



دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر
پروژه درس مقدمه‌ای بر بیوانفورماتیک

عنوان:

تحلیل و شناسایی نشانگرهای بیولوژیکی سرطان سینه

نگارش

عماد امام جمعه ۴۰۰۱۰۸۷۷۴

محمد لطفی ۴۰۰۱۰۵۲۲۸

امیرحسین کوچکیان ۴۰۰۱۰۵۱۹۹

۱۴۰۳ بهمن

چکیده: سرطان سینه یکی از جدی‌ترین بیماری‌های بدخیم در زنان است که به دلیل ناهمگونی زیستی و مقاومت درمانی، چالش‌های بزرگی در تشخیص و درمان ایجاد کرده است. روش‌های فعلی، که عمدتاً به بیومارکرهای^۱ مانند PR، ER و HER2 متکی هستند، در برخی زیرگروه‌های این سرطان، به ویژه سرطان سینه سه‌گانه منفی (TNBC)، کاربرد محدودی دارند و نمی‌توانند به طور دقیق پیش‌آگهی یا پاسخ به درمان را پیش‌بینی کنند.

در این مطالعه، برای شناسایی بیومارکرهای جدید، از تحلیل بیان ژنی RNA-Seq و یادگیری ماشین استفاده شده است. ابتدا، ژن‌های دارای بیان افتراقی بین نمونه‌های سرطانی و طبیعی شناسایی شده و سپس، با یک فرآیند انتخاب ویژگی دقیق، مهم‌ترین ژن‌های مرتبط با سرطان استخراج شدند. نقش عملکردی این ژن‌ها در مسیرهای زیستی کلیدی بررسی شده و تأثیر آن‌ها بر ویژگی‌های بالینی و میزان بقای بیماران ارزیابی شده است. یافته‌های این پژوهش می‌توانند به بهبود روش‌های تشخیصی و طراحی درمان‌های شخصی‌سازی شده برای سرطان پستان کمک کنند.

۱ مقدمه

سرطان سینه یکی از شایع‌ترین بدخیمی‌های زنان است که ناشی از تغییرات ژنتیکی، اپی‌ژنتیکی و محیطی در سطح سلولی و مولکولی است. در شرایط عادی، رشد و تقسیم سلولی توسط یک شبکه‌ی پیچیده از مسیرهای سیگنالینگ کنترل می‌شود که شامل تنظیم چرخه‌ی سلولی، مکانیسم‌های ترمیم DNA و تنظیم بیان ژنی توسط فاکتورهای رونویسی است. یکی از دلایل اصلی ایجاد سرطان، اختلال در این مکانیسم‌های تنظیمی است که می‌تواند منجر به افزایش تکثیر سلولی، مهار آپوپتوز، و ایجاد توانایی تهاجم و متاستاز شود [۱].

در سطح ژنتیکی، جهش‌های سوماتیک و تغییرات کپی‌تعداد از مهم‌ترین عوامل سرطان‌زاوی هستند. برخی از این تغییرات به افزایش بیان پروتئین‌های تنظیمی چرخه‌ی سلولی منجر می‌شوند که باعث تقسیم مداوم و غیرقابل‌کنترل سلول‌ها می‌شود. برای مثال، افزایش بیان سیکلین D1 که ناشی از تغییر در ژن CCND1 است، می‌تواند منجر به فعالیت بیش از حد کمپلکس CDK4/6 و در نتیجه، افزایش ورود سلول به فاز S چرخه‌ی سلولی شود، که یکی از ویژگی‌های اصلی تومورزاوی است [۲]. همچنین، در برخی موارد، تنظیم نادرست فاکتورهای رونویسی باعث افزایش بیان ژن‌های مرتبط با تکثیر سلولی و مهار مسیرهای آپوپتوزی مانند مسیرهای وابسته به p53 می‌شود [۳].

سرطان سینه به دلیل ناهمگونی زیستی گسترده‌ای که دارد، از نظر مولکولی به چندین زیرگروه

^۱ برای بررسی و تعریف دقیق نشانگرهای بیولوژیکی (Biomarker) به آآ-۱ مراجعه کنید.

مشخص تقسیم‌بندی می‌شود. در حالی که در گذشته تنها چهار زیرگروه اصلی تعریف شده بود، مطالعات اخیر بر اساس الگوهای بیان ژنی، شش زیرگروه مولکولی متمایز را معرفی کرده‌اند [۴] :

.Luminal A, Luminal B, Basal-like, Claudin-low, Normal-like, HER2-positive

تومورهای Luminal A که با بیان بالا در گیرنده‌های هورمونی (PR و ER) و سطح پایین Ki-67 مشخص می‌شوند، کمترین میزان تهاجم و بهترین پیش‌آگهی را دارند. در مقابل، تومورهای Luminal B دارای نرخ تکثیر بالاتر، میزان Ki-67 افزایش یافته و اغلب بیان HER2 هستند که آن‌ها را نسبت به درمان‌های هورمونی مقاوم‌تر می‌کند و پیش‌آگهی آن‌ها را بدتر می‌کند. در HER2-positive، افزایش بیان ژن HER2 رخ می‌دهد. این تومورها معمولاً نسبت به درمان‌های ضد HER2 مقاومت کمی نشان می‌دهند. در تومورهای Basal-like که با سرطان سینه سه‌گانه منفی (TNBC) شناخته می‌شود اما بیان گیرنده‌های هورمونی و HER2 وجود ندارد که منجر به تومورهایی تهاجمی و مقاوم به شیمی‌درمانی می‌شود و درمان آن را یک فرآیند چالشی می‌کند. زیرگروه Claudin-low با بیان پایین ژن‌های چسبندگی سلولی و ویژگی‌های سلول‌های بنیادی تومور، درمان‌های استاندارد را به خوبی پاسخ نمی‌دهد و اغلب به مقاومت دارویی دچار می‌شود. در نهایت، Normal-like که شبیه بافت طبیعی سینه است، به طور کلی با پیش‌آگهی بهتر و پاسخ بهتر به درمان‌های استاندارد همراه است.

این تنوع مولکولی نشان‌دهنده‌ی چالش‌های موجود در شناسایی بیومارکرهای دقیق برای تشخیص و درمان سرطان پستان است، چرا که بسیاری از بیومارکرهای موجود در زیرگروه‌های مقاوم‌تر مانند Basal-like و Claudin-low قادر به پیش‌بینی پاسخ به درمان نمی‌باشند. به طور دقیق‌تر روش‌های فعلی برای تشخیص و درمان سرطان پستان عمدتاً به بیومارکرهای پروتئینی مانند PR، ER و HER2 مقتکی هستند که به عنوان استانداردهای بالینی برای تعیین نوع درمان بیماران مورد استفاده قرار می‌گیرند. با این حال این بیومارکرها دارای محدودیت‌هایی هستند. اول، همانطور که در ابتدا گفته شد این بیومارکرها قادر به پوشش کامل ناهمگونی زیستی سرطان پستان نیستند. دوم، وجود ناهمگونی داخل توموری باعث کاهش دقت این بیومارکرها در پیش‌بینی پاسخ بیمار به درمان می‌شود. سوم، پیشرفت مقاومت دارویی در بیماران Luminal A و B نشان می‌دهد که بیومارکرهای فعلی به تنها‌ی قابلی قادر به پیش‌بینی دقیق رفتار سرطان نیستند.

برای رفع چالش‌های موجود در شناسایی بیومارکرهای دقیق‌تر، نیاز به بیومارکرهای جدید با دقت بالاتر و پوشش بهتر ناهمگونی‌های زیستی سرطان پستان احساس می‌شود. در این مطالعه، با استفاده از رویکردی جامع و چند مرحله‌ای، تلاش کرده‌ایم که بیومارکرهای جدیدی شناسایی کنیم که بتوانند دقیق‌تر و بهتر سرطان پستان را تشخیص دهنند. این فرآیند شامل تحلیل داده‌های بیان ژنی به منظور شناسایی ژن‌های دارای تغییرات افتراقی معنادار، استفاده از روش‌های یادگیری ماشین برای انتخاب

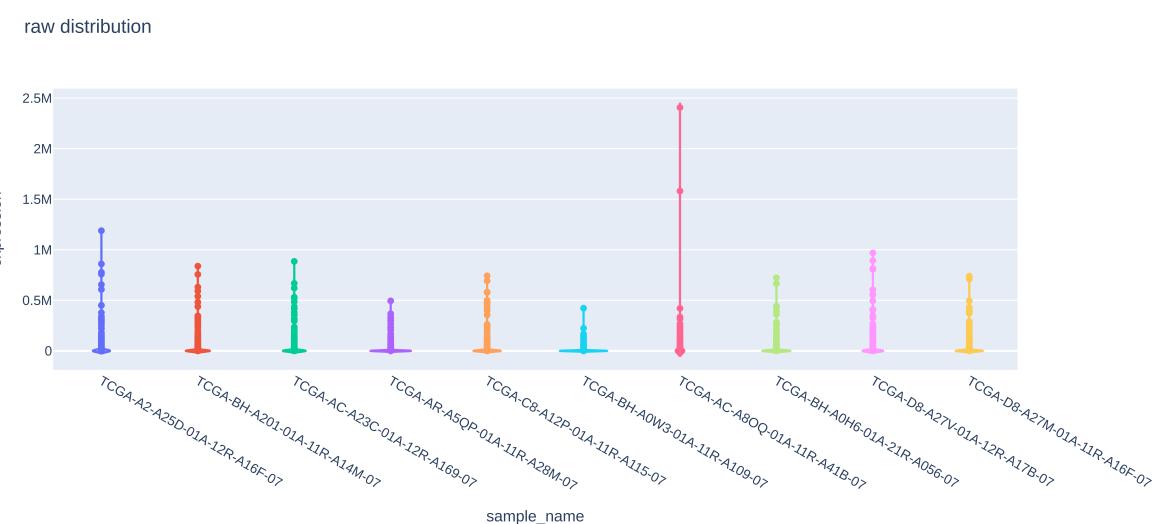
ویژگی‌های کلیدی، بررسی عملکرد زیستی ژن‌های منتخب در مسیرهای سیگنالینگ و فرآیندهای سلولی مختلف، و در نهایت، تحلیل پیش‌آگهی بهمنظور بررسی ارتباط این ژن‌ها با ویژگی‌های بالینی بیماران و میزان بقای آن‌ها است. با انجام این مراحل، ما قادر به شناسایی بیومارکرهایی خواهیم بود که می‌توانند در تشخیص سرطان پستان و پیش‌بینی پیش‌آگهی دقت بیشتری داشته باشند.

