

00_Covfilecreator.R

ساخت جدول هممتغیر استاندارد (**CombinedCovariates.csv**) با پاکسازی/نرمال‌سازی مقادیر، تشخیص **Platform** و ثبت خانه ی خالی

NA. 

01_Covfilecreator.R

نسخه ی بهروزشده ی همان هدف: تولید **CombinedCovariates.csv** با همان قراردادها و کنترل‌های کیفیت .

02_Pvalue.R

آزمون وابستگی زوجی بین کوارییتات (Fisher/Chi-square) با مونت‌کارلو Spearman + برای متغیرهای ترتیبی (همراه با **level_counts**)

MANIFEST.  و **CSV** 

03_geo_combat_umap_O.R

برای زیرمجموعه ی **UMAP** با حذف نمونه‌های بافت نامشخص (Normal/Tumor) فقط **Microarray** و ترسیم **ComBat**

Batch×Tissue.  و جدول **GEO_ComBat_matrix.csv**

04_normalize_and_merge.R

فقط روی بچه‌های چندنمونه‌ای (singleton) ها دست‌نخورده، حفظ سطح **Unknown** در بافت، و **UMAP/PCA** در صورت کفايت

نمونه؛ خروجی **GEO_ComBat_matrix.csv**  و **GEO_UMAP_4panel.png**.

05_all_combat.R

سراسری با تفکیک **ComBat** مدل انعطاف‌پذیر (Sex/Stage/Age) **stratify-by-tissue** و **multi vs singleton** با حذف تدریجی در تعارض؛

ALL_UMAP_4panel.png.  و **All_ComBat_matrix.csv** خروجی

09_qc_outliers_scaling.R

QC پیش/پس: نمودار چگالی و باکس، **Tissue** رنگ‌شده با **PCA/UMAP** ، و شناسایی آوت‌لایرها با **RPCA + IQR** گزارش

retained/removed.  و فهرست نمونه‌های **Outliers_report.csv**

04_edge.R

تحلیل **DGE** برای **Tumor vs Normal** (سراسری و درون‌گروهی)، تولید **Volcano** و **Heatmap** و ذخیره ی نتایج  **CSV**.

امتیازدهی مسیرها با limma (Tumor–Normal) روى امتیازها (GOBP) . انجام **GSVA** (H, KEGG, REACTOME, GOBP)

آماده‌سازی خروجی‌های شکه و تلفیق 

06_gsva.R

12_wgcna_modules.R

ساخت شبکه **WGCNA** روی ژن‌های پُرواریانس، استخراج رنگ مدول‌ها/ایگن‌مدول‌ها، همبستگی **Module-Trait** و (در صورت وجود) تحلیل بقا؛ خروجی‌های CSV/ نمودار .

13_network_variants.R

شبکه‌های جایگزین در **Tumor/Normal** با **ARACNE/CLR/MRNET/PCIT** و محاسبه **Jaccard** همپوشانی یال‌ها؛ خروجی **YAL**‌ها به ازای هر روش/گروه  .

11_networks.R

بازسازی شبکه‌های **ARACNE** به تفکیک زیرگروه‌های دموگرافیک/پاتولوژیک با **residualization** مبتنی بر کوواریتی‌های بقایی معنی‌دار؛ خروجی **YAL**‌ها، درجات، و گراف پیش‌نمایش  .

12_0network.R

نسخه **MI/ARACNE** ساده و سریع شبکه‌سازی برای **sanity-check** و **baseline**؛ بازگذاری سبک کواریتی‌ها و گراف‌های پیش‌نمایش  .

13_diffcoexp.R

برای شناسایی هم‌بیانی تفاضلی بین **Tumor** و **Normal** (Spearman + FDR) **DGCA**  .

14_integration_network_dge_gsva.R

یکپارچه‌سازی شواهد: هاب‌های شبکه **ME×GSVA** (ARACNE)؛ خروجی **DE** و همبستگی **Hub-Drivers** و هیتمپ مسیرهای شاخص  .

