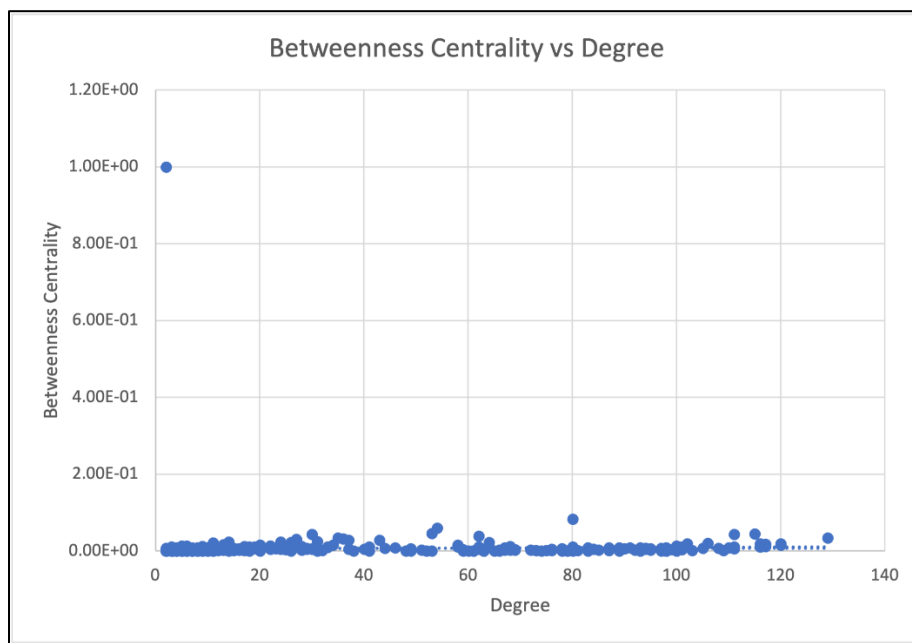
شکل 23: شبکه ppi رسم شده که گره ها بر اساس درجه رنگ آمیزی شده اند. (درجه بالاتر، پررنگ تر)

شبکه رسم شده دارای 838 گره و 6676 یال است. اطلاعات دیگر پس از آنالیز شبکه به شکل زیر نمایش داده می‌شود. همچنین جهت دسترسی به جدول گره‌های این شبکه پس از آنالیز، بر روی [این لینک](#) کلیک کنید.



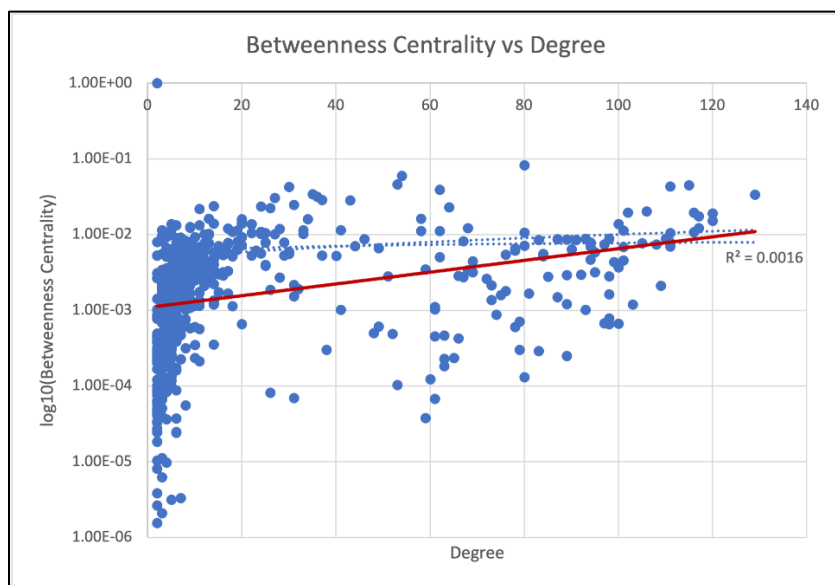
شکل 24: خلاصه آنالیز داده های شبکه ppi

جهت بحث در ارتباط با ژن هایی که بالاترین درجه را دارند و مقایسه آن ها بوسیله معیار Betweenness Centrality، لازم است تا نمودار درجه بر حسب Betweenness Centrality رسم شود.