۲ داده‌ها و پایگاه‌ها

در این بخش به توضیح داده‌های مورد استفاده و همچنین پایگاه‌های استفاده شده برای ژن‌های منتخب می‌پردازیم.

۱-۲ داده‌های بخش تحلیل بیان ژن افتراقی

برای بررسی تفاوت بیان ژن‌ها از داده‌های RNA-seq استفاده می‌کنیم که به وسیله‌ی فناوری توالی یابی نسل جدید (NGS) استخراج شده‌اند. همچنین از پایگاه داده The Cancer Genome Atlas برای تهیه‌ی داده‌های RNA-seq استفاده می‌کنیم. این پایگاه داده یکی از جامع‌ترین منابع اطلاعاتی در زمینه سرطان است که شامل داده‌های RNA-seq می‌باشد. با استفاده از این پایگاه داد نمونه‌های سرطانی و سالم مربوط به سرطان سینه (BRCA) برای تحلیل افتراقی بدست می‌آیند. در این بررسی، ۱۰ نمونه از افراد سالم به عنوان نمونه کنترل و ۳۹ نمونه از افراد دارای سرطان سینه مورد استفاده قرار گرفته است. توزیع اولیه این نمونه‌ها به شکل زیر می‌باشد.



شکل ۱: توزیع داده‌ها بدون پردازش

۲-۲ پایگاه GEPIA

پایگاه داده‌ی GEPIA برای بررسی بیان ژنی و تحلیل افتراقی بین نمونه‌های سرطانی و سالم است که داده‌های آن از پایگاه‌هایی چون TCGA و GTEx استخراج شده‌اند. این ابزار امکان تحلیل بیان ژنی را در سطح RNA-Seq فراهم می‌کند و از روش‌های آماری کلاسیک برای مقایسه بیان ژن‌ها بین سرطان و نمونه‌های طبیعی استفاده می‌کند. تحلیل‌های ممکن در GEPIA شامل مقایسه‌ی بیان ژنی بین سرطان و بافت سالم، بررسی همبستگی بین ژن‌ها، و تحلیل بقا با استفاده از Kaplan-Meier است. از این پایگاه در ادامه برای بررسی تفاوت بیان ژن‌های منتخب در بافت سالم و سرطانی به ازای تعداد نمونه‌های بیشتر استفاده می‌شود [۵].

۳-۲ پایگاه UALCAN

علاوه بر پایگاه داده GEPIA، پایگاه داده UALCAN نیز یکی از ابزارهای مهم برای تحلیل بیان ژنی در انواع سرطان‌ها، از جمله سرطان سینه، محسوب می‌شود. این پایگاه داده که از اطلاعات RNA-seq و داده‌های بالینی TCGA استفاده می‌کند، امکان مقایسه‌ی بیان ژن‌ها بین نمونه‌های سرطانی و نرمال، تحلیل تأثیر بیان ژن بر بقای بیماران، بررسی ارتباط بیان ژن با زیرگروه‌های مختلف سرطان و تحلیل متیلاسیون DNA را فراهم می‌کند.

۴-۲ پایگاه GeneCards

این پایگاه داده اطلاعاتی درباره‌ی عملکرد زیستی ژن‌ها، نقش آن‌ها در بیماری‌ها، مسیرهای زیستی مرتبط، و داروهای هدفمند را جمع‌آوری می‌کند. این پایگاه داده داده‌های خود را از منابع متعددی مانند UniProt, NCBI, Reactome, KEGG استخراج کرده و اطلاعاتی درباره‌ی موقعیت کروموزومی ژن، پلی‌مورفیسم‌های شناخته‌شده، تغییرات اپی‌ژنتیکی، و شبکه‌های تعاملی ژن ارائه می‌دهد. در بررسی ما از این پایگاه داده برای تشخیص مکان پروتئین ژن‌های منتخب در سلول استفاده می‌شود [۶].

۵-۲ پایگاه bc-GenExMiner

کی از پایگاههای داده‌ی اختصاصی برای سرطان سینه است که داده‌های بیان ژنی را برای انواع مختلف سرطان سینه ارائه می‌دهد. این پایگاه داده امکان بررسی ارتباط بیان ژن با زیرگروههای سرطان سینه مانند HER2-positive را فراهم می‌کند. مهم‌ترین خروجی‌های این پایگاه داده شامل Box Plot برای مقایسه‌ی بیان ژنی در زیرگروههای مختلف سرطان سینه، نمودارهای همبستگی، و مقادیر p-value برای آزمون‌های آماری مربوط به تغییرات بیان ژن است. از این اطلاعات می‌توان برای ارزیابی ارتباط تفاوت بیان با انواع سرطان سینه استفاده کرد و چنانچه از نظر آماری تفاوتی در هیچ نوعی دیده نشد می‌توان آن ژن را در نظر نگرفت [۴].

۶-۲ پایگاه cBioPortal

یکی دیگر از تحلیل‌های حیاتی در رسیدن به بهترین مجموعه از ژن‌ها به عنوان بیومارکر، استفاده از تغییرات ژنتیکی و جهش‌ها است آ-۲. با استفاده از این پایگاه داده می‌توان تغییرات و جهش‌ها را در پایگاههای مختلف به ازای ژن‌های دلخواه بررسی کرد. در بررسی ما از ۱۰۹۶۷ نمونه در ۳۲ بررسی استفاده شده است که نمونه‌های آنها از TCGA بدست آمده است [۵].

۷-۲ پایگاه Kaplan-Meier Plotter

از این پایگاه برای تحلیل بقا استفاده می‌کنیم. به این معنا که آیا میزان بیان یک ژن خاص بر طول عمر بیماران تأثیر دارد یا خیر. خروجی اصلی این پایگاه داده یک Kaplan-Meier Survival Curve است که بقا را بین دو گروه از بیماران مقایسه می‌کند: بیمارانی که بیان ژن بالایی دارند و بیمارانی که بیان ژن کمتری دارند. مقدار Hazard Ratio نیز در این خروجی ارائه می‌شود؛ اگر مقدار HR بزرگ‌تر از ۱ باشد، یعنی بیان زیاد ژن منجر به کاهش بقا شده است و اگر کمتر از ۱ باشد، نشان‌دهنده‌ی تأثیر محافظتی آن ژن است. علاوه بر این، مقدار p-value از آزمون Log-rank مشخص می‌کند که آیا تفاوت بین دو گروه معنادار است یا خیر [۶].

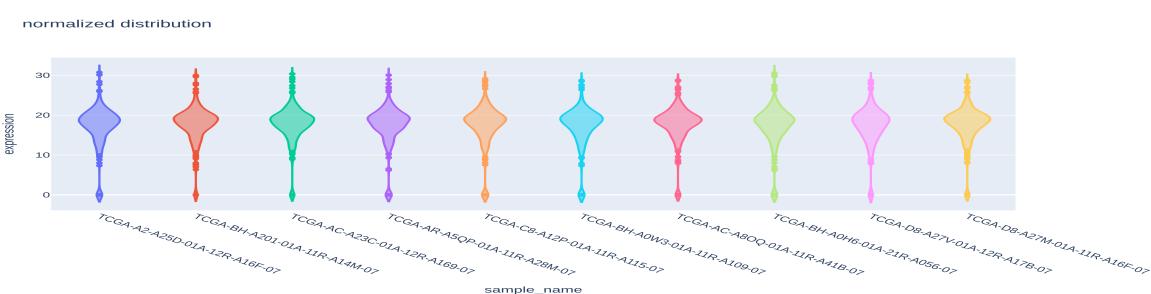
۸-۲ پایگاه Reactome

برای بررسی مسیرهای زیستی از این پایگاه استفاده می‌کنیم. مهم‌ترین خروجی این پایگاه داده نقشه‌های مسیرهای زیستی و نمودارهای ارتباطی ژن‌ها است. اگر یک ژن در مسیرهای کلیدی سرطان سینه دخیل نباشد، ممکن است اهمیت کمتری برای بیومارکر بودن داشته باشد [۱۰].

۳ پیش‌پردازش داده‌های RNA-seq

یکی از مهم‌ترین مراحل در پرداش داده‌های بیان ژن پیش‌پردازش داده‌ها می‌باشد. در استفاده از داده‌های RNA-seq ممکن است به دلیل تعداد خوانش متفاوت هر نمونه (sequencing depth) و یا طول متفاوت ژن‌ها سوگیری‌هایی اتفاق بیفتد که در تحلیل افتراقی تاثیر منفی بگذارند. به همین دلیل برای قابل مقایسه کردن ژن‌ها از روش UQ-FPKM استفاده شده است. برای بررسی دقیق‌تر این روش به بخش آ-۳ مراجعه کنید. بنابراین برای اینکه بتوان به طور موثری بیان ژن‌ها را بررسی کرد از این تکنیک استفاده می‌کنیم. داده‌های بدست آمده از پایگاه TCGA با استفاده از این روش نرمالایز می‌شوند. در ادامه نیز از \log_2 transform استفاده شده است. این تبدیل مقادیر بزرگ را فشرده و مقادیر کوچک را گسترش می‌دهد، که توزیع داده‌ها را متقارن‌تر کرده و آن‌ها را برای تحلیل آماری مناسب‌تر می‌سازد. همچنین، مقیاس بدست آمده از این تبدیل در ژنومیک مناسب است، زیرا تغییرات بیان ژن معمولاً به صورت مضارب ۲ بیان می‌شوند.