15_survival.R

تحلیل بقا کلاسیک **Kaplan-Meier** بر حسب دموگرافیک، **Cox** چندمتغیره تنظیم‌شده، و (اختیاری) **KM** بر پایه نمرات مسیر **GSVA**.  .

16_survival_advanced.R

پیش‌بینی بقا با **Cox LASSO** بر پایه **Module Eigengenes** (اختیاری هاب‌ژن‌ها)، محاسبه **Risk score** و **timeROC** بر اساس **KM** و **GSVA** ترتیبل ریسک  .

17_external_validation.R

اعتبارسنجی خارجی: بازخمین **MES** با رنگ‌های آموزش، اعمال ضرایب **CoxLASSO**، و گزارش **AUC** زمان‌مند و نمرات ریسک در گهورت مستقل  .

Covfilecreator.R_00

این اسکریپت یک جدول هممتغیر استاندارد می‌سازد تا برای تمام نمونه‌ها در کل پروژه قابل استفاده باشد. خروجی دقیقاً شامل ستون‌های زیر است

Sample
Platform
Tissue
Sex
Stage
LocationCode
LaurenCode
Age_Code

همه مقادیر مفقود به صورت خانه خالی نوشته می‌شود و عبارت NA در فایل خروجی دیده نمی‌شود

ورودی‌ها و مسیرها

پوشه پایه

E دو نقطه اسلش یک خط تیره ده بار GC زیرخط 29 زیرخط 4 زیرخط mainDATAFRAME خط تیره پنج بار یک فایل‌های ورودی

شامل سطرهای توصیفی ویژگی‌ها و ستون‌های نمونه‌ها Demo.xlsx
در صورت وجود برای استخراج فهرست نمونه‌های RNA_seq.xlsx
در صورت وجود برای استخراج فهرست نمونه‌های میکرواری Micro.xlsx
در صورت وجود برای تعیین ترتیب نمونه‌ها بر اساس سرستون‌های ماتریس بیان Full_matrix.csv
وجود Demo.xlsx اجباری است و در صورت نبودن اسکریپت متوقف می‌شود

وابستگی کتابخانه‌ها

برای خواندن اکسل readxl
سایر توابع پایه R برای پردازش متن و داده

نمای کلی جریان داده

یک انتخاب ترتیب نمونه بر اساس Full_matrix یا ترکیب سرستون‌های RNA و Micro RNA
خواندن جدول Demo و یافتن سطرهای مرتبط با هر ویژگی
پاکسازی و نormal‌سازی مقادیر رسته‌ای و عددی
پر کردن خلاًها با استفاده از فلگ‌های یک به یک در صورت وجود استنتاج مرحله کلی از TNM هنگام نیاز
اعمال قواعد دامنه نظیر خالی کردن Stage برای بافت Normal
تشخیص Platform برای هر نمونه
ساخت دیتا فریم نهایی و ذخیره به صورت CSV با خانه‌های خالی به جای NA

توابع کمکی و نقش هر کدام

norm_key

نرم‌السازی کلیدهای متنی با تبدیل به حروف کوچک و تک فاصله و برش فضای اضافی. برای تطبیق اسمی سطر در Demo

find_row

گرفتن بردار نام و پژوهی‌ها و یک فهرست نام‌های کاندید و برگرداندن اندیس اولین تطبیق. هسته بازیابی سطر درست از Demo

clean_tissue

تبدیل انواع نگارش بافت به دو مقدار استاندارد Normal و Tumor. هر متن مرتبط با نرم‌السازی متن non adjacent normal یا tumor به Normal و هر متن حاوی سرطان به Tumor. در غیر این صورت مقدار مفقود

clean_sex

تبدیل نگارش‌های جنسیت مانند Female و male و f به Male و male و f به Female. موارد نامعتبر مفقود

clean_stage_overall

نرم‌السازی مرحله کلی بیماری به یکی از چهار سطح IV، III، II و I حروف زیر مرحله مانند C، B، A حذف می‌شوند. نگارش‌های عددی یک تا چهار نیز نگاشت می‌شوند

