گزارش پروژه‌ی نهایی درس مقدمه‌ای بر بیوانفورماتیک
هستی‌شناسی ژن‌های شاخص (با بیان بالا) در نمونه‌های سرطان خون (لوکمیا)

اساتید درس: دکتر علی شریفی زارچی، دکتر سمیه کوهی
نگارنده: کیان باختری

bakhtari.kian@gmail.com

۱۸ بهمن ۱۴۰۰

۱ مقدمه

لوسمی حاد مغز استخوان (Acute Myeloid Leukemia) هشتمین سرطان کشنده‌ی جهان است. علی‌رغم این که در میان سرطان‌ها رایج نیست و حدود یک درصد از این بیماری‌ها را تشکیل می‌دهد، شانس درمان به نسبت پایین‌تری دارد و فقط ۲۶ درصد از افرادی که در سنین بالای ۲۰ سال به این بیماری دچار می‌شوند تا بیش از پنج سال پس از تشخیص بیماری زنده می‌مانند. در افراد زیر ۲۰ سال این عدد به ۶۸ درصد می‌رسد. مغز استخوان مبتلایان این بیماری مقدار زیادی سلول‌های خونی غیر عادی تولید می‌کند که روی گلبول‌های قرمز و سفید و همچنین پلاکت‌های خون اثر منفی می‌گذارد. درمان‌های معمول برای انواع سرطان مانند شیمی‌درمانی یا پرتو درمانی برای درمان این بیماری نیز استفاده می‌شوند.

هستی‌شناسی (Ontology) ژن‌های شاخص در سلول‌های مغز استخوان بیمارانی که به لوسمی حاد مغز استخوان مبتلا هستند می‌تواند روابط پنهان ژنتیکی که منجر به ظهور این بیماری می‌شوند را آشکار کرده و منجر به ایجاد روش‌های تازه‌ای برای پیشگیری و درمان این بیماری شود. در این پروژه، یک مجموعه‌ی داده‌ی ریزآرایه حاوی اطلاعات چندین نمونه‌ی سالم و مبتلا به لوسمی حاد مغز استخوان بررسی می‌شود و تلاش می‌شود تا ژن‌هایی که به گونه‌ی معنی‌داری در نمونه‌های بیمار بیان شده‌اند شناسایی شوند. همچنین سعی شده است تا با تطبیق دادن ژن‌های مذکور با پایگاه داده‌های زیستی موجود، هستی‌شناسی این ژن‌ها مورد بررسی قرار گرفته و مروری شود بر مقالاتی که تا به امروز ارتباطی معنی‌دار میان این ژن‌ها و این بیماری گزارش کرده‌اند. در کنار این گزارش، یک فایل به زبان R وجود دارد که تمامی تحلیل‌ها و نمودارهای ارائه شده در این گزارش توسط کدهای موجود در فایل مذکور به دست آمده‌اند.

۲ مجموعه‌ی داده‌ها

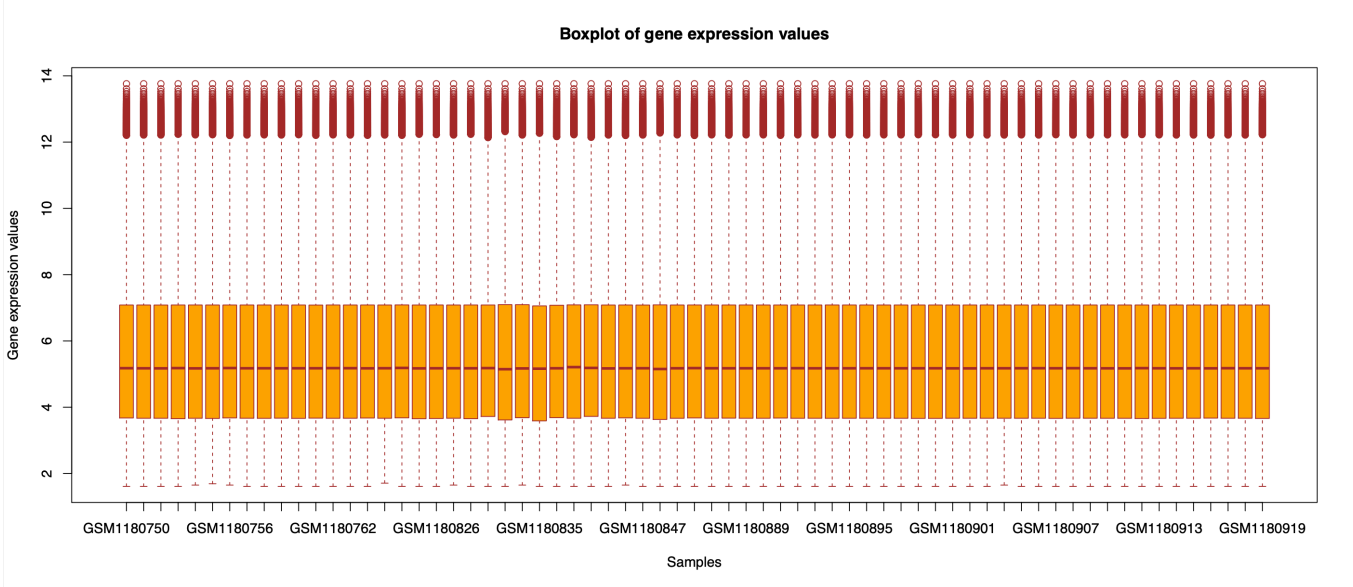
داده‌هایی که در این پروژه مورد بررسی قرار گرفته‌اند، توسط دانشگاه مریلند تهیه شده و در وب‌سایت Gene Expression Omnibus در این لینک قرار گرفته‌اند. این مجموعه داده حاوی ۱۷۰ پروفایل بیان ژن می‌باشد که با استفاده از ریزآرایه (microarray) به دست آمده‌اند. از میان این ۱۷۰ نمونه، ۴۹ نمونه‌ی سالم و ۱۸ نمونه‌ی مبتلا به لوسمی حاد مغز استخوان جدا شده و مورد بررسی قرار گرفته‌اند.