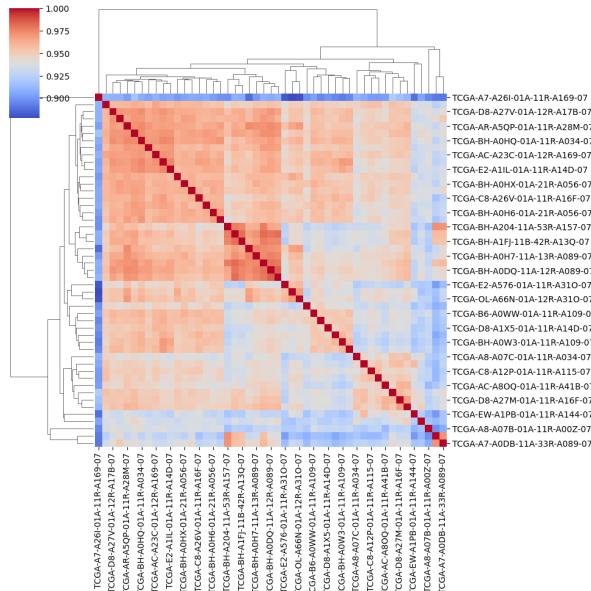
همانطور که در ۱ دیدیم، توزیع داده‌ها تفاوت بسیاری دارد که می‌تواند موجب سوگیری شود. با استفاده از تکنیک‌های گفته شده می‌توان توزیع نمونه‌ها را تطابق داد تا امکان مقایسه عادلانه بین نمونه‌ها و ژن‌ها فراهم شود.



شکل ۲: توزیع داده‌ها بعد از پردازش

در ادامه هیئت‌مپ کورلیشن با خوشبندی سلسله مراتبی ارائه شده است که شباهت یا تفاوت نمونه‌های مختلف را بر اساس بیان ژن‌ها نشان می‌دهد. در این نمایش می‌توان گروه‌بندی‌های مجزای

نمونه‌های سرطانی و سالم را شناسایی کرد که بیانگر تفاوت‌های قابل توجه در بیان ژن‌ها بین این دو گروه است. این موضوع به ما این بینش را می‌دهد که اطلاعات ارزشمندی از طریق تحلیل تفاوت بیان ژن‌ها قابل استخراج است. بنابراین در مسیر یافتن نشانگرهای سرطان سینه، یک نقطه شروع مناسب می‌تواند تحلیل تفاوت بیان ژن‌ها در این نمونه‌ها باشد.

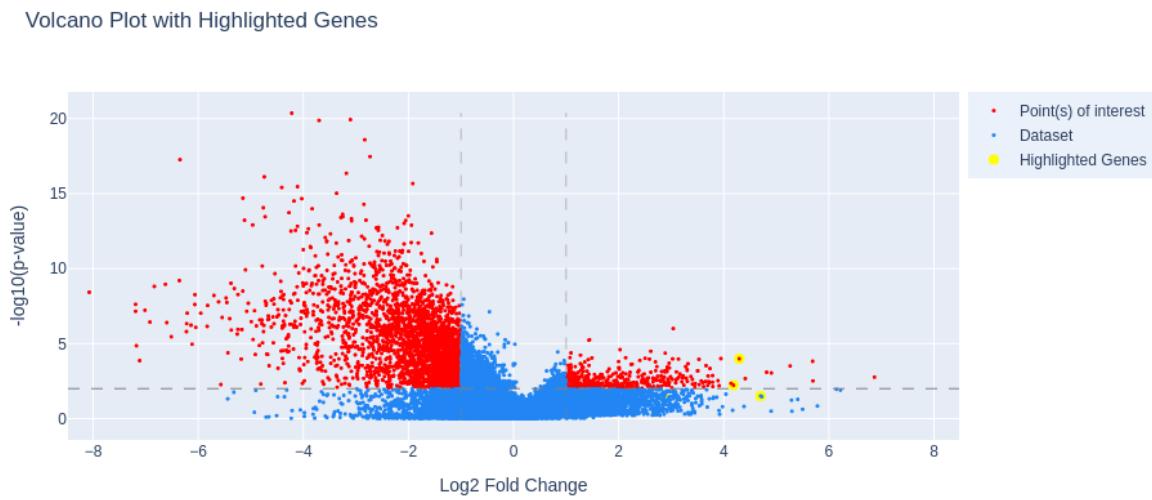


شکل ۳: نمودار شباهت نمونه‌ها

۴ تحلیل و نتایج

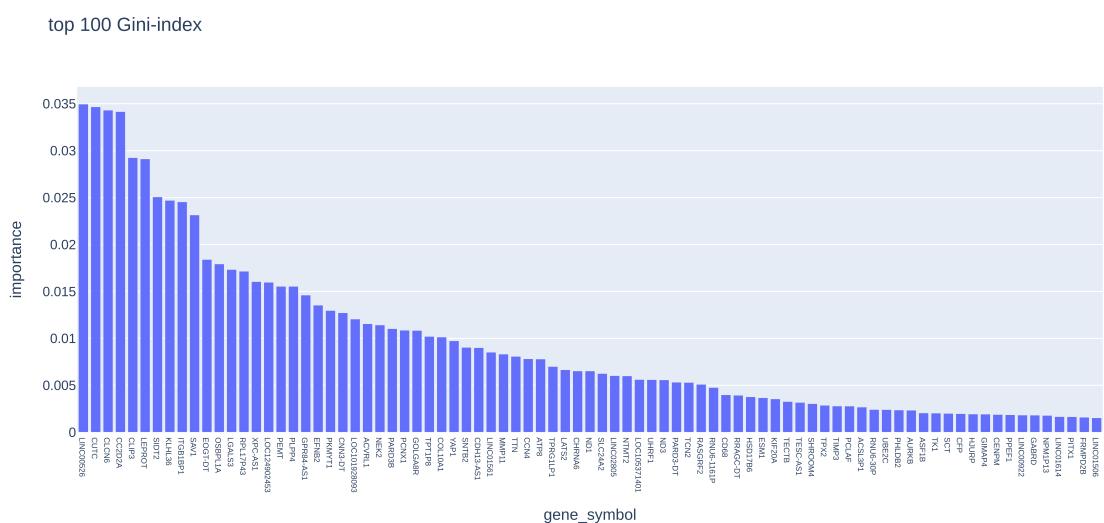
۱-۴ پیدا کردن ژن‌ها با تفاوت بیان معنادار

در مرحله اول 500 ژن که دارای تفاوت بیان معنادار هستند، پیدا می‌شوند. با استفاده از $\log FC$ تفاوت بیان یک ژن در دو دسته‌ی کنترل و سرطانی اندازه‌گیری می‌شود. از آستانه 1 ± 1 برای جداسازی ژن‌ها با تغییر بیان معنادار استفاده می‌کنیم. همچنین علاوه بر $\log FC$ از آزمون t برای مقایسه میانگین بیان ژن‌ها در گروه‌های مختلف انجام می‌شود تا با محاسبه p -value برای اصلاح p -value Benjamini-Hochberg بروز گشته شود. بهمنظور کاهش نرخ کشف کاذب (FDR) از روش Benjamini-Hochberg برای اصلاح p -value‌ها استفاده می‌شود. توضیحات بیشتر در آنچه موجود است. برای این معیار از آستانه 1% استفاده می‌کنیم.



شکل ۴: نمودار آتش‌فشن

در گام بعدی برای بررسی بیشتر اثرگذاری ژن‌ها در سالم یا سرطانی بودن نمونه‌ها از یک الگوریتم یادگیری ماشین به نام جنگل تصادفی آ-۵ استفاده می‌کنیم. در واقع در این روش از ژن‌ها به عنوان تعدادی فیچر استفاده می‌کنیم تا براساس آنها بتوانیم نمونه‌های سالم و سرطانی را کلاس بندی بکنیم. در این الگوریتم فیچرها براساس شاخصی به نام Gini index اولویت بندی شده‌اند. این اولویت نشان می‌دهد که هر ژن به چه میزان در تشخیص کلاس نمونه می‌تواند اثرگذار باشد. براساس این توضیحات، ۱۰۰ ژن با بیشترین مقدار Gini index انتخاب می‌شوند تا بر روی آنها بررسی بیشتر انجام شود.



شکل ۵: صد ژن با بیشترین میزان Gini index

۲-۴ بررسی دقیق ژن‌های منتخب

در بخش قبل با استفاده از روش‌های آماری و یادگیری ماشین توانستیم مجموعه‌ای از ژن‌ها را استخراج کنیم که در بین نمونه‌های سالم و سرطانی بیشترین تفاوت بیان را دارند. با مشاهداتی که در بخش‌های گذشته مثل نمودار ۳ داشتیم، به این نتیجه رسیدیم که ژن‌ها و مسیرهای زیستی وجود دارند که می‌توانند تفاوت‌های یک سلول سرطانی و سالم را ایجاد کنند. با این حال ارائه ژن‌های یافت شده به عنوان بیومارکرهای سرطان سینه حتی با تست‌های سختگیرانه کاری درست نخواهد بود. به بیان دیگر، تحلیل DEG تنها یک نقطه‌ی شروع برای شناسایی بیومارکرهای است و برای اعتبارسنجی آن‌ها باید مراحل تکمیلی انجام شود. این مراحل شامل بررسی اهمیت زیستی ژن، تأثیر بر بقا، تغییرات ژنتیکی، ویژگی‌های بالینی و ارتباط با زیرگروه‌های بیماری است. ژن‌هایی که صرفاً در DEG معنادار هستند اما در سایر تحلیل‌ها عملکرد ضعیفی دارند، معمولاً بیومارکرهای مناسبی نیستند. برخی از ویژگی‌های مورد نیاز یک بیومارک در آ-۱ موجود است. بنابراین لازم دانستیم در این پژوهه بررسی فراتر از تفاوت بیان ژن داشته باشیم.