clean_M و clean_N و clean_T

استخراج زیرشاخص‌های TNM از متن آزاد. حذف نویسه‌های غیر عددی حروفی و جستجوی الگوی تی ان ام T، تیس و T4 تا T1 با تفکیک 4 و b4 و a و b3 و 3 و صفر تا سه با تفکیک 3 و 3 و صفر و یک M. نتیجه برای استنتاج مرحله کلی استفاده می‌شود

clean_location_code

نگاشت متن محل تومور به کد سه مقداری یک برای ناحیه پروگریمال و قسمت فوقانی یا GEJ و کاردیا و فاندوس دو برای تنه معده و میانی سه برای آنترو و بخش دیستال و پایینی ورودی عددی یک تا سه نیز پذیرفته می‌شود

clean_lauren_code

نگاشت رده لورن به کد diffuse یک برای intestinal دو برای indeterminate یا mixed سه برای ورودی عددی یک تا سه نیز پذیرفته می‌شود

clean_age_code

نگاشت سن یا کد سن به سه رسته یک برای کمتر از چهل دو برای چهل تا شصت

سه برای بالای شست

اگر متن عددی باشد تبدیل می‌شود و اگر کد نوشتاری مانند 40 یا 40-60 باشد نیز نگاشت می‌شود

merge_chr

ادغام مقدار تمیز شده با مقدار خام برای ستون‌های متنی. اگر مقدار تمیز شده خالی یا مفقود باشد مقدار خام پس از پاکسازی فضای خالی جایگزین می‌شود

merge_code

ادغام مقدار کد تمیز شده با مقدار خام برای ستون‌های عددی. اگر مقدار تمیز مفقود باشد تبدیل عددی امن از خام اعمال می‌شود

Truthy

تعابیر مقادیر باینری متنی به درست و نادرست برای فلگ‌ها. رشته‌هایی مانند یک و true و yes به درست و صفر و false و no و رشته خالی به نادرست و غیر از این مفقود

align

هم‌ترازی یک بردار به ترتیب نمونه‌ها. با اندیس دهی بر اساس نام ستون‌ها همان ترتیب ستون‌های بیان یا ترکیب RNA و Micro RNA حفظ می‌شود

fill_from_flags

پر کردن مقادیر مفقود در یک ستون هدف با استفاده از دو بردار فلگ مثبت و منفی نمونه‌ها

پر کردن sex_female و sex_male از Sex

پر کردن tissue_normal و tissue_tumor از Tissue

which_flag

استنتاج مرحله کلی از چهار فلگ I Stage تا IV Stage تنها زمانی مقدار تعیین می‌شود که دقیقاً یکی از فلگ‌ها درست باشد

منطق انتخاب نمونه‌ها

اگر موجود باشد سرستون‌های آن به عنوان ترتیب نهایی نمونه‌ها استفاده می‌شود و ستون اول را به عنوان ژن کار می‌گذارد

اگر موجود نباشد سرستون‌های Micro.xlsx و RNA_seq.xlsx خوانده می‌شود و مجموعه یکتا ساخته می‌شود. ترتیب برابر است با

نمونه‌های رنا به همان ترتیب و سپس نمونه‌های میکروواری که در رنا نبوده‌اند

اگر هیچ نمونه‌ای یافته نشود اجرای اسکریپت متوقف می‌شود

Demo و بازیابی سطرهای ویژگی

Demo.xlsx به دیتا فریم تبدیل می‌شود

ستون اول شامل نام ویژگی‌ها است

تابع grab_row با یک کلید اصلی و چند نام جایگزین تلاش می‌کند یک سطر را پیدا کرده و مقادیر آن سطر را به شکل یک بردار با نام‌های نمونه‌ها بازگرداند