۲-۱ کنترل کیفیت داده

در بخش کنترل کیفیت داده اطمینان حاصل شد که داده‌ها نرمال‌سازی شده و داده‌ی پرت نداشته باشند. نمودار جعبه‌ای پروفایل بیان ژن برای ۶۷ نمونه‌ی مورد بررسی در شکل ۱ قابل مشاهده است. از این نمودار این طور برداشت می‌شود که داده‌ها از پیش نرمال شده‌اند و توزیع بیان ژن‌ها در نمونه‌های مختلف تقریباً یکسان است. در صورتی که اختلاف غیر قابل قبولی در توزیع مذکور وجود می‌داشت، می‌توانستیم با استفاده از Quantile Normalization عمل نرمال‌سازی را انجام دهیم. همچنین داده‌ها مقدار گم‌شده یا نامعلوم نداشتند و به این ترتیب از کیفیت قابل قبولی برای شروع تحلیل برخوردار بودند.

۳ کاهش ابعاد

با توجه به ابعاد بالای داده‌ها (۶۷ نمونه و ۳۲۳۲۱ پروب ریزآرایه) نیاز است که برای تفکیک نمونه‌های سالم و سرطانی در فضای چند ده هزار بعدی ژن‌ها، کاهش ابعاد صورت گیرد. در این بررسی از دو روش مرسوم کاهش ابعاد، یعنی PCA و tSNE استفاده شده است.

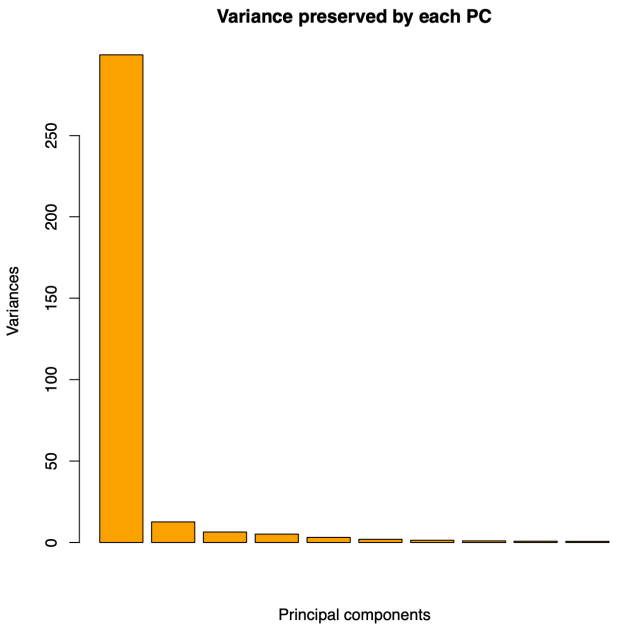


شکل ۱: نمودار جعبه‌ای پروفایل بیان ژنی برای ۶۷ نمونه‌ی مورد استفاده. داده‌ها نرمال شده و آماده‌ی استفاده و تحلیل هستند.

PCA ۱-۳

در روش PCA ابعاد به گونه‌ای کاهش داده می‌شوند تا ضمن رسیدن به ابعاد دلخواه (معمولاً ۲ یا ۳) حداکثر واریانس ممکن حفظ شود. به هدف نمایش دادن نمونه‌ها در دو بعد، ابعاد فضای نمونه‌ها را به ۲ کاهش دادیم. در شکل ۲ مشاهده می‌شود که هر مولفه‌ی اصلی (prin-cipal component) چه مقدار از واریانس را در راستای خودش حفظ کرده است و در شکل ۳ حاصل نگاشت نقاط به فضای دو بعدی کاهش یافته قابل مشاهده است. همان طور که از این نمودارها پیدا است، بخش قابل توجهی از واریانس در مولفه‌ی اول نگه‌داری شده است. این به آن علت است که میزان بیان برخی ژن‌ها در همه‌ی نمونه‌ها پایین و میزان بیان گروهی دیگر از ژن‌ها در تمامی نمونه‌ها بالا است. به همین جهت در راستای دور شدن از مبدأ مختصات مقدار قابل توجهی واریانس دیده می‌شود. برای رفع مشکل، مقدار میانگین بیان هر ژن در نمونه‌های متفاوت از مقادیر بیان آن ژن در هر نمونه کسر شده است تا تنها اختلافی که باقی می‌ماند اختلاف معنی‌دار در بیان آن ژن باشد. پس از اعمال این تغییر، شکل‌های ۴ و ۵ به دست می‌آیند که نمایانگر کارایی این روش بوده و به خوبی تفکیک قابل قبول‌تری را ارائه می‌دهند.

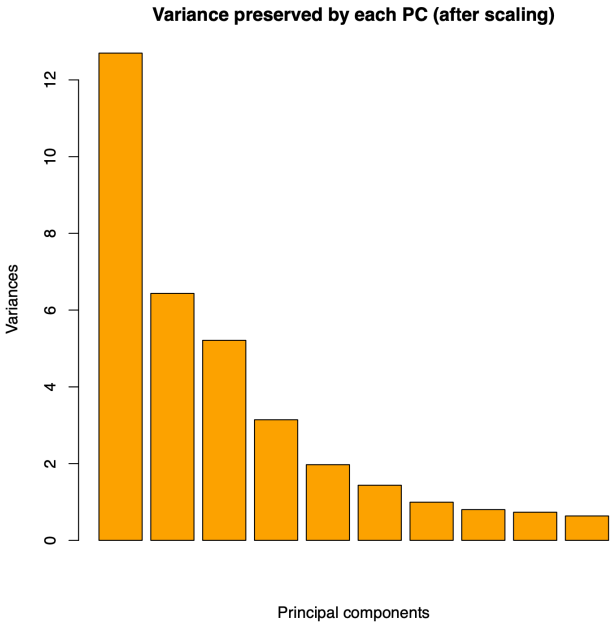
حال برای آن که بتوان دید با استفاده از کاهش ابعد به روش PCA سلول‌های سالم چه مقدار از سلول‌های سرطانی جدا می‌شوند، فضای ژن‌ها را به دو بعد کاهش داده و نگاشت نمونه‌ها به فضای کاهش بعد داده شده را رسم می‌کنیم. در شکل ۶ می‌توان مشاهده کرد که نمونه‌های سالم از نمونه‌های سرطانی تا حدودی جدا شده‌اند، به این معنا که سلول‌های سرطانی با تراکم بیشتری نزدیک به یکدیگر قرار گرفته‌اند. در این شکل گروه صفر سلول‌های سرطانی و گروه ۱ سلول‌های سالم را نشان می‌دهد.