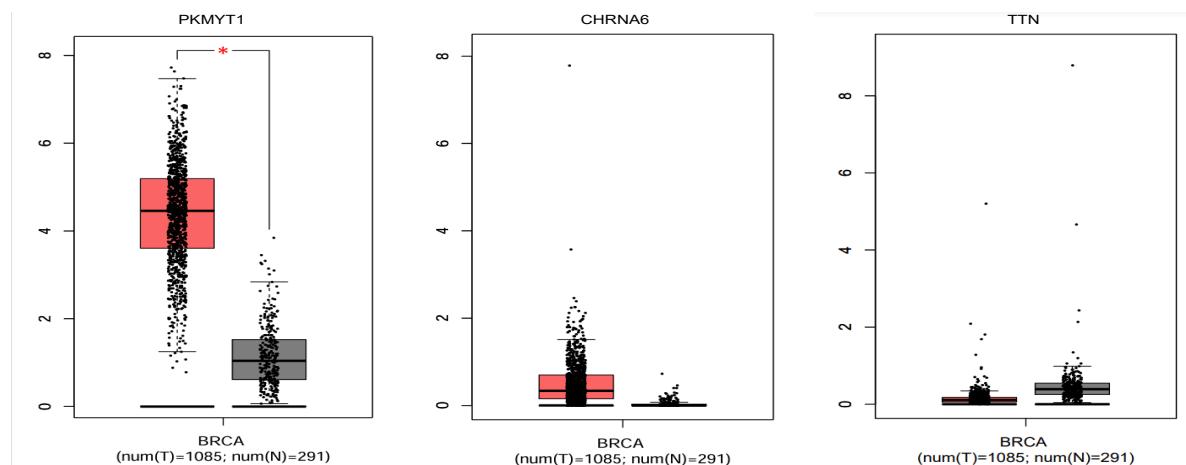
توجه شود در ادامه تمام نتایج برای سه ژن نمایش داده می‌شود و تحلیل‌ها بر روی آنها انجام می‌شود. توجه شود همانطور که در بخش‌های آینده خواهیم دید، بررسی نشان می‌دهند که PKMYT1 و CHRNA6 قابلیت تبدیل شدن به بیومارک را دارند. انتخاب ژن دیگر برای نشان دادن مراحل تحلیل است. در هر مرحله ممکن است براساس ضعیف بودن ژن آن را حذف کنیم.

۱-۲-۴ بررسی نتایج GEPIA و UALCAN

این پایگاه همانطور که پیش‌تر گفته شد امکانات گوناگونی را در رابطه با بیان ژن‌ها در نمونه‌های سرطانی در اختیار ما قرار می‌دهد. با استفاده این پایگاه می‌توانیم تفاوت بیان ژن‌های مختلف را در حجم بزرگتری از نمونه‌ها بسنجدیم تا میزان پایداری آنها تعیین شود.

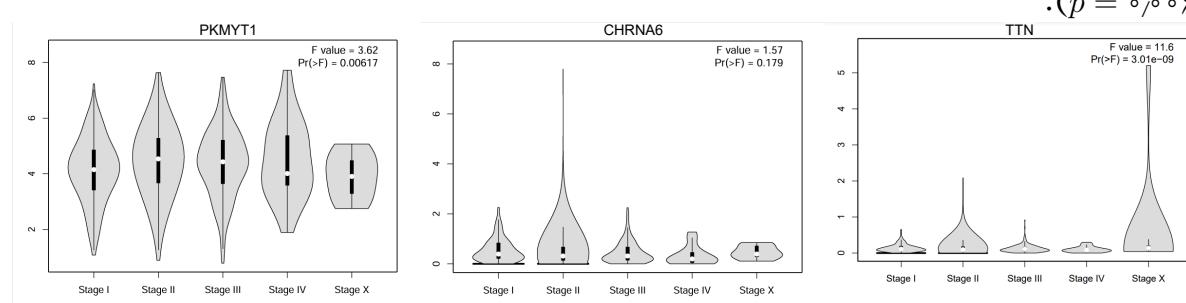
در اولین بررسی، اختلاف بیان ژن‌های منتخب را در بافت‌های سالم و سرطانی بررسی می‌کنیم. با توجه به نمودار PKMYT1 دارای اختلاف بیان بالایی است در حالی که این موضوع در دو ژن دیگر دیده نمی‌شود. با این حال با استفاده از پایگاه UALCAN مشخص می‌شود که سطح بیان در بافت سرطانی به طور معناداری بالاتر از بافت سالم برای ژن CHRNA6 است.

Gene	p-value
PKMYT1	1E-12
CHRNA6	1.315800E-03
TTN	2.139800E-01



شکل ۶: سطح بیان ژن‌ها در بافت‌های سالم و سرطانی

از دیگر امکانات این پایگاه‌ها می‌توان به بررسی سطح بیان ژن در مراحل مختلف بیماری اشاره کرد. بر این اساس مجدداً دیده می‌شود که PKMYT1 به طور معناداری دارای تفاوت بیان است . ($p = 0.006$)



شکل ۷: میزان اختلاف بیان ژن‌ها در مراحل مختلف سرطان سینه

براساس این بررسی‌ها تا اینجا بنظر می‌رسد که PKMYT1 می‌تواند در نقش یک بیومارکر تشخیصی ظاهر شود چون تفاوت بیان آن در بافت‌های سالم و سرطانی به صورت معناداری مشهود است. با این حال این مرحله بسیار شبیه به تحلیل داده‌های RNA-seq بود و نمی‌توان بیومارکر بودن ژنی را نتیجه گرفت. در این مرحله ژن‌هایی که تفاوت بیان معناداری در این دو پایگاه نداشتند حذف شدند.

۲-۲-۴ برسی نتایج GeneCard

این پایگاه اطلاعات گوناگونی را درباره ژن مورد نظر در اختیار ما قرار می‌دهد. در این بخش اما ما از این پایگاه به منظور شناسایی محل زیرسلولی پروتئین‌ها استفاده می‌کنیم.

بررسی محل زیرسلولی پروتئین‌ها در شناسایی بیومارکرها اهمیت زیادی دارد، زیرا موقعیت مکانی یک پروتئین در سلول می‌تواند نقش آن را در بیماری و قابلیت آن به عنوان یک هدف درمانی مشخص کند. پروتئین‌هایی که در غشای سلولی قرار دارند، معمولاً برای طراحی داروهای هدفمند بسیار مهم هستند، زیرا به راحتی در دسترس مولکول‌های دارویی مانند آنتی‌بادی‌ها یا مهارکننده‌های سطحی قرار می‌گیرند. در مقابل، پروتئین‌هایی که در هسته یا سیتوپلاسم قرار دارند، ممکن است در مسیرهای سیگنالینگ درون‌سلولی یا تنظیم بیان ژنی نقش داشته باشند و به عنوان بیومارکرهای پیش‌آگهی یا پیش‌بینی‌کننده‌ی پاسخ به درمان مورد استفاده قرار گیرند. با استفاده از پایگاه GeneCards، می‌توان مکان احتمالی پروتئین‌های شناسایی‌شده را تعیین کرد.

محل قرارگیری پروتئین ژن‌های TTN، PKMYT1 و CHRNA6 به ترتیب با احتمال بالا در هسته و سیتوزل، غشای پلاسمایی، هسته و اسکلت سلولی می‌باشد.

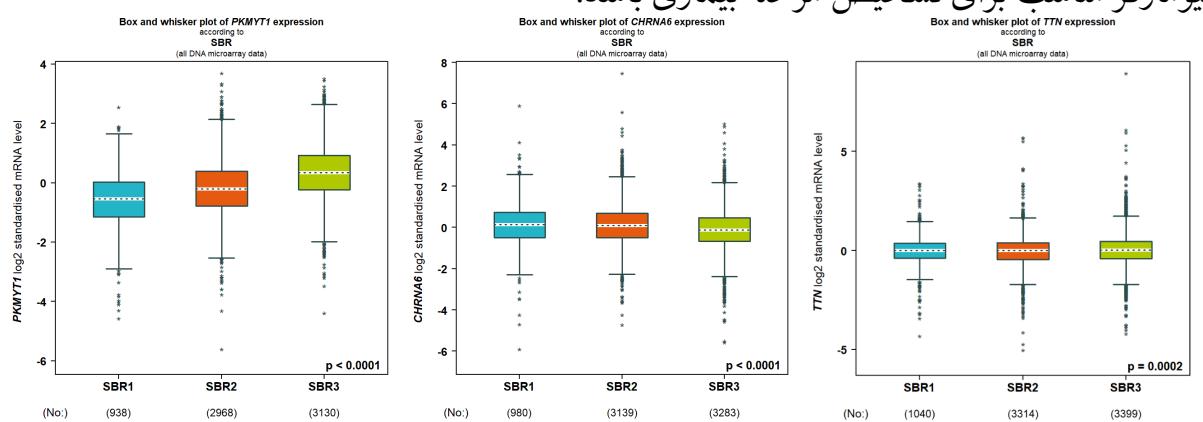
اینطور به نظر می‌رسد که PKMYT1 نقش مهمی در چرخه‌ی سلولی دارد. با این حال به دلیل دسترسی سخت آن امکان کنترل آن میسر نیست و بیشتر می‌تواند نقش تشخیصی داشته باشد. ژن CHRNA6 در غشای پلاسمایی بیان می‌شود و به عنوان بخشی از گیرنده‌های نیکوتینی استیل‌کولین نقش دارد. این موقعیت باعث می‌شود که این پروتئین بهتر در دسترس آنتی‌بادی‌ها و روش‌های تشخیصی مبتنی بر سطح سلول قرار گیرد. ژن TTN با اینکه در هسته است اما در مسیرهای مربوط به چرخه‌ی سلولی شرکت نمی‌کند^{۶-۲-۴}. با توجه به حضور این پروتئین در اسکلت سلولی می‌توان نتیجه گرفت که نقش این ژن بیشتر مرتبط با ساختار و انعطاف سلول است. یافته‌های موجود در بخش ۶-۲-۴ نیز نشان می‌دهند که این ژن در مسیرهای مربوط به عضلات و انقباض عضلانی حضور دارد.