شکل 25: Betweenness Centrality vs Degree plot

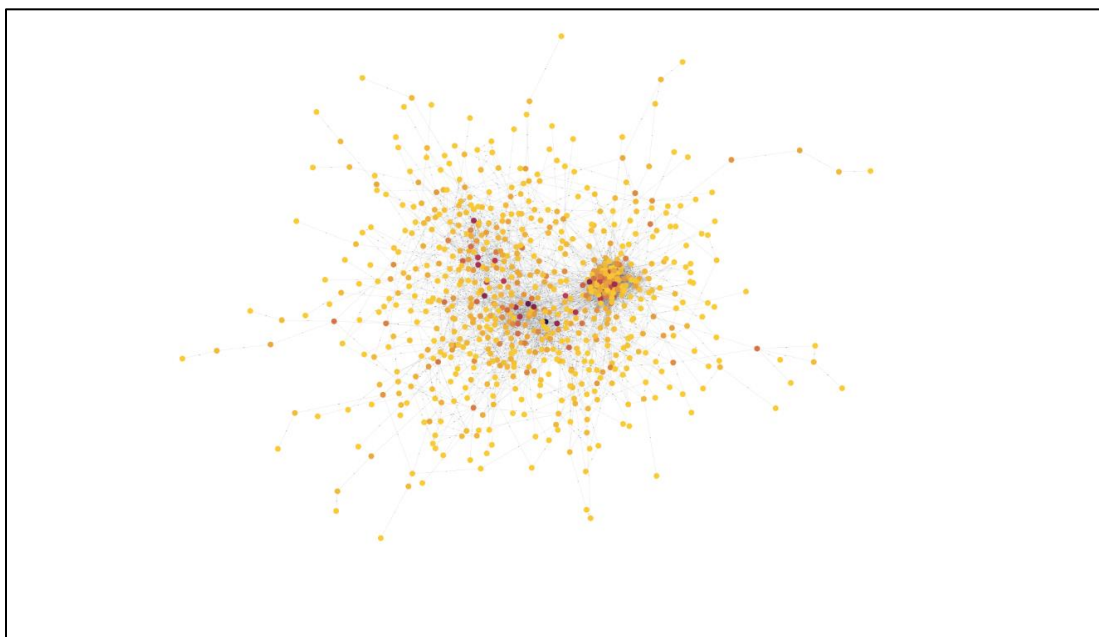
جدول بالا اطلاعات زیادی به ما نمیدهد زیرا Scale این دو معیار با هم هماهنگ نیستند. یکی از آنها مقادیر بالاتر از 100 را هم شامل میشود و یکی دیگر از معیارها مقادیر زیر 1 را شامل میشود. بنابراین برای درک بیشتر ارتباط بین این دو معیار در شبکه بدست آمده Betweenness Centrality را بر حسب لگاریتم بر پایه 10 رسم میکنیم تا شکل زیر حاصل شود.



شکل 26: $\text{Log}_{10}(\text{Betweenness Centrality})$ vs Degree plot

طبق شکل بالا نمی‌توان ارتباط خاصی بین این دو معیار در نظر گرفت؛ هرچند که به طور کلی می‌توان گفت با افزایش درجه مقدار Betweenness centrality افزایش می‌یابد.

همچنین می‌توان گره‌های شبکه را بر حسب Betweenness centrality رنگ کرد تا با مقایسه آن با رنگ آمیزی قبلی که بر حسب درجه بود به تحلیل ارتباط بین این دو معیار پرداخت.



شکل 27: رنگ آمیزی گره‌های شبکه ppi بر حسب مقدار Betweenness Centrality آنها

همانطور که در نمودار ها و تصاویر مربوط به رنگ آمیزی شبکه مشخص است ارتباط واضحی بین دو معیار Degree و Betweenness Centrality دیده نمیشود.