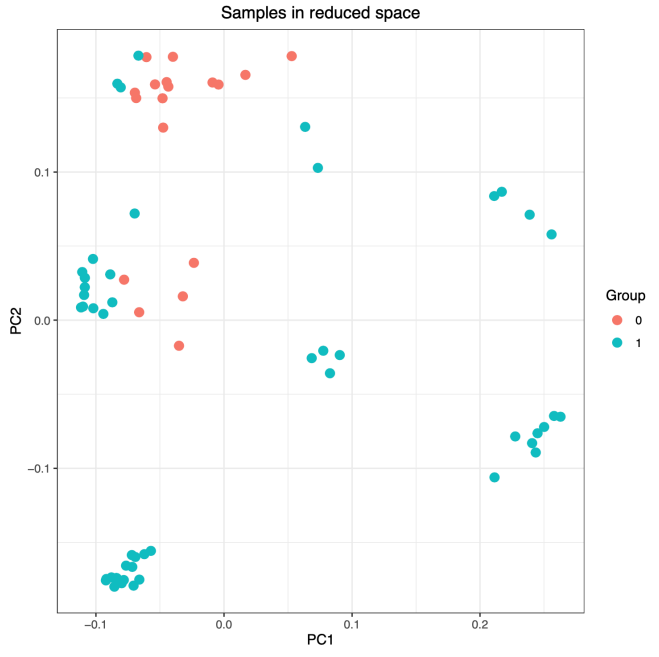
شکل ۲: میزان واریانس نگه‌داری شده در راستای هر مولفه‌ی اصلی (PC) بدون انجام scaling

شکل ۳: ژن‌های نگاشت شده به فضای کاهش بعد یافته (بدون انجام scaling)

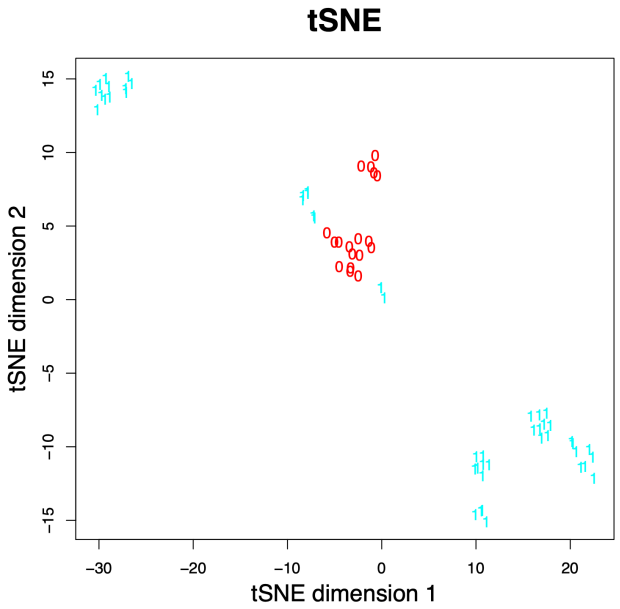
شکل ۵: ژن‌های نگاشت شده به فضای کاهش بعد یافته (با انجام scaling)



شکل ۴: میزان واریانس نگه‌داری شده در راستای هر مولفه‌ی اصلی (PC) با انجام scaling



شکل ۶: نمونه‌های نگاشت شده به فضای کاهش یافته‌ی دو بعدی با استفاده از PCA . در این شکل گروه صفر نشان‌دهنده‌ی سلول‌های سرطانی و گروه یک نشان‌دهنده‌ی سلول‌هاس سالم هستند.



شکل ۷: نگاشت نمونه‌ها به فضای کاهش یافته با استفاده از الگوریتم t-SNE . گروه صفر (قرمز رنگ) گروه سلول‌های سرطانی و گروه یک (آبی رنگ) گروه سلول‌های سالم هستند.

۳-۲ t-SNE

کاهش ابعاد به روش t-SNE بلقوه می‌تواند نتایج بهتری نسبت به روش PCA به دست دهد. در این روش فاصله‌ی توپولوژیکی میان دو نقطه مد نظر قرار گرفته و تلاش می‌شود تا نگاشتی دو یا سه بعدی از نقاط ارائه شود که فاصله‌ی هر دو نقطه در فضای جدید متناسب با فاصله‌ی آن‌ها در فضای اصلی باشد. در شکل ۷ نتیجه‌ی اجرای الگوریتم t-SNE با پارامتر perpelexity برابر با ۱۰ نشان داده شده است. این الگوریتم با مقادیر perpelexity برابر با ۵ و ۲۰ هم آزموده شد ولی نتایج بهتری به دست نیامد. در شکل ۷ نقاط صفر به رنگ قرمز نشان‌دهنده‌ی سلول‌های بیمار و نقاط یک به رنگ آبی نشان‌دهنده‌ی سلول های سالم هستند. مشاهده می‌شود که سلول‌های سرطانی چقدر نزدیک به یکدیگر تجمعی چشم‌گیر دارند که به تبع به این معنا است که الگوریتم t-SNE از روش PCA نتایج بهتری به دست داده است.