۳-۲-۴ برسی نتایج bc-GenExMiner

از این پایگاه به منظور بررسی ارتباط بیان ژن‌ها با پارامترهای کلینیکی-پاتولوژیکی (Clinico-Pathological Parameters) استفاده شده است. پارامترهای کلینیکی-پاتولوژیکی مجموعه‌ای از ویژگی‌های بیمار و بافت توموری هستند که برای بررسی و پیش‌بینی شدت بیماری، پاسخ به درمان و پیش‌آگهی سرطان استفاده می‌شوند. یکی از دلایل اهمیت این بررسی در پیدا کردن بیومارکرهای

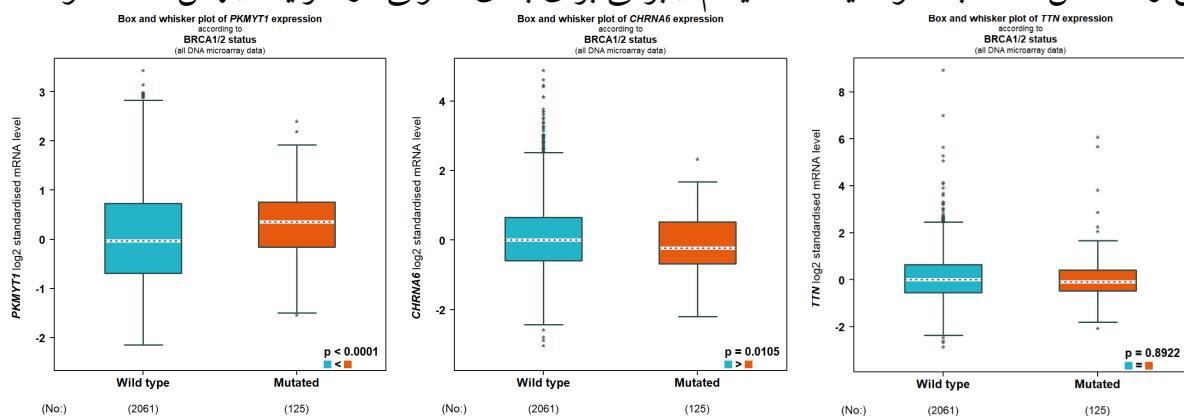
سرطان سینه، بررسی بیان ژن‌ها در تقسیم‌بندی‌های مختلف بیماری است. در این بررسی دو دسته مختلف تحلیل شده است.

در تقسیم‌بندی اول از SBR استفاده شده است. برای بررسی دقیق آن به آن راجعه کنید. با توجه به نمودارهای پایین می‌توان دید که میزان بیان PKMYT1 با بدخیم شدن سرطان افزایش یافته است. از طرفی این روند در CHRNA6 تقریباً بر عکس بوده است. در ژن TTN اما میزان بیان تقریباً ثابت باقی مانده است. تفسیر این مشاهده می‌تواند این باشد که میزان بیان PKMYT1 می‌تواند یک بیومارکر مناسب برای تشخیص مرحله بیماری باشد.



شکل ۸: سطح بیان ژن‌ها با توجه به تقسیم‌بندی SBR

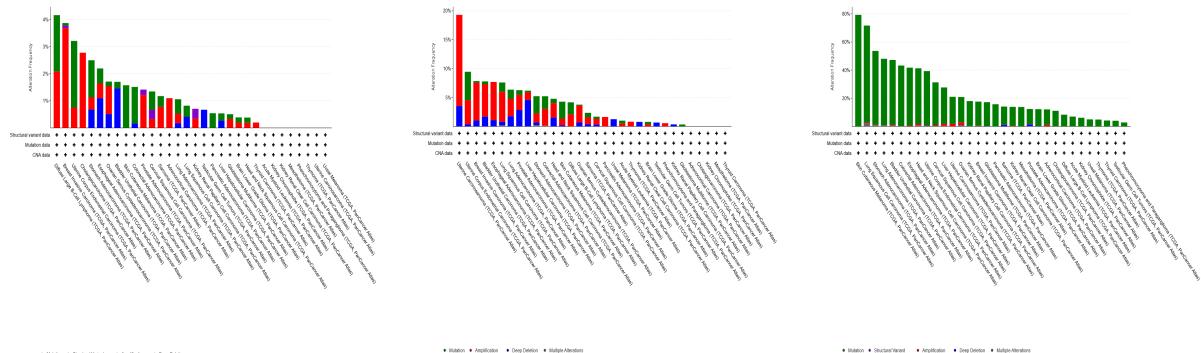
در تقسیم‌بندی دیگر از BRCA1/2 استفاده می‌کنیم. دو ژن BRCA1 و BRCA2 دو ژن مهم در ترمیم آسیب DNA هستند. اگر یکی از این ژن‌ها دچار جهش شود، احتمال ابتلا به سرطان سینه افزایش می‌یابد. براین اساس گروه wild-type افرادی هستند که این دو ژن در آنها جهش نداشته است و گروه mutated افراد دارای جهش هستند. مجدداً همان روند تقسیم‌بندی قبلی قابل مشاهده است. این یافته نشان می‌دهد که PKMYT1 ممکن است در مکانیسم‌های مولکولی مرتبط با سرطان‌های ارشی نقش داشته باشد. افزایش بیان PKMYT1 در بیماران جهش‌یافته BRCA1/2 نشان می‌دهد که این ژن ممکن است به عنوان یک مکانیسم جبرانی برای بقای سلولی در شرایط جهش فعال شود.



براساس بررسی‌ها، ارتباط بالایی بین بیان ژن PKMYT1 و فرم‌های تهاجمی سرطان سینه و همچنین حالت ژنتیکی آن وجود دارد. این یعنی بیان بیشتر این ژن می‌تواند باعث شدیدتر شدن بیماری شود. از سوی دیگر دیده شد که بیان CHRNA6 با بدتر شدن وضعیت بیماری کاهش می‌یابد. این می‌تواند نشان دهندهٔ نقش محافظتی این ژن در سرطان سینه باشد و احتمال استفاده از آن به عنوان یک بیومارکر پیش‌آگه‌ی را افزایش می‌دهد. با این حال در مورد ژن TTN به نظر می‌رسد که ارتباطی بین بیان آن و پیشرفت بیماری وجود ندارد.

۴-۲-۴ بررسی نتایج cBioPortal

در cBioPortal، تغییرات ژنتیکی شامل جهش‌ها، تغییرات ساختاری، افزایش تعداد نسخه و حذف عمیق بررسی می‌شوند. در پیوست آ-۲ به طور دقیق‌تر درباره اهمیت این بررسی توضیح داده‌ایم. نکته‌ی مهم اول این است که در دو ژن PKMYT1 و CHRNA6، سرطان سینه از نرخ تغییر بالایی برخوردار است. نکته‌ی مهم دیگر این است که افزایش تعداد نسخه بخش عمدی این تغییرات را برای آنها تشکیل می‌دهد. نکته‌ی جالب توجه دیگر بالا بودن نرخ حذف عمیق در CHRNA6 است. با توجه به نکات گفته شده در پیوست آ-۲، این می‌تواند نشان دهنده نقش سرکوب‌کننده آن باشد. با بررسی ژن TTN می‌بینیم که در تمام سرطان‌ها دارای نرخ تغییر بالایی است و تقریباً تمام این تغییرات را جهش تشکیل می‌دهد. در نگاه اول به نظر می‌رسد که این ژن یک بیومارکر سرطان است اما با بررسی بیشتر این ژن مشخص شد که این جهش‌ها از نوع Passenger Mutations است. این تغییرات در طول رشد سلول‌های سرطانی رخ می‌دهند، اما تأثیر مستقیمی بر سرطان ندارند. علت این موضوع نیز این است که TTN به دلیل اندازهٔ بزرگ خود، بیشتر از سایر ژن‌ها در معرض جهش‌های تصادفی قرار دارد. بنابراین این جهش‌ها نمی‌توانند اطلاعات مفیدی در رابطه با سرطان سینه به ما دهند.