برای هر ویژگی مهم یک بار grab_row فراخوانی می‌شود

Tissue

Sex
Stage
M و T
LocationCode
LaurenCode
Age_Code
Age

فلگ‌های اختیاری برای پر کردن خلا

sex_male
sex_female
stage_iv<stage_i
tissue_tumor
tissue_normal

در پایان هر بردار با align به ترتیب نهایی نمونه‌ها هم‌تراز می‌شود

پاکسازی و نرمال‌سازی اولیه

برای Stage و Sex و Tissue نسخه تمیز شده تولید می‌شود
Tissue_c
Sex_c
Stage_c

پر کردن خلا با فلگ‌ها

اگر Sex_c برای یک نمونه مفقود باشد و sex_male درست باشد مقدار Male گذاشته می‌شود و بر عکس برای Female اگر c مفقود باشد از فلگ‌های تومور و نرمال استفاده می‌شود

استنتاج از فلگ‌های مرحله

برای نمونه‌هایی که Stage_c مفقود است خروجی which_flag محاسبه می‌شود و اگر یک فلگ به تنها‌یی فعال باشد همان سطح به تخصیص می‌یابد

استنتاج اینم TNM از Stage

اگر Stage_c همچنان مفقود باشد و مقدار M برابر M1 باشد مرحله IV تعیین می‌شود
این قاعده بالینی سازگار است زیرا متاستاز دور دست به معنی مرحله چهار است

قاعده ویژه برای نمونه‌های نرمال

برای نمونه‌هایی که Tissue Normal برابر است ستون Stage عمداً خالی نگه داشته می‌شود
این کار مانع از تعریف مرحله برای نمونه‌های غیر توموری می‌شود

کدگذاری محل تومور و رده لورن و سن

با تابع تمیزکننده به عدد یک تا سه نگاشت می‌شود
با به عدد یک تا سه LaurenCode_c

با اولویت بردار Age_Code خام و در صورت نبودن از Age_Code استخراج می‌شود

ادغام مقادیر تمیز با خام

برای هر ستون
merge_chr با Stage و Sex و Tissue
merge_codey با Age_Code و LaurenCode و LocationCode

هدف این است که ابتدا مقدار استاندارد شده مصرف شود و فقط اگر آن مقدار خالی بود نسخه خام جایگزین شود

تعیین Platform برای هر نمونه

سرستون‌های Micro.xlsx و RNA_seq.xlsx بدون داده خوانده می‌شود

اگر نام نمونه در RNA دیده شود RNAseq Platform برابر

اگر در Microarray دیده شود Micro Platform برابر

اگر در هیچ کدام دیده نشود مقدار خالی ثبت می‌شود

ساخت دیتافریم نهایی

ستون‌ها دقیقاً به ترتیب زیر ساخته می‌شود

Sample

Platform

Tissue

Sex

Stage

LocationCode

LaurenCode

Age_Code

پارامتر check.names برابر False تا نام ستون‌ها دست نخورده بماند

خروجی و شیوه نوشتن

مسیر ذخیره

در پوشه پایه CombinedCovariates.csv

گرینه na برابر رشته خالی تا هیچ مقدار NA نوشته نشود

در پایان پیام تایید چاپ می‌شود که شامل تعداد سطر و ستون و مسیر فایل خروجی است

قراردادهای معنایی ستون‌ها

نام یکتا برای هر نمونه Sample

یکی از دو مقدار Microarray یا RNAseq Platform یا خالی

تنهای دو مقدار Normal یا Tumor یا خالی

تنهای دو مقدار Male یا Female یا خالی

فقط یکی از I یا II یا III یا IV برای نمونه‌های توموری. برای نمونه‌های نرمал خالی

کد سه مقداری ناحیه در معده. یک پروگزیمال. دو تنه. سه دیستانل

کد سه مقداری رده لورن. یک منتشر. دو روده‌ای. سه مختلط

کد سه مقداری سن. یک کمتر از چهل. دو بازه چهل تا شصت. سه بیشتر از شصت

کنترل‌های ورودی و توقف امن

عدم وجود Demo.xlsx باعث توقف با پیام خطای می‌شود
اگر Full_matrix موجود باشد ولی عنوان ستون‌ها کمتر از دو عدد باشد یعنی ساختار نادرست و توقف انجام می‌شود
اگر هیچ نمونه‌ای پیدا نشود توقف انجام می‌شود
در مرحله همترازی Demo اگر سطیری یافت نشود برداری از مقادیر مفقود برگردانده می‌شود که در مراحل بعدی مدیریت می‌شود