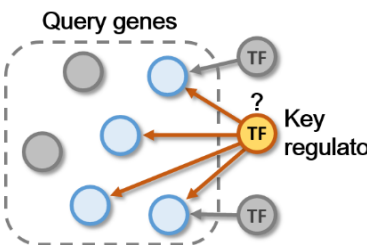
اگر Degree و Betweenness Centrality با یکدیگر همبستگی نداشته باشند، نشان می دهد که موقعیت گره ها از نظر اتصال و نفوذ در شبکه به طور مستقیم مرتبط نیست. در اینجا چند تفسیر احتمالی وجود دارد:

1. نقش ساختاری: گره های با درجه بالا دارای ارتباطات مستقیم زیادی هستند که نشان دهنده اهمیت آنها از نظر اتصال محلی است. از سوی دیگر، گره هایی با Betweenness Centrality بالا به عنوان پل یا واسطه در شبکه عمل می کنند و جریان اطلاعات را بین سایر گره ها تسهیل می کنند. وقتی این دو معیار همبستگی ندارند، نشان می دهد که گره هایی با درجه مرکزیت بالا ممکن است لزوماً در کوتاه ترین مسیر بین گره های دیگر قرار نگیرند یا تأثیر زیادی بر جریان اطلاعات نداشته باشند.

2. مدولار بودن شبکه: در برخی از شبکه ها، گره ها ممکن است ساختارهای مدولار یا خوشه ای را نشان دهند. در چنین مواردی، درجه، اتصال محلی را در ماژول ها اندازه گیری می کند، در حالی که Betweenness Centrality، اتصال بین ماژول را نشان می دهد. اگر شبکه دارای ماژول های متمایز با اتصالات محدود بین آنها باشد، گره های با مرکزیت درجه بالا در داخل ماژول ها ممکن است Betweenness Centrality بالایی نداشته باشند، زیرا آنها ماژول های مختلف را پل نمی کنند.

3. پویایی شبکه: همبستگی بین Degree و Betweenness Centrality می تواند تحت تأثیر ماهیت پویای شبکه باشد. در شبکه های در حال تکامل یا شبکه هایی با تغییر اتصالات در طول زمان، رابطه بین این دو معیار می تواند متفاوت باشد. گره هایی که دارای درجه بالایی در یک Snapshot از شبکه هستند، ممکن است با تکامل ساختار شبکه، Betweenness Centrality بالایی را حفظ نکنند.

برای رسم شبکه Gene Regulatory نیز از دیتابیس TRRUST استفاده میکنیم. این دیتابیس با ورودی گرفتن لیستی از ژن ها به ما مجموعه ای از مهمترین تنظیم کننده ها به همراه ژن های هدف آن ها را خروجی میدهد که با استفاده از این لیست می-توانیم شبکه تنظیم تنظیم ژن را در سایتواسکیپ رسم کنیم.



Query genes

Key regulator

2. Find key regulators for query genes

Submit a set of genes for a function/pathway/phenotype. (Min=5, Max=500)

Each gene name must be separated by comma, tab, white space or new line.
Input format: Entrez Gene ID (79923) or Gene Symbol (NANOG)

Species: ☒ Human ☐ Mouse

SFRP2
SFRP2
DNMT3B
IFITM1
TERF1
TERF1


Examples Submit Reset

Example gene sets

#1: 33 DEGs perturbed by ESR1 knockdown in human breast tumors.
[Muthukaruppan et al., Clin Breast Cancer, 2017](#)

شکل 28: ورودی دیتابیس TRRUST جهت پیدا کردن Key Regulators


پس از سابمیت ژن های ورودی، این دیتابیس لیست تنظیم کننده های کلیدی را خروجی میدهد. چون تعداد کل ژن های ما 1180 تاست و این دیتابیس بیشتر از 500 ژن نمیتواند ورودی بگیرد، 500 ژن نخست upregulated را ورودی میدهیم.



TRRUST

version 2

Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining



About TRRUST

Search

Download

Search result of candidate key regulators ↓

Valid query genes (362):

SFRP2, DNMT3B, IFITM1, TERF1, SEMA6A, USP44, FABP6, AASS, SCG3, TRIM71, TOX3, SYNE2, GALNT12, CXCL12, GPRC5B, SOX2, GLDC, FLVCR1, DNA2, FOXH1, FZD5, ZGRF1, SAMHD1, ARTN, PLP1, CDCA7, SMC4, EPHA1, BCL11A, DCAF12L1, BEND3, PRIMA1, PIF1, JARID2, CENPU, PTRPZ1, NEMP1, WDHD1, CHRM3, CNTNAP2, CDCA2,

Invalid query genes (44):

LOC645321, PWAR6, LOC101060391, TEX15, WBSCR17, JAKMIP2-AS1, FLJ45482, LOC729732, LINC00458, ZNF738, LINC00665, LINC00664, GABPB1-AS1, ADAMTS19, LINC00698, DGCR5, KCNMB2-AS1, PWAR5, TUNAR, ERVMER34-1, FAM201A, LINC01405, PXN-AS1, D21S2088E, ZNF678, LOC105372840, LOC101927746,