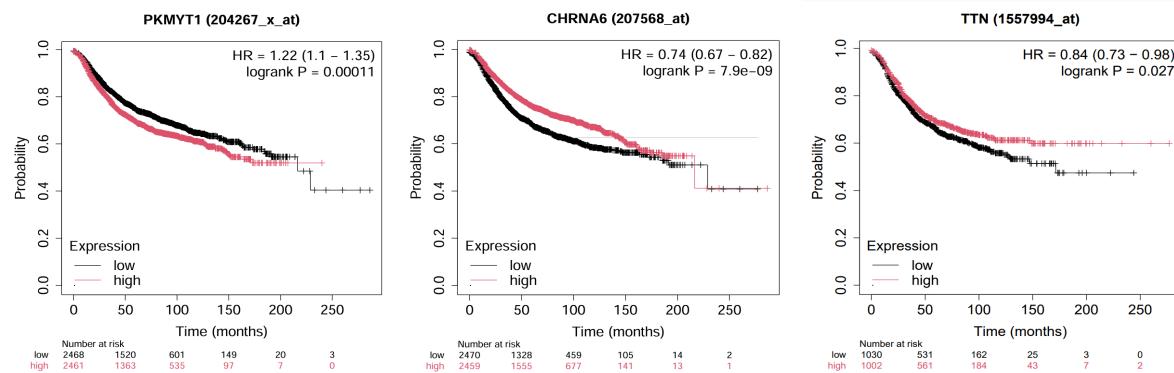


شکل ۹: نرخ تغییرات ژن‌ها در سرطان‌های مختلف

۵-۲-۴ برسی نتایج Kaplan-Meier Plotter

تحلیل بقا نقش مهمی در ارزیابی پتانسیل پیش‌آگهی یک ژن دارد. با استفاده از نمودارهای Kaplan-Meier Survival و اطلاعاتی چون Hazard Ratio و p-value می‌توان تشخیص داد که بیان چه ژن‌هایی بر بقا بیماران اثر منفی یا مثبت دارد.

نمودارها نشان می‌دهند که ژن PKMYT1 با بقا ارتباط منفی دارد. این نتیجه نشان می‌دهد که افزایش بیان PKMYT1 با کاهش بقا در بیماران سرطانی همراه است. به همین دلیل می‌تواند یک بیومارکر پیش‌آگهی منفی باشد، بهویژه برای بیماران مبتلا به سرطان که بیان این ژن در آن‌ها بالاتر است. در مقابل، بیان بالای CHRNA6 با بقا بهتر و طول عمر بیشتر بیماران ارتباط مثبت دارد. از این‌رو، CHRNA6 می‌تواند یک بیومارکر پیش‌آگهی مثبت باشد، به این معنی که افزایش بیان آن ممکن است با پیش‌بینی بهبود وضعیت بیمار و پاسخ بهتر به درمان‌های موجود همراه باشد. از طرفی TTN به‌طور معناداری با بقا و پیشرفت سرطان ارتباط ندارد. این نتیجه براساس p-value بزرگ آن نسبت به دو ژن دیگر و همچنین HR نزدیک به یک بdst می‌آید.



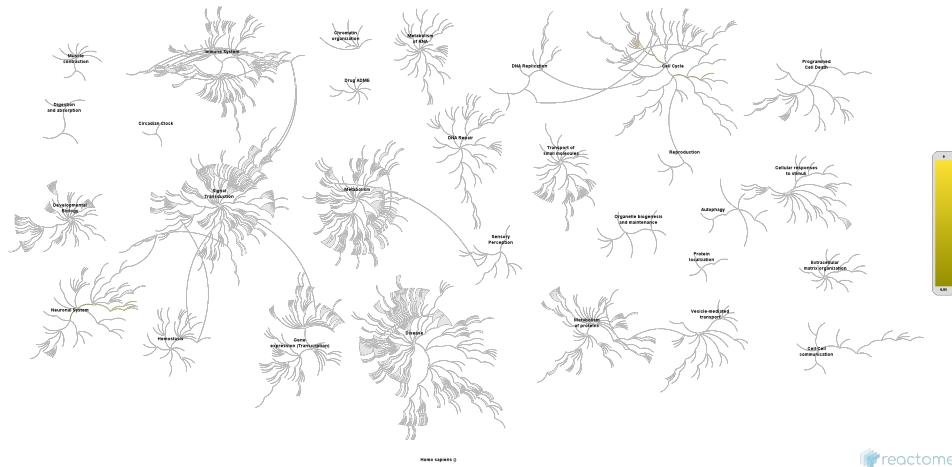
۶-۲-۴ برسی نتایج Reactome

به عنوان آخرین تحلیل لازم است که ژن‌های باقی‌مانده را از نظر مسیرهای زیستی بررسی کنیم. PKMYT1 یک کیناز است که به‌طور مستقیم CDK1 را در مرحله G2/M فسفریله و غیرفعال می‌کند. این عمل مانع از ورود زودرس سلول به مرحله میتوز می‌شود و به عنوان یک مکانیسم کنترلی برای اطمینان از تکمیل صحیح مراحل قبلی چرخه سلولی عمل می‌کند. مطالعات متعدد نشان داده‌اند که افزایش بیان PKMYT1 ر سلول‌های سرطانی می‌تواند منجر به تکثیر غیرقابل‌کنترل و در نتیجه رشد تومور شود. پژوهش [۱۱] نشان داده است که افزایش بیان با کاهش بقا در بیماران مبتلا به سرطان سینه مرتبط است، که این موضوع تأکیدی بر نقش آن به عنوان یک هدف درمانی بالقوه در کارسينومای سینه دارد. همچنین، تحقیق [۱۲] نشان داده است که افزایش بیان با پیشرفت

تومور و کاهش بقا در بیماران مبتلا به کارسینوم سلول سنگفرشی مری (ESCC) مرتبط است، که نشان‌دهنده‌ی تأثیر گستردگی این ژن در سرطان‌های مختلف است.

در مقابل، ژن *CHRNA6* که زیر واحد آلفای گیرنده‌ی نیکوتینیک استیل‌کولین را کد می‌کند، در تنظیم انتقال پیام‌های عصبی و عملکرد کanal‌های یونی نقش دارد. این پروتئین با استیل‌کولین و نیکوتین فعال می‌شود و در تنظیم انتقالات دوپامینزیک مشارک است. علاوه بر این، این ژن با مسیر سیگنالینگ *ErbB* و انتقال پیام‌های داخل‌سلولی در ارتباط است، که این مسیرها در تنظیم رشد سلولی و تمایز نقش حیاتی دارند. به دلیل نقش این مسیرها در بسیاری از سرطان‌ها، تغییرات در *CHRNA6* می‌تواند تأثیر مهمی در رفتار تومور داشته باشد. تحلیل پایگاه‌های داده‌ی کلینیکو-پاتولوژیک نشان داده است که افزایش *HER2* و کاهش *PR* با افزایش *CHRNA6* همراه است. همچنین، مشاهده شده است که جهش در *BRCA1/2* با کاهش *CHRNA6* ارتباط دارد که این ژن را یک گزینه مناسب برای بیومارکرهای سرطان سینه تبدیل می‌کند. علاوه بر این موارد، بررسی‌های *co-expression* *CHRNA6* نشان داده است که با ژن‌های *TLR7* و *OLR1* همبستگی بالایی دارد. از سوی دیگر، اگرچه افزایش *CHRNA6* با تنظیم مسیرهای پیام‌رسانی مهمی همراه است، اما برخلاف *PKMYT1*، بقای بیماران را افزایش می‌دهد و ممکن است تأثیر محافظتی در برخی انواع سرطان سینه داشته باشد.

یکی از مزایای تحلیل مسیرهای زیستی در این است که می‌توان به راحتی ژن‌های با ارتباط کم را شناسایی کرد. به عنوان مثال ژن *TTN* بزرگ‌ترین پروتئین شناخته‌شده در انسان است که نقش کلیدی در ساختار و عملکرد عضلات مخطط دارد. این پروتئین در تنظیم کشش‌پذیری و پایداری سارکومرها (واحدهای انقباضی عضله) مشارک است. بررسی مسیرهای مربوط به سرطان سینه هیچ ارتباطی بین این ژن و بیماری سرطان سینه را نشان نمی‌دهد.



شکل ۱۰: نمودار نقش ژن‌های منتخب در مسیرهای زیستی بدن

۵ نتیجه‌گیری

در این بررسی از آنالیز چند مرحله‌ای برای کشف بیومارکرهای سرطان سینه استفاده شد. در این بررسی از روش‌های آماری به منظور تحلیل تفاوت بیان استفاده شد. در ادامه با استفاده از روش‌های یادگیری ماشین استخراج ویژگی انجام شد. در ادامه با استفاده از پایگاه‌های متعدد ویژگی‌های زیستی و کلینیکی ژن‌های منتخب بررسی شدند تا در نهایت دو ژن PKMYT1 و CHRNA6 به عنوان دو ژن با پتانسیل بالا به عنوان نشانگرهای بیولوژیکی سرطان سینه ارائه شوند.

References

- [1] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: The next generation,” *Cell*, vol.144, no.5, pp.646–674, 2011.
- [2] M. C. Casimiro, M. Crosariol, E. Loro, Z. Li, and R. G. Pestell, “Cyclins and cell cycle control in cancer and disease,” *Genes & Cancer*, vol.3, no.11-12, pp.649–657, 2012.
- [3] C. V. Dang, “Myc on the path to cancer,” *Cell*, vol.149, no.1, pp.22–35, 2012.
- [4] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, and Z. Hu, “Supervised risk predictor of breast cancer based on intrinsic subtypes,” *Journal of Clinical Oncology*, vol.27, no.8, p.1160, 2009.
- [5] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, “Gepia: a web server for cancer and normal gene expression profiling and interactive analyses,” *Nucleic Acids Research*, vol.45, no.W1, pp.W98–W102, 2017.
- [6] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, R. Nudel, T. I. Stein, O. Karni, D. Oz-Levi, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet, “The genecards suite: From gene data mining to disease and drug target discovery,” *Current Protocols in Bioinformatics*, vol.54, pp.1.30.1–1.30.33, 2016.
- [7] P. Jézéquel, L. Campion, M. P. Joalland, S. Barillé-Nion, V. Verrièle, L. Biron-May, F. Dravet, J. M. Classe, S. Lassalle, and C. Guette, “bc-genexminer: An easy-to-use online platform for gene prognostic analyses in breast cancer,” *Breast Cancer Research and Treatment*, vol.131, p.765–775, 2012.
- [8] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, “The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data,” *Cancer Discovery*, vol.2, pp.401–404, 2012.
- [9] . Nagy, G. Munkácsy, and B. Győrffy, “Validation of mirna prognostic power in hepatocellular carcinoma using expression data of independent datasets,” *Scientific Reports*, vol.8, p.9227, 2018.
- [10] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, “The reactome pathway knowledgebase,” *Nucleic Acids Research*, vol.48, no.D1, pp.D498–D503, 2020.
- [11] X. Liu, Y. Zhang, J. Li, and X. Wang, “Overexpression of pkmyt1 is associated with poor prognosis in breast cancer patients,” *Journal of Cancer Research and Clinical Oncology*, vol.146, no.6, pp.1551–1561, 2020.