۴ همبستگی

بررسی همبستگی میان نمونه‌ها می‌تواند نشان دهد که نمونه‌های سالم با یکدیگر و یا نمونه‌های سرطانی با یکدیگر چقدر شباهت در پروفایل بیان ژن دارند. با رسم نقشه‌ی حرارتی یا همان heatmap برای ماتریس هبستگی، می‌توان دید که نمونه‌های سالم و سرطانی هر یک با هم‌نوعان خود رابطه‌ی نزدیک‌تری دارند تا با گروه دگر. در شکل ۸ و ۹ به ترتیب نقشه‌ی حرارتی ماتریس همبستگی برای تمامی ۱۷۰ نمونه‌ی موجود در مجموعه‌ی داده و برای ۶۷ نمونه‌ی مورد بررسی نمایش داده شده است. کاهش ابعاد می‌تواند نقش موثری در نمایان شدن همبستگی میان نمونه‌ها داشته باشد. در شکل ۱۰ نقشه‌ی حرارتی ماتریس همبستگی برای نمونه‌ها قابل مشاهده است که پس از کاهش ابعاد با روش PCA به دست آمده است. می‌توان دید که مقادیر همبستگی معنی‌دار تری در این نقشه وجود دارد. در شکل ۱۱ نقشه‌ی حرارتی ماتریس همبستگی پس از کاهش ابعاد با روش t-SNE به نمایش درآمده است که حتی همبستگی‌های قوی را نشان می‌دهد.

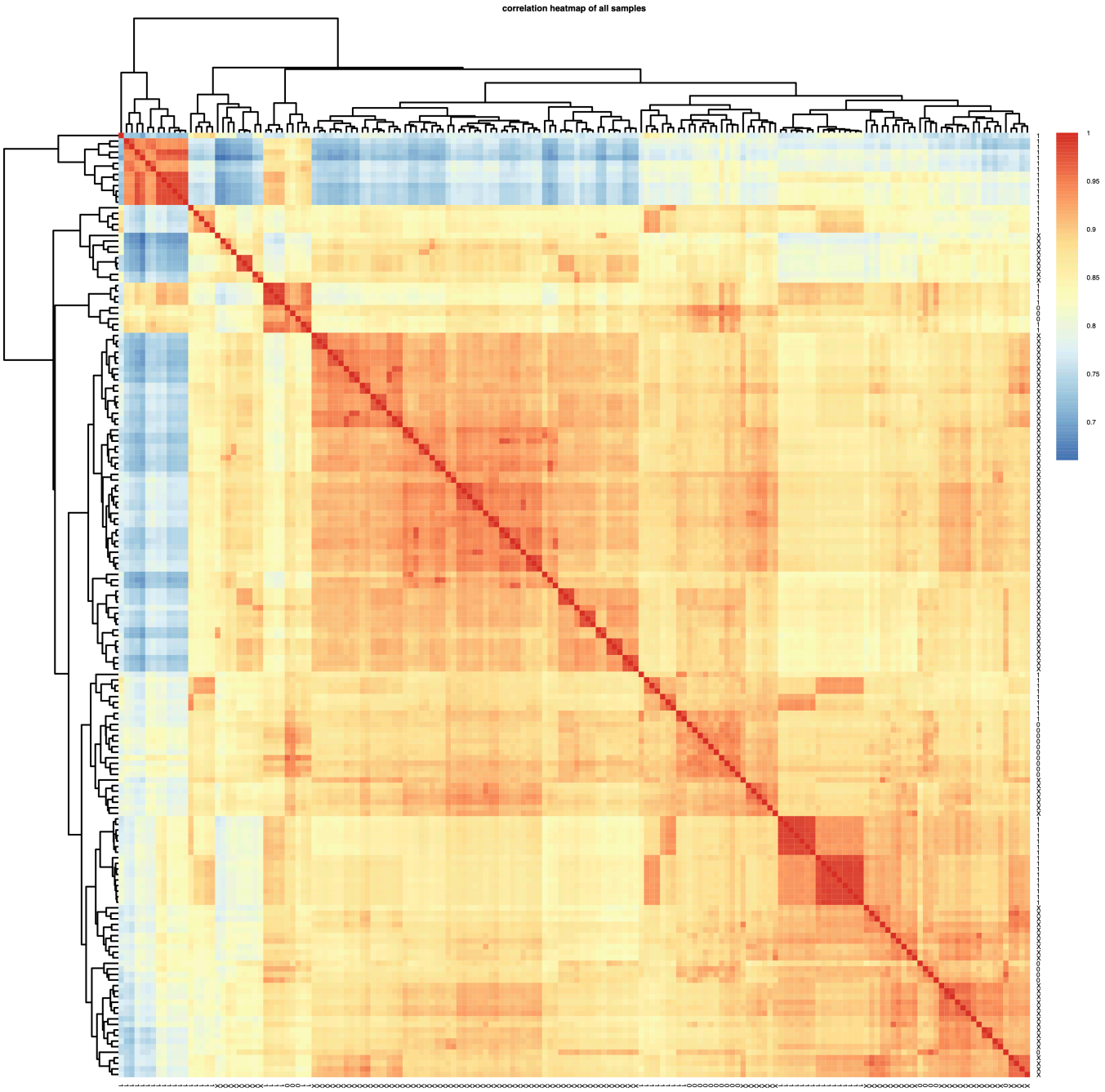
۵ تمایز در بیان ژن‌ها

با استفاده از محاسبه‌ی adjusted P value برای نمونه‌های سالم و سرطانی و همچنین محاسبه‌ی LogFC می‌توان به جدولی دست یافت که نشان می‌دهد هر کدام از ژن‌های حاضر در مجموعه‌ی داده آیا تفاوت معنی‌داری در میزان بیان‌شان در این دو گروه وجود دارد یا خیر. برای آن که معیاری مناسب برای تشخیص معنادار بودن تفاوت بیان داشته باشیم، مقدار adj-P-Val را کمتر از 0.05 و میزان LogFC را بزرگتر از ۱ یا کمتر از ۱- در نظر می‌گیریم. در شکل ۱۲ ۱۵ سطر ابتدایی جدولی به نمایش درآمده است که اطلاعات مربوط به ژن‌هایی که در دو گروه تفاوت بیان معنی‌دار دارند را نشان می‌دهد. از نتایج و داده‌های این جدول استفاده می‌کنیم تا همه‌ی ژن‌هایی را بیابیم که بین دو گروه نمونه‌ها تفاوت بیان معنی‌داری دارند. در ادامه به هستی‌شناسی این ژن‌ها پرداخته می‌شود.