Query genes included in TRRUST (89):

ALPL, ASPM, BCL11A, BLM, BRCA2, C3, CALB1, CCL26, CD24, CDA, CDC25A, CDC6, CDCA7, CDK1, CHEK2, CHGA, CHST4, CNTNAP2, COX6A1, CR2, CRABP1, CXCL12, CXCL5, CXCR4, DBF4, DHFR, DNMT3B, EDNRB, EPB41L4B, ERBB3, EXO1, FABP6, FANCA, FDFT1, FGF19, FGFR1, FOXQ1, GABRB3, GDA, GDF3, HLA-DRA, HMMR, HSD11B2,

#	Key TF	Description	# of overlapped genes	P value	FDR
1	E2F1	E2F transcription factor 1	12	1.06e-05	0.000572
2	NANOG	Nanog homeobox	5	6.39e-05	0.00143
3	POU5F1	POU class 5 homeobox 1	5	7.94e-05	0.00143
4	ARID3A	AT rich interactive domain 3A (BRIGHT-like)	3	0.000134	0.00181
5	SOX2	SRY (sex determining region Y)-box 2	4	0.000741	0.008

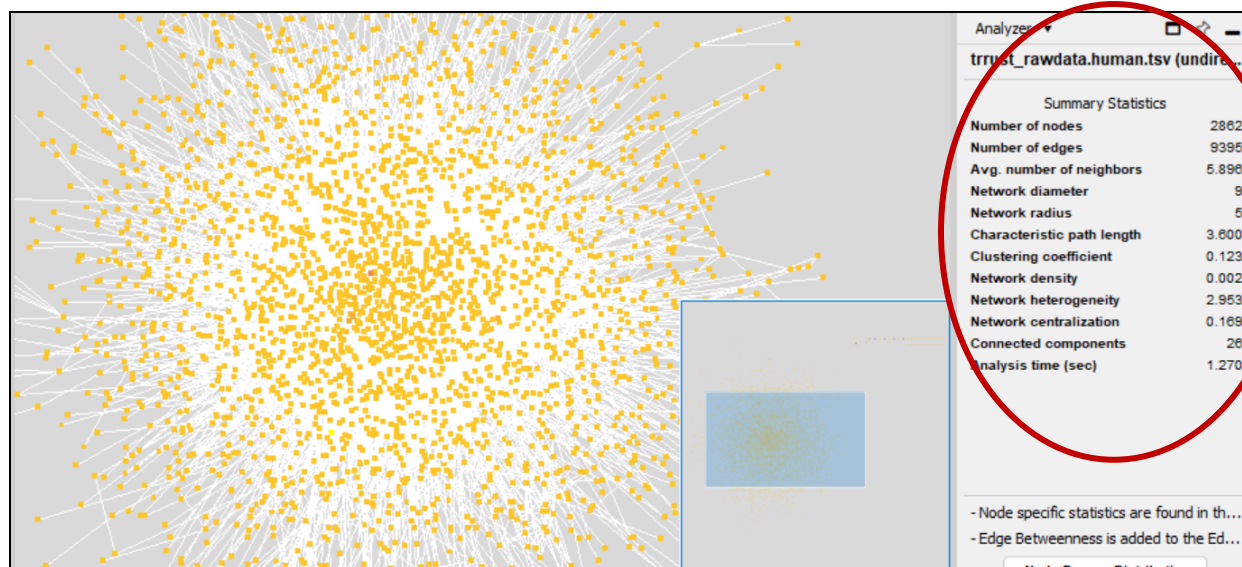
شکل 29: جدول تنظیم کننده های خروجی (از طریق این لینک جدول کامل Key Regulators قابل مشاهده است)

اگر روی هر کدام از این تنظیم کننده ها کلیک کنیم، لیستی از ژن های مورد هدف این تنظیم کننده به همراه نوع تنظیم کنندگی و همچنین Gene Ontology ژن های هدف را نشان خواهد داد.