- [12] W. Zhang, C. Wang, and J. Liu, “Pkmyt1 upregulation promotes tumor progression and correlates with poorer overall survival in patients with esophageal squamous cell carcinoma (escc),” *Oncotarget*, vol.10, no.31, pp.3022–3031, 2019.
- [13] National Center for Advancing Translational Sciences (NCATS), “Biomarkers,” n.d. Accessed: 2025-02-03.
- [14] Atlas Antibodies, “7 types of biomarkers,” n.d. Accessed: 2025-02-03.

آ توضیحات تکمیلی

آ-1 Biomarker

بیومارکر یک ویژگی زیستی است که به طور عینی اندازه‌گیری شده و به عنوان یک شاخص از یک فرآیند زیستی طبیعی، یک فرآیند بیماری‌زا، یا پاسخ به یک مداخله درمانی (دارویی، جراحی، یا سایر درمان‌ها) مورد استفاده قرار می‌گیرد [۱۳]. از ساده‌ترین بیومارکرهای بدن می‌توان به دمای بدن اشاره کرد.

بیومارکرها بسته به هدف تحقیقاتی و بالینی خود به دسته‌های مختلفی تقسیم می‌شوند [۱۴]:

- بیومارکرهای حساسیت/ریسک (Susceptibility/Risk Biomarkers): بیومارکرهایی هستند که احتمال ابتلای یک فرد به یک بیماری خاص در آینده را پیش‌بینی می‌کنند.
- بیومارکرهای تشخیصی (Diagnostic Biomarkers): برای تشخیص یک بیماری یا تمايز آن از سایر بیماری‌ها استفاده می‌شوند. این نوع بیومارکرها می‌توانند به صورت پروتئین، RNA، یا متابولیت‌های خاصی باشند که در حضور بیماری تغییر می‌کنند.
- بیومارکرهای پیش‌آگهی (Prognostic Biomarkers): بیومارکرهایی که پیش‌بینی می‌کنند بیماری در بیماران مبتلا چگونه پیشرفت خواهد کرد. این نوع بیومارکرها برای تعیین شدت بیماری، احتمال عود، و میزان بقا بیمار استفاده می‌شوند.
- بیومارکرهای پایشی (Monitoring Biomarkers): بیومارکرهایی که به صورت مداوم برای ارزیابی وضعیت بیماری یا پاسخ بیمار به درمان اندازه‌گیری می‌شوند.
- بیومارکرهای پیش‌بینی‌کننده (Predictive Biomarkers): بیومارکرهایی که پیش‌بینی می‌کنند کدام بیماران به یک درمان خاص بهتر پاسخ خواهند داد.
- بیومارکرهای فارماکوکوئینامیکی (Pharmacodynamic Biomarkers): بیومارکرهایی که نشان می‌دهند یک پاسخ بیولوژیکی در اثر دریافت یک دارو یا مواجهه با یک عامل محیطی رخ داده است.

- بیومارکرهای ایمنی/سمیتی (Safety Biomarkers): بیومارکرها یکی که نشان‌دهنده‌ی احتمال وجود، یا میزان سمیت و عوارض جانبی ناشی از مواجهه با یک دارو یا عامل محیطی هستند.

با توجه به این نکات، هر ویژگی زیستی نمی‌تواند یک بیومارکر باشد بلکه باید دارای خواص زیر باشد:

- **ویژگی‌های زیستی:**
 - ارتباط مستقیم با بیماری. به عنوان مثال یک ژن به عنوان یک بیومارکر باید در مسیرهای زیستی بیماری نقش داشته باشد.
 - بیان افتراقی قابل توجه در نمونه‌های سالم و بیمار.
 - بیان ژن باید در مطالعات مستقل و مجموعه داده‌های مختلف پایدار باشد.
 - یک نشانه‌ی معناداری بیومارکر برای یک وضعیت داشتن تغییرات در سطح ژنتیکی است، مانند جهش، افزایش یا کاهش تعداد نسخه.
- **ویژگی‌های کلینیکی:** بیومارکر باید در تشخیص، پیش‌بینی، یا درمان بیماری مفید باشد و کاربرد بالینی داشته باشد. همچنین ژن باید به‌طور اختصاصی در بیماری مورد نظر بیان شود و در سایر بیماری‌ها یا شرایط طبیعی تغییر نداشته باشد (specificity).

• **ویژگی‌های تحلیلی:**

- قابل اندازه‌گیری باشد.
- قابلیت بازتولید نتایج را داشته باشد.
- پایداری در شرایط مختلف.

براساس موارد ذکر شده، یکی از گزینه‌های قابل در نظر گرفتن برای بیومارکر، ژن‌هایی هستن که تفاوت معنای معناداری در نمونه‌های سالم و سرطانی دارند. با این حال هر ژنی که تست‌های مربوط به تفاوت بیان را پاس کند نمی‌تواند به عنوان بیومارکر مورد استفاده قرار بگیرد. به همین دلیل در این پروژه از پایگاه داده‌های گوناگون استفاده شده است.

آ_۲ Genetic alteration and somatic mutation analysis

سرطان معمولاً زمانی ایجاد می‌شود که جهش‌های نقطه‌ای، تغییر در تعداد نسخه‌های ژنی و بازارایی‌های کروموزومی، عملکرد طبیعی سلول را مختل می‌کنند. جهش‌های نقطه‌ای، تغییر یک نوکلئوتید منفرد

در توالی DNA است که می‌تواند منجر به تولید پروتئین‌های غیرطبیعی شود. افزایش یا کاهش تعداد نسخه‌ی ژنی می‌تواند بیان ژن را تغییر دهد و در مواردی که یک ژن مرتبط با رشد سلولی تحت افزایش تعداد نسخه قرار بگیرد، می‌تواند منجر به سرطان‌زایی شود. از سوی دیگر، حذف یک ژن سرکوبگر تومور ممکن است سلول را از مکانیسم‌های کنترل رشد و ترمیم DNA محروم کند.

در میان انواع تغییرات ژنتیکی و جهش‌های سوماتیکی سه مورد زیر دارای اهمیت بالایی هستند:

- افزایش تعداد نسخه (Amplification): افزایش تعداد نسخه‌ی ژنی زمانی رخ می‌دهد که یک ژن خاص در سلول‌های سرطانی بیش از حد تکثیر شود، که اغلب باعث افزایش بیان پروتئین آن Amplification می‌شود. این نوع تغییر در سرطان سینه بسیار مهم است زیرا ژن‌هایی که دچار می‌شوند، معمولاً در تنظیم رشد سلولی و تقسیم غیرقابل کنترل نقش دارند.
- جهش‌های ژنی (Mutations): جهش‌های ژنی زمانی رخ می‌دهند که تغییرات نقطه‌ای در DNA باعث تولید پروتئین‌های تغییر یافته یا غیرفعال شود. در سرطان سینه ژن‌های سرکوبگر تومور مانند TP53 اغلب دچار جهش‌های غیرقابل بازگشت می‌شوند، که منجر به کاهش کنترل بر رشد سلولی و افزایش تقسیم‌های نامنظم سلول‌های سرطانی می‌شود. اگر یک ژن در درصد زیادی از بیماران دچار جهش باشد، ممکن است به عنوان بیومارکر پیش‌آگهی استفاده شود. اما بر خلاف Amplification، جهش‌های منفرد همیشه به افزایش بیان ژن منجر نمی‌شوند و نمی‌توانند به تنهایی معیاری قوی برای بیومارکرهای تشخیصی یا درمانی باشند.
- حذف عمیق (Deep Deletion): حذف عمیق یا Deep Deletion زمانی رخ می‌دهد که یک ژن مهم که نقش سرکوب تومور دارد، به طور کامل یا تا حد زیادی حذف شود. این تغییر باعث می‌شود که عملکرد ژن از بین بود و کنترل سلولی ضعیف شود.

با توجه به این توضیحات مهم است که ژن‌ها را از نظر جهش و یا تغییرات در بافت‌های سرطانی بررسی کنیم.

آ_۳_UQ-FPKM

در روش FPKM با استفاده از رابطه‌ی زیر بیان ژن‌ها را نرم‌الایز می‌کنیم:

$$FPKM = \frac{\text{ExonMappedFragments} \times 10^9}{\text{TotalMappedFragments} \times \text{ExonLength}}$$

این روش باعث می‌شود که بیان ژن‌ها نسبت به طول ژن و کل خوانش‌های نمونه‌ی موردنظر تعديل شود و در نتیجه، امکان مقایسه‌ی بهتر بین ژن‌های مختلف فراهم شود.

با وجود مزایای FPKM، این روش دارای محدودیت‌هایی است. اول، اگر یک یا چند ژن بسیار پر بیان باشند، مقدار FPKM سایر ژن‌ها را به‌طور مصنوعی کاهش می‌دهند، حتی اگر بیان واقعی آن‌ها تغییر نکرده باشد. دوم، این روش به شدت تحت تأثیر عمق توالی‌یابی قرار دارد، به این معنا که اگر یک نمونه تعداد کل خوانش‌های بیشتری داشته باشد، مقادیر FPKM آن به‌طور سیستماتیک کمتر از نمونه‌ای با خوانش‌های کمتر خواهد بود. این امر مقایسه‌ی بین نمونه‌ها را نادرست می‌کند و می‌تواند تحلیل‌های آماری را جهت‌دار کند.