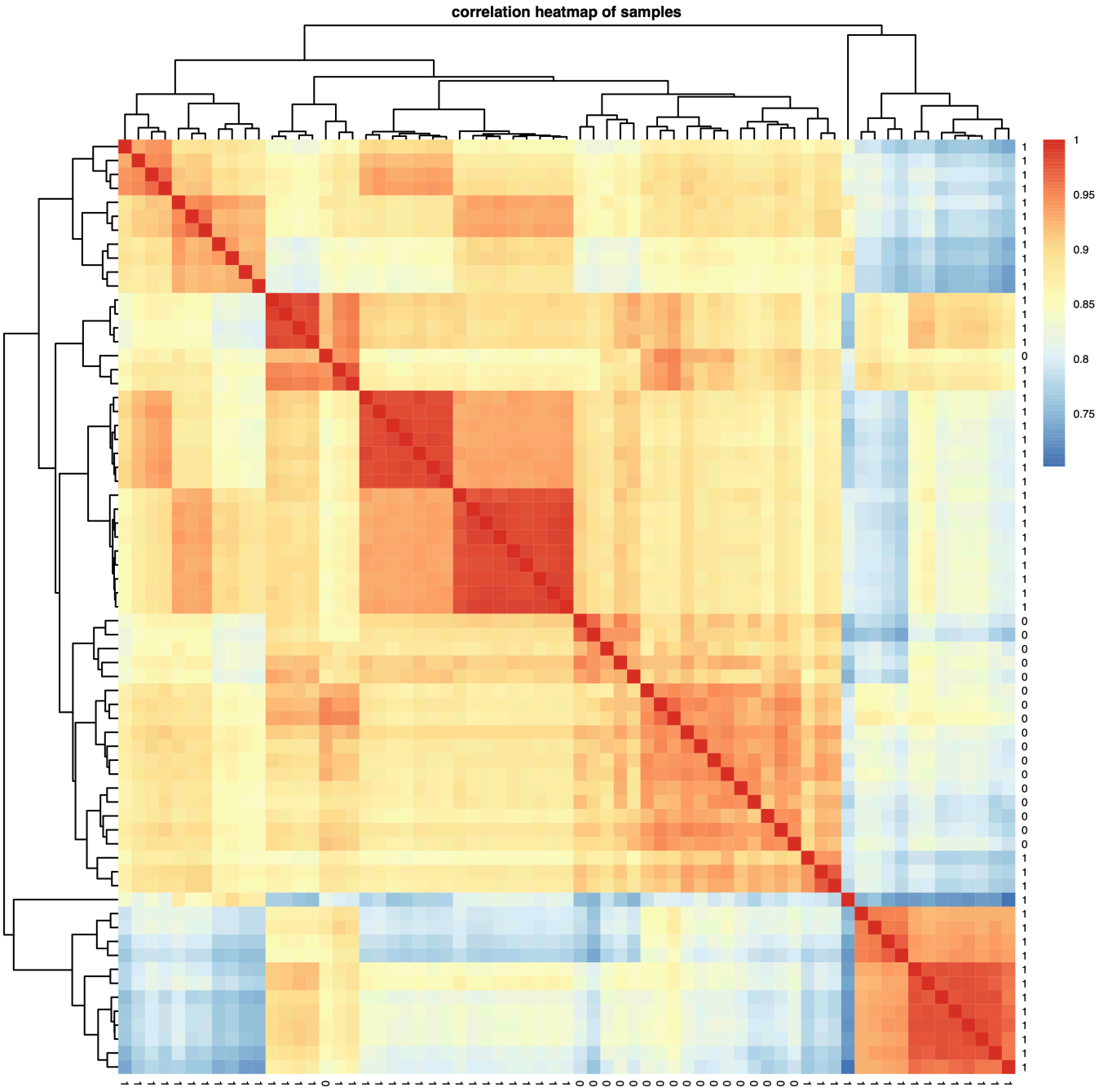
۶ هستی‌شناسی

ژن‌های به دست آمده از قسمت قبل را با استفاده از پایگاه داده‌ی En-richr با بایو مارکر های شناخته شده تطبیق می‌دهیم تاامکان مطالعه‌ی pathway ها و هستی‌شناسی این ژن‌ها فراهم شود. در کنار کدهای این پروژه، دو فایل با نام‌های aml-up و aml-down وجود دارند. این فایل‌ها به ترتیب حاوی ژن‌هایی هستند که به تشخیص و تحلیلی که در این پروژه صورت گرفته در بیماری لوسمی حاد مغز استخوان ژن‌های تاثیرگذار و شاخصی هستند. پس از تطابق ژن‌های موجود در فایل aml-up با پایگاه داده‌ی Enrichr چندین بایو مارکر مرتبط با بیماری لوسمی حاد مغز استخوان در نتایج ظاهر می‌شوند. به عنوان مثال E2F4 ENCODE یک transcription factor است که با تعداد زیادی از ژن‌های پیدا شده توسط بررسی‌های این پروژه ارتباط دارد و با مقدار P از مرتبه‌ی ده به توان ۷۹- بیان بسیار متفاوتی در گروه‌های سالم و سرطانی دارد. همان طور که در شکل ۱۳ دیده می‌شود، به پیشنهاد Enrichr ستون اول ماتریس خروجی بایو مارکر قابل توجهی بوده که همان E2F4 ENCODE می‌باشد. در این مقاله که در سال ۲۰۲۰ در ژورنال PubMed منتشر شده است، این موضوع مورد بررسی و گزارش قرار گرفته است که E2F4 ENCODE یک سرکوب‌کننده‌ی بلقوه برای لوسمی حاد مغز استخوان است و می‌توان از آن برای کاربردهای درمانی بهره برد.