TF	Target	Mode of Regulation	References (PMID)	Gene Ontology biological process of target genes
E2F1	CDC6	Unknown	18541154	DNA replication checkpoint; regulation of cyclin-dependent protein serine/threonine kinase activity; G1/S transition of mitotic cell cycle; regulation of transcription involved in G1/S transition of mitotic cell cycle; mitotic cell cycle; DNA replication;
E2F1	CDCA7	Activation	16580749	regulation of cell proliferation
E2F1	CDK1	Activation	14618416	G2/M transition of mitotic cell cycle; activation of MAPK activity; microtubule cytoskeleton organization; DNA replication; DNA repair; protein phosphorylation; DNA damage checkpoint; G2/M transition of mitotic cell cycle; replicative cell aging; double-strand break repair; regulation of transcription, DNA-templated; protein phosphorylation;
E2F1	CHEK2	Activation	15024084	G1/S transition of mitotic cell cycle; DNA replication
E2F1	DBF4	Activation	18503552	regulation of transcription involved in G1/S transition of mitotic cell cycle; tetrahydropterin biosynthetic process; negative regulation of translation; axon regeneration; response to methotrexate; dihydrofolate metabolic process;
E2F1	DHFR	Activation	14618416	DNA repair; protein complex assembly; interstrand cross-link repair
E2F1	FANCA	Activation	18438432	MAPK cascade; skeletal system development; neuron migration; protein phosphorylation; positive regulation of cell proliferation; fibroblast growth factor receptor signaling pathway;
E2F1	FGFR1	Activation	19303924	

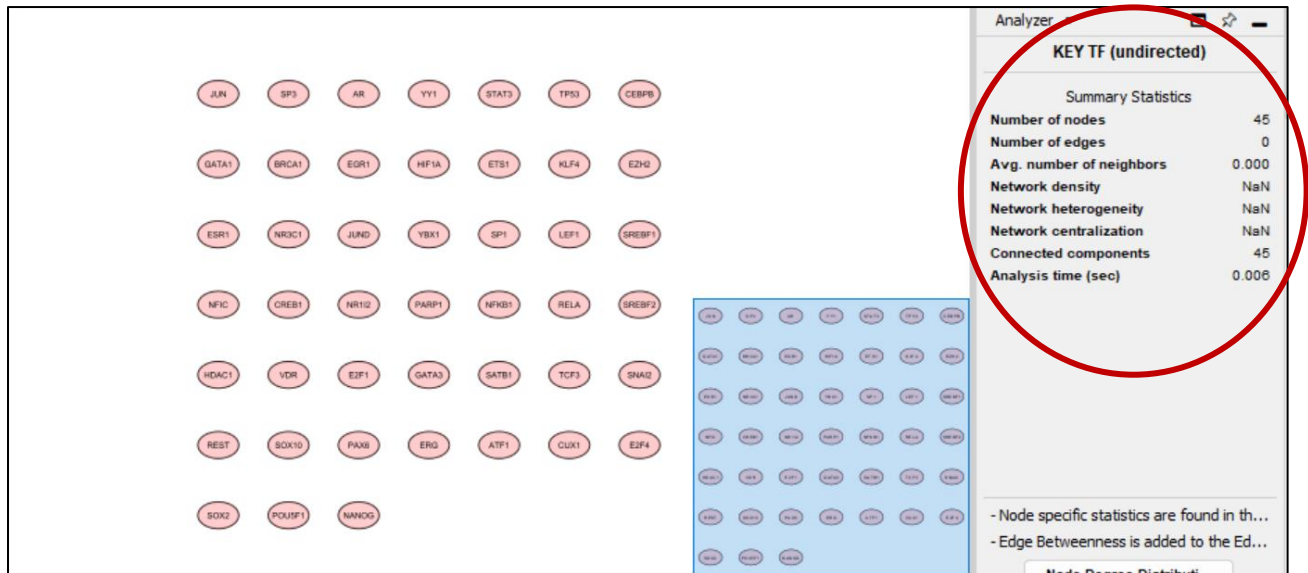
شکل 30: جدول مربوط به ژن های هدف تنظیم کننده E2F1

برای رسم شبکه تنظیم ژن نیز با توجه به عدم دسترسی به دیتابیس Regnetwork، ابتدا شبکه کامل Transcription Factor ها و ژن های هدف آن ها در انسان را از دیتابیس TRRUST دانلود کرده و شبکه آن را در سایتواسکیپ رسم میکنیم.