نکات کیفیت داده و تعارض‌ها

از TNM فقط در صورت M1 به IV ارتقا می‌یابد و از T و N برای سطوح دیگر استفاده نمی‌شود تا از استنتاج نادرست جلوگیری شود

فلگ‌های مرحله فقط زمانی معتبرند که دقیقاً یک فلگ فعال باشد
برای جنسیت و بافت اگر هر دو فلگ متناقض باشند مقدار تعیین نمی‌شود مگر آن که یکی درست و دیگری نادرست باشد

از هدر فایل‌های اکسل استخراج می‌شود و نیاز به انطباق دقیق نام نمونه‌ها میان فایل‌ها دارد

پیشنهادهای پایش و خطایابی

بررسی یکتایی ستون Sample در CombinedCovariates
تعداد مقادیر خالی در ستون‌های کلیدی شامل Stage و Tissue
همسوبی کامل نام نمونه‌ها با ماتریس بیان و با فایل‌های RNA و Micro
بازبینی چند مورد نمونه به صورت دستی برای صحت نگاشتهای Lauren و Location و Age_Code

پیچیدگی محاسباتی

ورود و خروج به اندازه شمار نمونه‌ها و سطرهای Demo خطی است
توابع پاکسازی رشته‌ای سربار ناچیز دارند
حافظه مصرفی عمدتاً به اندازه تعداد نمونه‌ها و شمار ویژگی‌های استخراجی است

مثال بسیار کوچک از ورودی و خروجی

فرض کنید سه نمونه با نام‌های S یک و S دو و S سه موجود است
برای Tumor Normal Tissue برابر Female Male یا برابر Sex
برای Stage Tumor یک مقدار مانند II و برای Normal خالی LocationCode
یک یا دو یا سه LaurenCode
یک یا دو یا سه Age_Code
در خروجی CombinedCovariates.csv دقیقاً همین ستون‌ها با همین مقادیر و خانه‌های خالی به جای مقادیر مفقود دیده می‌شود

Covfilecreator.R_01

ساخت یک فایل هم‌متغیر استاندارد برای تمام نمونه‌ها با ستون‌های دقیق زیر

Sample

Platform

Tissue

Sex

Stage

LocationCode

LaurenCode

Age_Code

هر مقدار مفقود به صورت خانه خالی در خروجی نوشته می‌شود و رشته NA دیده نمی‌شود

وروودی‌ها

مسیر پایه درایو E پوشه GC زیرخط 29 زیرخط 4 زیرخط mainDATAFRAME

شامل ردیف ویژگی‌ها و ستون نمونه‌ها Demo.xlsx

در صورت وجود برای فهرست نمونه‌های رنا RNA_seq.xlsx

در صورت وجود برای فهرست نمونه‌های میکروواری Micro.xlsx

در صورت وجود برای تحمیل ترتیب نمونه‌ها بر اساس سرستون‌های ماتریس بیان Full_matrix.csv

وجود Demo.xlsx اجباری است و در نبود آن اجرا متوقف می‌شود

وابستگی

بسته readxl برای خواندن اکسل

توابع پایه R برای پردازش متن و داده

خروجی

با ستون‌های دقیق که در بخش هدف آمده است CombinedCovariates.csv

خانه‌های خالی به جای مقادیر مفقود

پیام پایانی شامل تعداد سطر و ستون و مسیر ذخیره

نمای کلی فرایند

یک انتخاب ترتیب نمونه بر اساس Full_matrix در اولویت یا ترکیب سرستون‌های Micro RNA و RNA