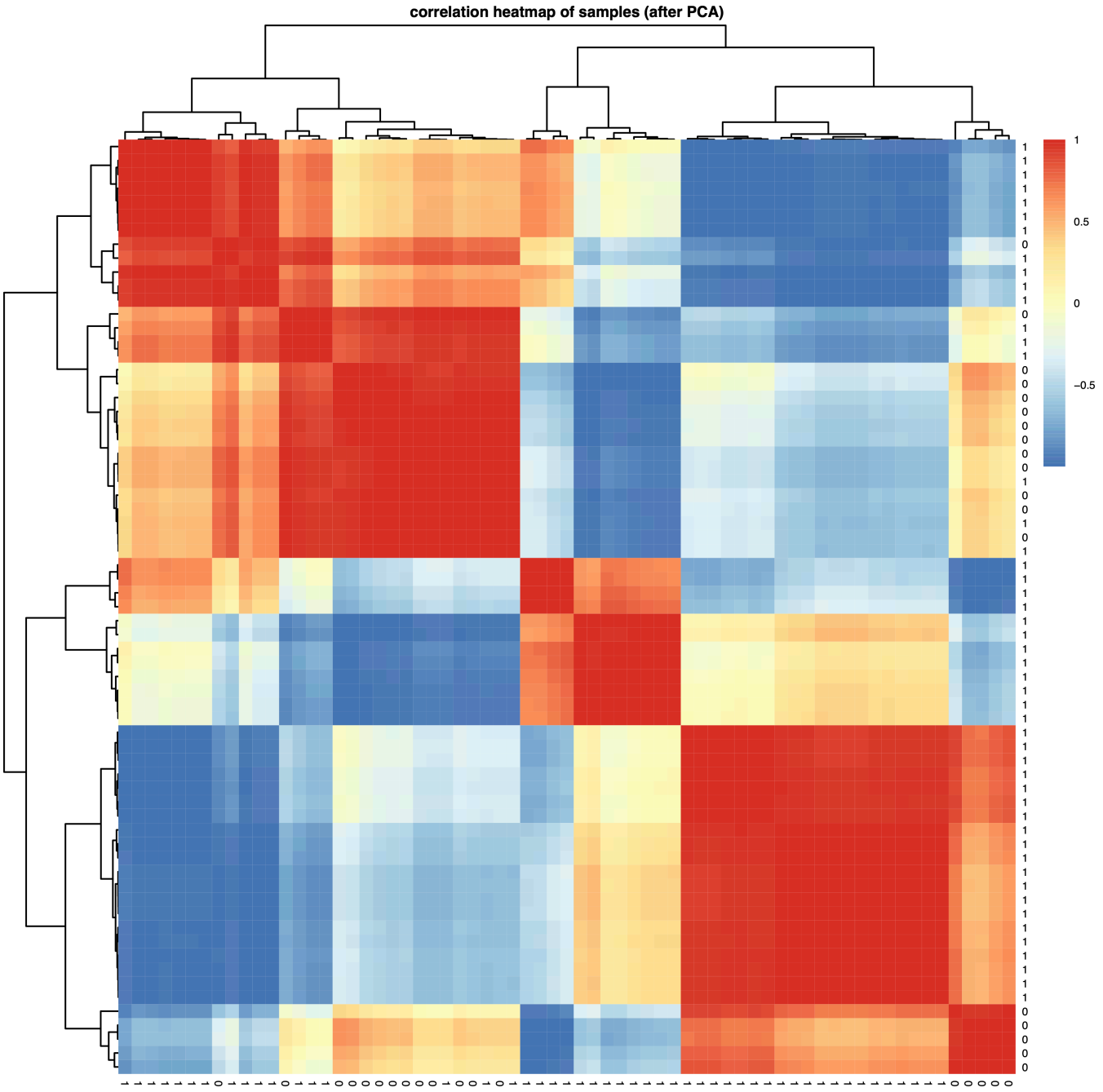
در نمونه‌ی دیگری، از نتایج تحلیل pathway ها به دست آمد که پروتئین FOXM1 یک بایو مارکر تقویت‌کننده‌ی بیماری AML است.