برای رفع وابستگی بالای روش FPKM به تعداد خوانش نمونه‌ها و ژن‌ها با بیان بالا یا پایین از روش UQ-FPKM استفاده می‌کنیم. در این روش ابتدا ژن‌هایی که در تمام نمونه‌ها مقدار صفر دارند حذف می‌شوند. سپس، تعداد باقی‌مانده‌ی ژن‌ها بر چارک بالایی (۷۵ درصدی) مقادیر غیرصفر در همان نمونه تقسیم می‌شود تا فاکتور نرمال‌سازی برای هر نمونه محاسبه شود. در نهایت، این مقدار در میانگین چارک بالایی همه‌ی نمونه‌های مجموعه داده ضرب می‌شود تا مقایسه‌ی بین نمونه‌ها پایدارتر باشد.

Algorithm 1 UQ-FPKM Normalization

Require: Gene expression matrix E of size (Samples \times Genes)

Ensure: Normalized expression matrix E'

- 1: **Step 1: Remove Genes with Zero Expression in All Samples**
 - 2: $E_{\text{filtered}} \leftarrow$ Remove rows where all values are zero
 - 3: **Step 2: Compute Upper Quartile (75th percentile) for Each Sample**
 - 4: **for** each sample s in E_{filtered} **do**
 - 5: $V_s \leftarrow$ Non-zero values in sample s
 - 6: $UQ_s \leftarrow$ 75th percentile of V_s
 - 7: **end for**
 - 8: **Step 3: Compute Normalization Factor**
 - 9: $MeanUQ \leftarrow$ Mean of all UQ_s values
 - 10: **Step 4: Normalize Each Gene Expression Value**
 - 11: **for** each sample s in E_{filtered} **do**
 - 12: $NF_s \leftarrow UQ_s / MeanUQ$
 - 13: **for** each gene g in sample s **do**
 - 14: $E'_{s,g} \leftarrow E_{\text{filtered}_{s,g}} / NF_s$
 - 15: **end for**
 - 16: **end for**
 - 17: **Return** E'
-

آ_ False Discovery Rate ۴

در تحلیل بیان افتراقی، مقدار p-value نشان می‌دهد که آیا تغییر بیان یک ژن از نظر آماری معنادار است یا خیر. به طور دقیق‌تر p-value احتمال وقوع یک مقدار و حالات اکسترمی‌تر آن را نشان می‌دهد. هر چه cut off کوچک‌تر باشد یعنی شرایط رد فرض صفر سخت‌گیرانه‌تر می‌شود. با این حال وقتی این مقدار را برای تعداد زیادی ژن حساب می‌کنیم به طور متوسط $cutoff \times n$ تا خطای داریم. برای کاهش میزان خطای باید تغییری در مقدار p-value ها ایجاد کنیم. برای همین از روش Benjamini-Hochberg استفاده می‌کنیم. در این روش ابتدا p-value ها از کوچک به بزرگ مرتب می‌شوند. بزرگترین اندیس i به طوری که رابطه $\frac{i}{m} \alpha \leq p_i$ در آن برقرار باشد پیدا می‌شود. فرض صفر برای ن ژن اول رد می‌شود. بنابراین هنگامی که گفته می‌شود False Discovery Rate با آستانه 0.01 منظور قرار دادن $1 - \alpha = 0.90$ است. می‌توان نشان داد که با این کار داریم از بالا یک آستانه برای FDR قرار می‌دهیم.

آ_ Random Forest ۵

این الگوریتم از تعدادی درخت تصمیم گیری استفاده می‌کند که هر کدام بر روی مجموعه‌ای از داده‌های آموزشی با جایگذاری (Bagging) و زیرمجموعه‌ای از ویژگی‌ها آموزش دیده‌اند. نحوه‌ی تصمیم‌گیری در این مدل به صورت رای اکثریت است و از مزایای آن می‌توان به مقاوم بودن به بیش‌برازش، مقاومت به داده‌های پرت، تشخیص الگوهای غیر خطی و تعیین میزان اهمیت ویژگی‌ها اشاره کرد.

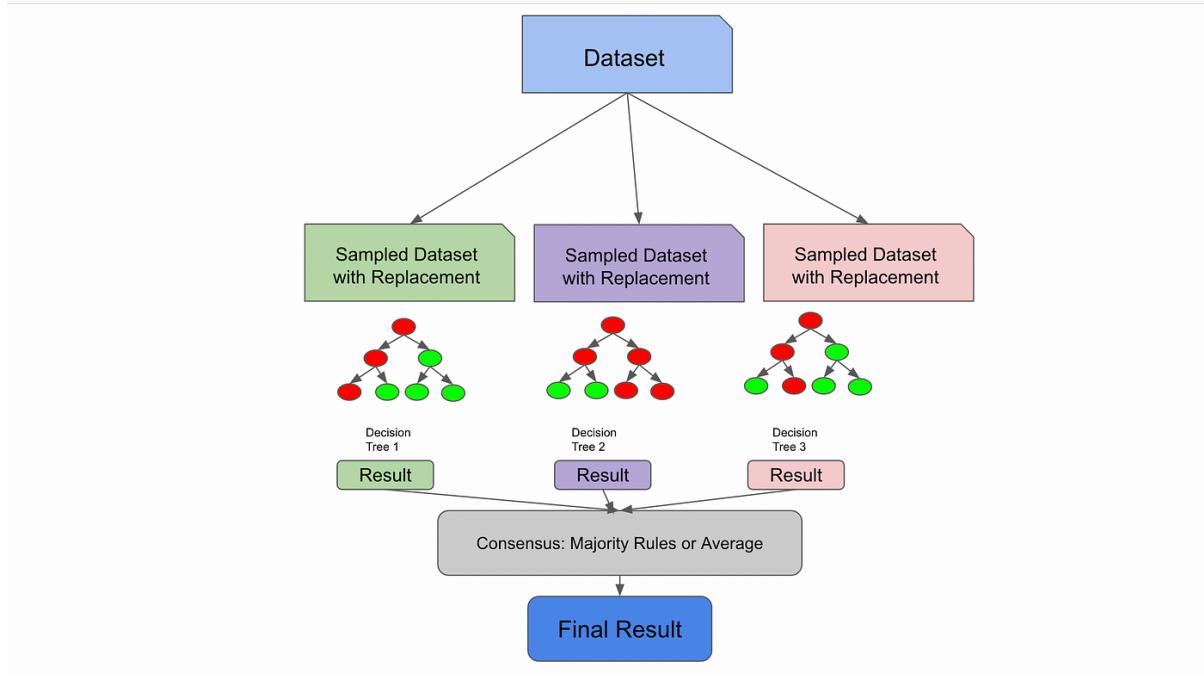
استفاده از مدل‌های یادگیری ماشین یا یادگیری عمیق به یک الگو رایج در مسائل تحلیل بیان ژن‌ها تبدیل شده است. از دلایل آن می‌توان به موارد زیر اشاره کرد:

- توانایی شناسایی روابط غیر خطی و میان ژنی. در روش‌هایی مثل p-value عموماً میزان اثربخشی هر ژن مستقلاب بررسی می‌شود.
- این الگوریتم در برابر نویز و داده‌های پرت مقاوم است در حالی که در روش‌های سنتی نیاز به پیش‌پردازش برای رفع این موارد است.
- ارائه یک رتبه‌بندی برای ژن‌ها.
- نیازی به گذاشتن فرض بر روی توزیع داده‌ها ندارد. بسیاری از روش‌های آماری فرضیاتی بر روی داده‌ها دارند و در صورت نقض این مفروضات قابلیت اطمینان آنها کاهش می‌یابد.

بنابراین ما نیز در ادامه تحلیل افتراقی خود از این روش‌ها استفاده کردیم تا بهتر بتوانیم ژن‌ها را تفکیک کنیم.

در هر گرهی درخت تصمیم برای انتخاب یک فیچر می‌توان از معیارهای مختلفی استفاده کرد. یکی از این معیارها Gini index است که برای اندازه‌گیری نابرابری و خلوص یک مجموعه داده استفاده می‌شود. در یک گره از درخت تصمیم فیچری را انتخاب می‌کنیم که بهتر بتواند گره‌هایی با خلوص بالا ایجاد کند. در واقع به دنبال فیچری هستیم که بتواند اختلاف Gini پدر از میانگین وزن دار Gini فرزندان را بیشینه کند. در یک جنگل تصادفی، هر بار که یک ژن انتخاب می‌شود، میزان اختلاف گفته شده برای آن ذخیره می‌شود و در نهایت میانگین این اختلاف‌ها به عنوان میزان اهمیت و اثرگذاری آن استفاده می‌شود.

$$\text{Gini} = 1 - \sum p_i^2$$



آ_۶ SBR

SBR یک سیستم درجه‌بندی هیستوپاتولوژیکی است که برای طبقه‌بندی سرطان سینه از نظر بدخیمی و شدت بیماری استفاده می‌شود. این سیستم سه فاکتور مهم بافت‌شناسی را بررسی می‌کند:

- تشکیل ساختارهای غددی: شباهت سلول‌های سرطانی به سلول‌های طبیعی.

- تفاوت‌های هسته‌ای (پلئومورفیسم): میزان غیرطبیعی بودن هسته سلول‌های سرطانی.

- نرخ تقسیم سلولی (تعداد میتوزها): تعداد سلول‌هایی که در حال تکثیر هستند.

براین اساس وضعیت تومور به سه حالت تقسیم می‌شود که از وضعیت نسبتاً عادی تا بدخیمی بالا و رشد بالا و رفتار تهاجمی گسترش می‌یابد.