شکل 31: تصویری از شبکه کامل تنظیم ژن در انسان به همراه آنالیز شبکه (رنگ آمیزی گره ها بر اساس Betweenness Centrality صورت گرفته است)

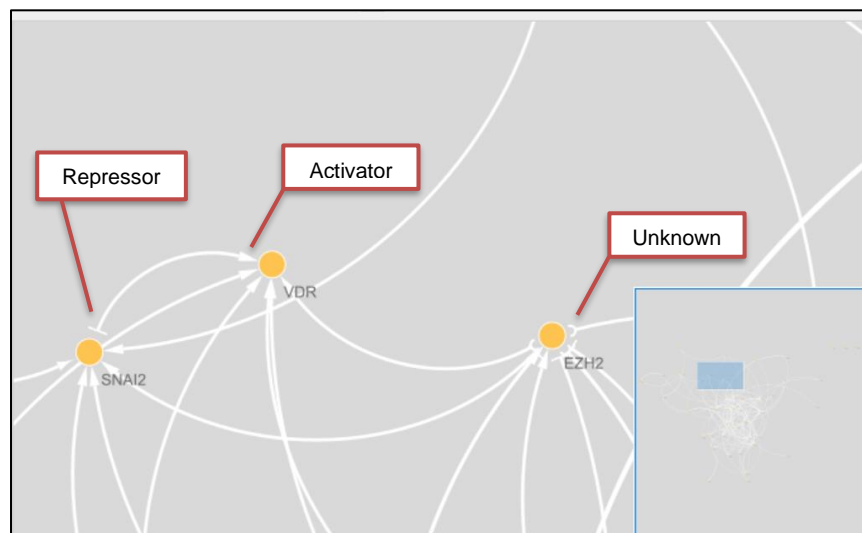
سپس شبکه مربوط به ارتباط Transcription Factor های کلیدی که از TRRUST خروجی گرفتیم را به وسیله سایتواسکیپ رسم میکنیم.



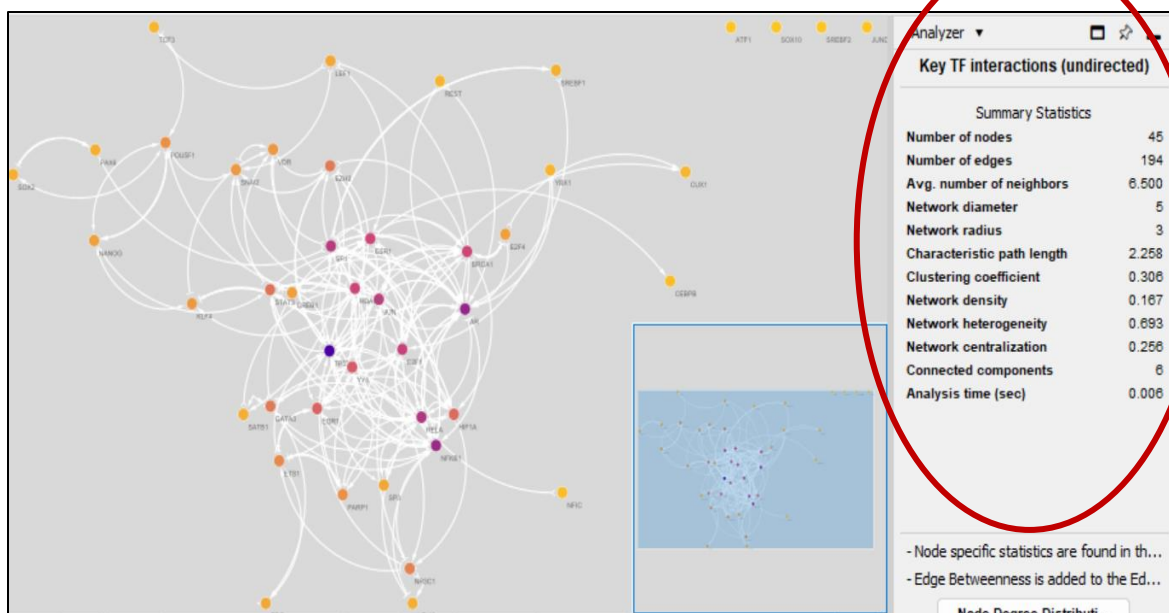
شکل 32: تصویر شبکه ارتباطی Key Transcription Factors به همراه آنالیز شبکه