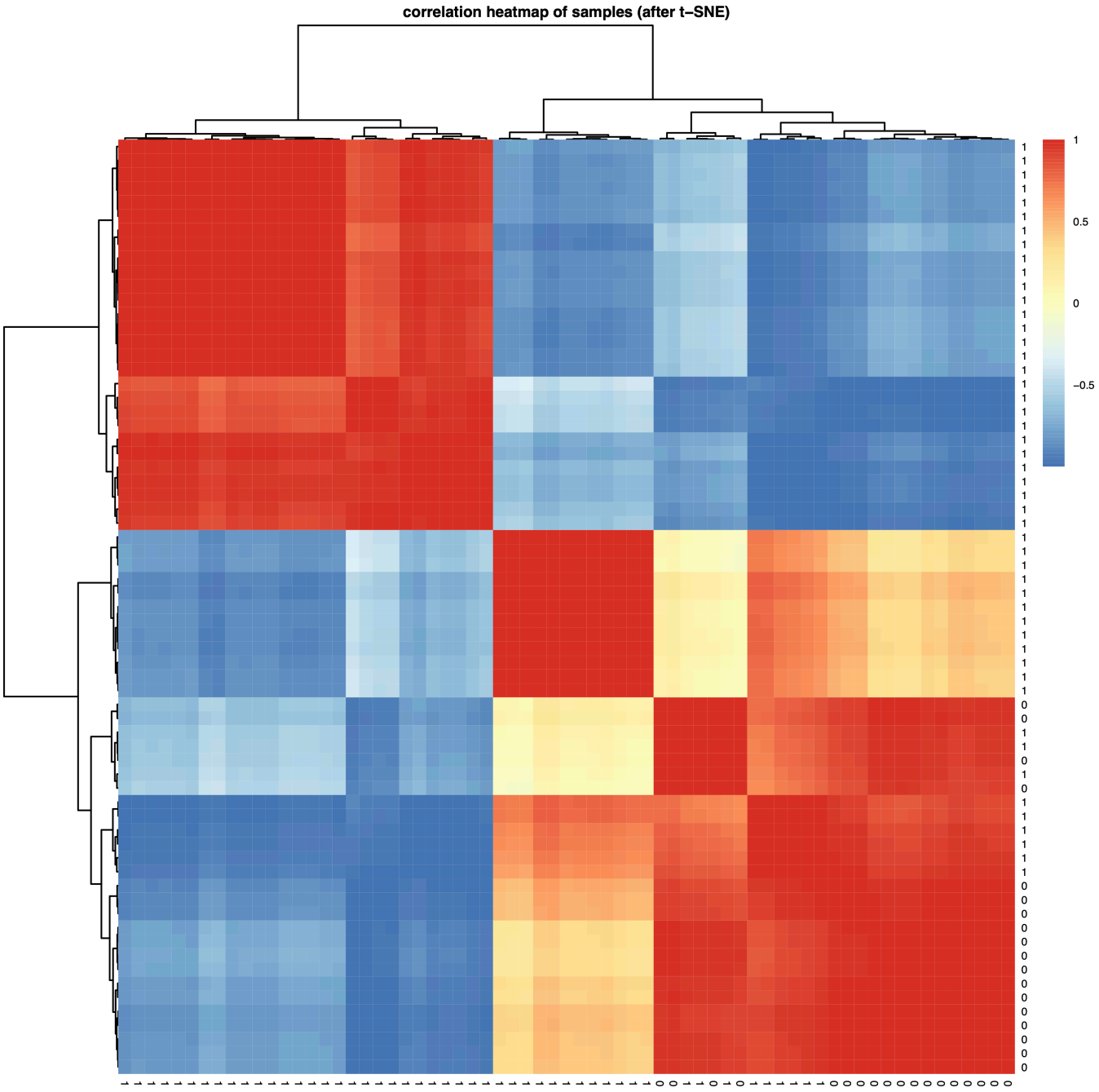
شکل ۸: نقشه‌ی حرارتی ماتریس همبستگی برای ۱۷۰ نمونه‌ی موجود در مجموعه‌ی داده



شکل ۹: نقشه‌ی حرارتی ماتریس همبستگی برای ۶۷ نمونه‌ی مورد بررسی



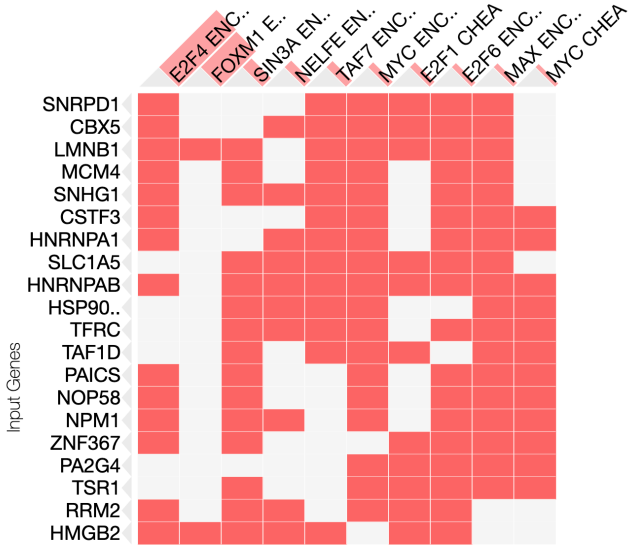
شکل ۱۰: نقشه‌ی حرارتی ماتریس همبستگی برای ۶۷ نمونه‌ی مورد بررسی پس از کاهش ابعاد با روش PCA



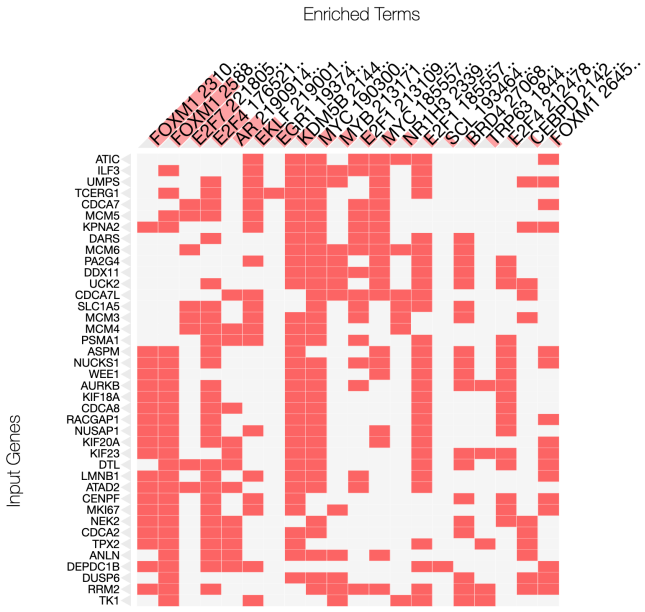
شکل ۱۱: نقشه‌ی حرارتی ماتریس همبستگی برای ۶۷ نمونه‌ی مورد بررسی پس از کاهش ابعاد با الگوریتم t-SNE

	Gene.symbol	ID	adj.P.Val	logFC
0	MPO	8016932	3.617813e-19	5.563501
1	FLT3	7970737	4.835716e-19	5.250065
2	KIAA0101	7989647	6.308160e-19	4.559135
3	BUB1B	7982663	1.664043e-18	2.756554
4	SUCNR1	8083422	1.938573e-18	2.996816
5	MCM10	7926259	3.712137e-18	2.318848
6	TPX2	8061579	4.695529e-18	3.156415
7	CIT	7966878	1.147946e-17	2.370751
8	CDC45	8071212	1.658665e-17	2.287501
9	IQGAP3	7921033	1.775540e-17	1.669697
10	POLQ	8089875	1.775540e-17	2.091978
11	CPXM1	8064539	6.592365e-17	3.776954
12	STK38	8126018	7.228461e-17	-1.880433
13	ANLN	8132318	7.504685e-17	2.641046
14	PRC1	7991406	7.504685e-17	3.080097
15	MELK	8155214	1.352640e-16	2.297318

شکل ۱۲: ۱۵ سطر ابتدایی جدولی که اطلاعات ژن‌های شاخص را در بر دارد



شکل ۱۳: تعدادی از بایو مارکر هایی که به پیشنهاد Enrichr با ژن‌های به دست آمده از آزمایش‌های این پروژه ارتباط معنی‌دار دارند.