بازیابی سطرهای مربوط به هر ویژگی از Demo با تطبیق نام قوی و غیرحساس به حروف

پاکسازی و نرمال‌سازی مقادیر متنی و کدی

پر کردن خلاً با استفاده از فلگ‌های یک به یک در صورت وجود

استنتاج مرحله کلی از فلگ‌ها و در صورت نیاز از مولفه TNM در M

قواعد دامنه شامل خالی کردن Stage برای نمونه‌های Normal

تشخیص Platform برای هر نمونه بر اساس حضور در فایل‌های Micro RNA یا RNA

ساخت جدول خروجی و نوشتمن به صورت CSV با خانه‌های خالی برای مقادیر مفقود

توابع کمکی و نقش آن‌ها

norm_key

یکنواختسازی کلیدهای متنی با کوچکسازی حروف و تک فاصله و حذف سفیدی اضافی
صرف برای تطبیق نام ردیفها در Demo

find_row

دربیافت بردار نام ویژگی‌ها به همراه فهرست نامهای کاندید
بازگرداندن اندیس اولین تطبیق
هسته یافتن ردیف درست در Demo

clean_tissue

تبدیل نگارش‌های مختلف بافت به Normal یا Tumor
عباراتی مانند non tumor و adjacent normal به Normal
عبارات حاوی tumor یا cancer یا carcinoma به Tumor
غیر از این مقادیر مفقود

clean_sex

تبدیل female و f به male و m به Female
مقادیر نامعتبر مفقود

clean_stage_overall

نگاشت مرحله کلی به یکی از I و II و III و IV
حذف حروف زیرمرحله مانند A و B و C
نگارش‌های عددی یک تا چهار نیز به چهار سطح استاندارد تبدیل می‌شوند

clean_M, clean_N و clean_T

استخراج شاخص‌های TNM از متن آزاد
حذف نویسه‌های نامعتبر و جستجوی الگو برای T و N و M
بازگرداندن مقادیر استاندارد مانند T1 تا T4 و N0 تا N3 و M0 یا M1

clean_location_code

نگاشت محل تومور به کد سه سطحی
یک برای ناحیه پروگزیمال یا GEJ یا کاردیا یا فاندوس یا قسمت بالایی
دو برای تنہ یا ناحیه میانی
سه برای آنترو یا قسمت دیستال یا بخش پایینی
ورودی عددی یک تا سه پذیرفته می‌شود

clean_lauren_code

نگاشت رده لورن به کد
یک برای diffuse
دو برای intestinal
سه برای mixed یا indeterminate
ورودی عددی یک تا سه پذیرفته می‌شود

clean_age_code

نگاشت سن یا کد سن به سه سطح
 یک برای کمتر از چهل
 دو برای بازه چهل تا شصت
 سه برای بالای شصت
 متون متدائل مانند 60-60+ یا 40-60+ نیز پشتیبانی می‌شوند

merge_chr

ادغام مقدار تمیز شده با مقدار خام برای ستون‌های متنی
 اگر مقدار تمیز خالی یا مفقود بود مقدار خام به شرط معنادار بودن جایگزین می‌شود

merge_code

ادغام مقدار تمیز شده با مقدار خام برای ستون‌های کدی
 در صورت مفقود بودن مقدار تمیز تلاش برای تبدیل امن مقدار خام به عدد صحیح انجام می‌شود

truthy

تعبیر رشته‌های باینری رایج مانند یک و true و yes و به درست و صفر و no و رشته خالی به نادرست
 سایر موارد مفقود
 این تابع برای فلگ‌ها در پر کردن خلاً استفاده می‌شود

align

هم‌ترازی یک بردار بر اساس ترتیب نمونه‌ها
 این کار تضمین می‌کند تمام بردارهای ویژگی دقیقاً در همان ترتیب ستون‌های بیان قرار گیرند

fill_from_flags

پر کردن مقادیر مفقود در یک ستون هدف با استفاده از فلگ مثبت و فلگ منفی
 کاربرد برای Sex از tissue_normal و tissue_tumor و برای sex_female و sex_male از tissue

which_flag

استنتاج مرحله کلی از چهار فلگ | Stage I\IV Stage I\II\III
 تنها زمانی مقدار تعیین می‌شود که دقیقاً یکی از فلگ‌ها درست باشد