پس از رسم این دو شبکه، شبکه کلی تنظیم ژن انسان با شبکه تنظیم کننده های کلیدی را Merge میکنیم تا ارتباط دقیق تر بین این تنظیم کننده ها با هم را بررسی کنیم. همانطور که در شبکه تنظیمی کاملی که برای انسان مشاهده کردیم، سه نوع یال در این شبکه های تنظیمی دیده می شود:

1. یال هایی که به یک کمان منتهی شده اند نشان دهنده اثر Activating دو گره بر هم است
2. یال هایی که به یک خط صاف منتهی شده اند نشان دهنده اثر Repressor دو گره بر هم است
3. یال هایی که به یک نیم دایره منتهی شده اند اثر و فعالیتشان ناشناخته است (Unknown)

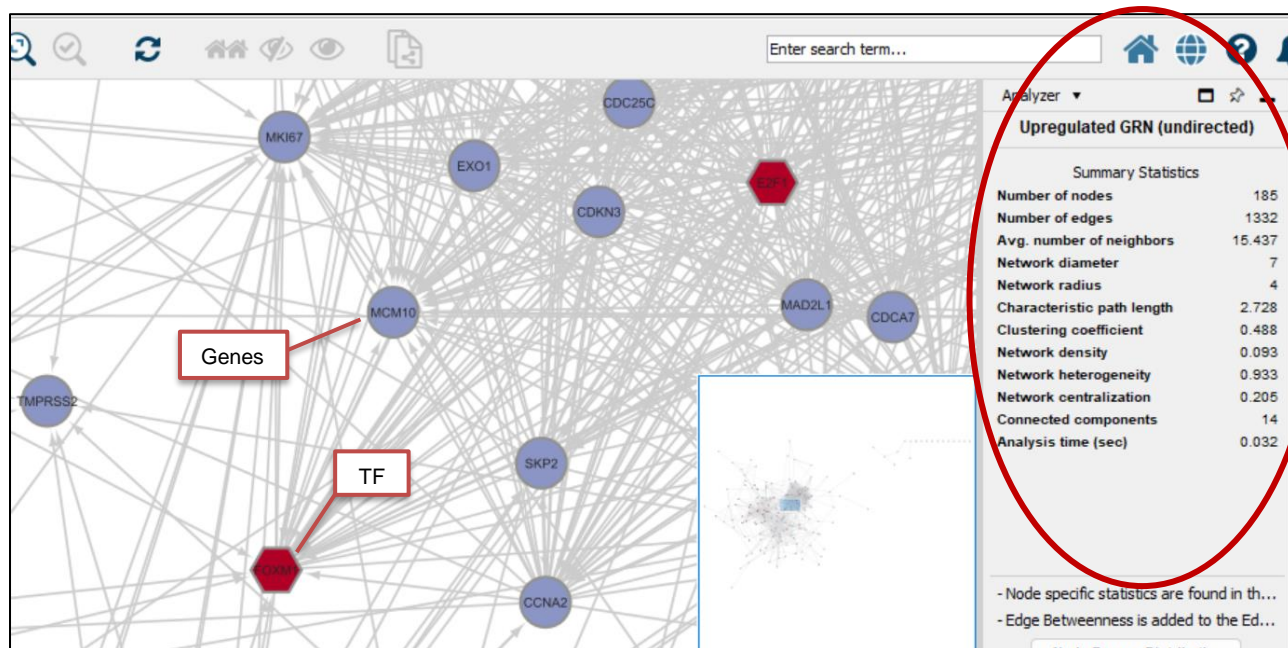


شکل 33: تصویر بخشی از شبکه Merge شده که شامل هر سه نوع یال می باشد.