شکل ۱۴: ارتباط بالای FOXM1 با لوسمی حاد مغز استخوان

بر اساس نتایج گزارش شده توسط این مقاله پروتئین مذکور در تحریک ژن‌هایی که باعث ایجاد مقاومت دارویی می‌شوند نقش داشته و از این طریق نوعی حمایت‌کننده‌ی لوسمی حاد مغز استخوان به شمار می‌رود. شکل ۱۴ بخشی از خروجی Enrichr را نشان می‌دهد که ستون اول آن FOXM1 است. شکل ۱۵ بخش دیگری از خروجی Enrichr را نشان می‌دهد. ستون سوم مربوط به مارکر EZH2 می‌باشد که یک آنزیم methyltransferase است و توسط نتایج آزمایش این پروژه با لوسمی حاد مغز استخوان ارتباط تقویت‌کنندگی دارد. این ارتباط قبلا در مقاله‌ای از Haematologica در سال ۲۰۲۰ به چاپ رسیده و از طریق این لینک قابل دسترسی است. در این مقاله پژوهشگران به این نتیجه رسیده‌اند که وجود جهش در EZH2 به گونه‌ی معنی‌داری با ابتلا به بیماری AML ارتباط دارد. این مطالعه روی بیش از ۱۶۰۰ بیمار مبتلا به لوسمی حاد مغز استخوان انجام شده و مشاهده شده است که کسر قابل توجه‌ای از این بیماران دارای جهش در مارکر مذکور هستند.

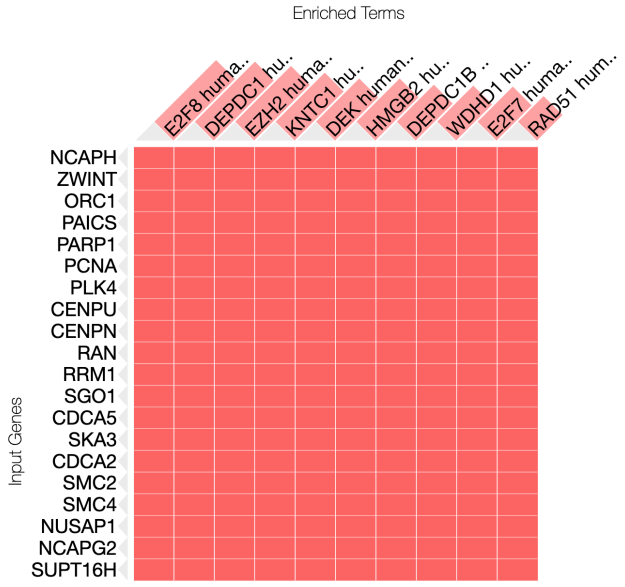
در سال گذشته مقاله‌ای توسط PubMed منتشر شد با این عنوان:

IRF8 is a Reliable Monoblast Marker for Acute Monocytic Leukemias

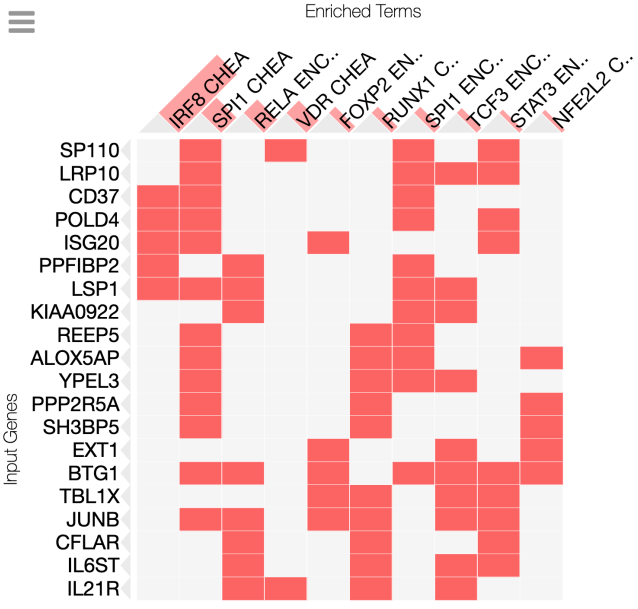
این مقاله در این لینک قابل دسترسی است. همان طور که از عنوان مقاله پیدا است، IRF8 به عنوان یک مارکر قوی برای لوسمی حاد مغز استخوان معرفی شده است. اما این مارکر در نتایج آزمایش و تحلیل انجام شده توسط این پروژه نیز ظاهر شده است. شکل ۱۶ بخش دیگری از خروجی Enrichr به داده‌های ما است که در ستون اول IRF8 قرار دارد و به عنوان یک مارکر با ارتباط قوی با داده‌های ما معرفی شده است.

۷ نتیجه‌گیری

همان طور که از انطباق نتایج و تحلیل‌های به دست آمده از آزمایش این پروژه با پایگاه‌های داده‌ی زیستی و مقایسه‌ی آن‌ها با مقالات اخیر در حوزه‌ی gene ontology این نتیجه حاصل شد که نتایج معنی‌دار و قابل اتکایی به دست آمده‌اند، می‌توان مشاهده کرد که حوزه‌ی بیوانفورماتیک می‌تواند تاثیری بسیار قابل توجه و انکار ناپذیر در پیشگیری و درمان سخت‌ترین بیماری‌ها داشته باشد.



شکل ۱۵: ارتباط بالای EZH2 با لوسمی حاد مغز استخوان (ستون سوم)



شکل ۱۶: ارتباط بالای IRF8 با لوسمی حاد مغز استخوان (ستون اول)