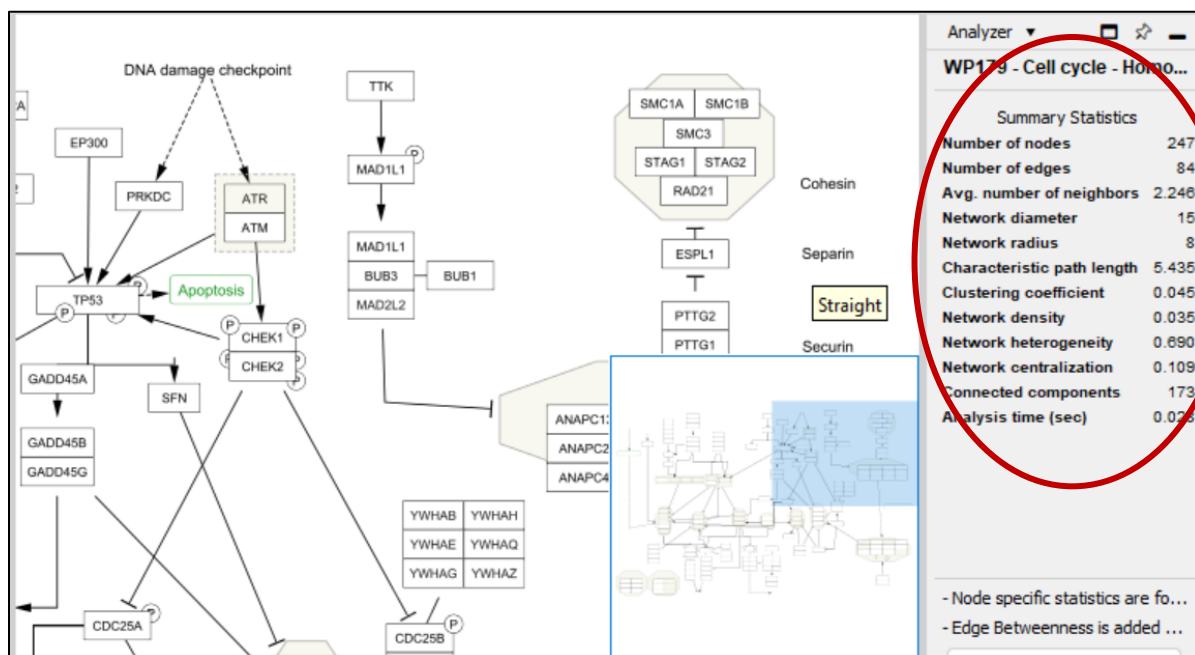
شکل 34: شبکه ارتباط بین TF ها باهم که حاصل Merge شبکه تنظیم ژن انسان و لیست Key TF ها از دیتابیس TRRUST است. گره ها در این شبکه بر اساس درجه رنگ آمیزی شده اند همچنین آنالیز شبکه نیز قابل مشاهده است.

در مرحله بعدی برای رسم GRN شبکه ppi ژن های upregulate شده که پیشتر رسم کردیم را با این شبکه بدست آمده Merge میکنیم تا رابطه بین این TF با ژن های هدف شان را بدست آوریم. نکته قابل توجه این است که در این شبکه نه تنها ارتباط بین TF باهم و TF بر ژن های هدف شان به تصویر کشیده شده، بلکه رابطه تنظیمی بین محصولات این ژن ها هم به نمایش درآمده است تا شبکه کاملی از تنظیم ژن در ارتباط با این دیتاست بدست آوریم.



شکل 35: تصویر از نمای نزدیک شبکه نهایی تنظیم ژن حاصل از دیتاست مورد بررسی به همراه آنالیز شبکه. در این شبکه TF ها با شکل و رنگ متفاوت از ژن ها جدا شده اند. ارتباط بین ژن ها و TF ها نیز با سه نوع یال مشخص شده است.

یکی دیگر از راه های فهمیدن رابطه تنظیمی بین Transcription Factor ها و ژن ها هدف آنها، استفاده از شبکه های آماده ای است که در خود سایتواسکیپ موجود است. برای این کار نخستین TF در جدول Key regulator ها که E2F1 است را با ژن های هدفش در Network Search سایتواسکیپ جستجو میکنیم. از بین شبکه های موجود، شبکه مربوط به Homo Sapiens Cell Cycle را Import میکنیم. طبق عملکرد هایی که در سوال قبلی از دیتابیس Enrichr استخراج کردیم همچنین جدول مربوط به Key Regulators و ژن های هدفشان و فعالیت این ژن ها، میتوان متوجه شد اکثریت ژن های دخیل در این دیتاست فعالیتی مرتبط با چرخه سلولی و تقسیم سلول را کنترل میکنند. بنابراین Regulatory Network که مرتبط با چرخه سلولی در انسان باشد میتواند اطلاعات خوبی از دیتاست مدنظر ما به ما ارائه میدهد.



شکل 36: نمایی نزدیک از شبکه تنظیمی چرخه سلولی در انسان به همراه آنالیز شبکه

مجموعه کامل شبکه های تنظیمی مورد استفاده در این سوال از طریق [این لینک](#) قابل دسترس است.