منطق انتخاب نمونه‌ها

اگر وجود داشته باشد Full_matrix.csv
 سرستون‌ها خوانده می‌شود
 ستون اول به عنوان ردیف ژن‌ها کنار گذاشته می‌شود
 نام نمونه‌ها همان سرستون‌های بعدی خواهد بود
 اگر Full_matrix موجود نباشد
 نام نمونه‌ها از هدر Micro.xlsx و RNA_seq.xlsx برداشته می‌شود
 ترکیب یکتا ساخته می‌شود

ابتدا نمونه‌های موجود در RNA و سپس نمونه‌های Micro RNA که در RNA نیستند
اگر هیچ نمونه‌ای یافت نشود اجرا متوقف می‌شود

استخراج ویژگی‌ها از Demo

به دیتافیریم تبدیل می‌شود Demo.xlsx

ستون اول نام ویژگی‌ها است

تابع grab_row با یک نام اصلی و چند نام جایگزین یک ردیف را پیدا می‌کند و به صورت برداری از مقادیر در ستون‌های نمونه بازمی‌گرداند

ویژگی‌های اصلی

Tissue

Sex

Stage

T

N

M

LocationCode

LaurenCode

Age_Code

Age

فلگ‌های اختیاری برای پر کردن خلا

sex_male

sex_female

stage_i

stage_ii

stage_iii

stage_iv

tissue_tumor

tissue_normal

پس از دریافت هر بردار تابع align اعمال می‌شود تا ترتیب نمونه‌ها یکسان بماند

پاکسازی اولیه

تولید نسخه‌های تمیز برای سه ستون کلیدی

Tissue_c

Sex_c

Stage_c

پر کردن خلا با فلگ‌ها

اگر Sex_c مفقود باشد و فلگ‌ها تضاد نداشته باشند مقدار Female یا Male از فلگ‌ها تعیین می‌شود

اگر Tissue_c مفقود باشد و فلگ‌ها تضاد نداشته باشند مقدار Normal یا Tumor از فلگ‌ها تعیین می‌شود

استنتاج مرحله از فلگ و از TNM

اگر Stage_c مفقود باشد و دقیقاً یکی از فلگ‌های مرحله فعال باشد همان سطح به Stage_c اختصاص می‌باید
اگر Stage_c همچنان مفقود باشد و مولفه M برابر M1 باشد مرحله IV تعیین می‌شود
این قانون بالینی متقن است و از استنتاج‌های پرخطر دیگر اجتناب می‌شود

قاعده نمونه‌های نرمال

برای نمونه‌های بافت Normal ستون Stage عمدتاً خالی می‌ماند
هدف جلوگیری از نسبت دادن مرحله به نمونه‌های غیرتوموری است

نگاشت کدها

از متن به کد عددی یک تا سه LocationCode_c
از متن به کد عددی یک تا سه LaurenCode_c
با اولویت بردار Age_Code خام و در نبود آن از Age_Code استخراج می‌شود

ادغام نهایی مقادیر

ستون‌های متنه Stage و Sex و Tissue merge_chr با
ستون‌های کدی LocationCode و LaurenCode merge_code با
منطق ادغام این است که مقدار استاندارد شده در اولویت است و اگر خالی باشد مقدار خام که قابل تبدیل است جایگزین شود

تشخیص Platform

نام ستون‌های Micro.xlsx و RNA_seq.xlsx بدون بارگذاری داده خوانده می‌شود
اگر نمونه در RNA دیده شود مقدار Platform RNAseq برابر
اگر نمونه در Micro دیده شود مقدار Platform Microarray برابر
اگر در هیچ‌کدام نبود خانه خالی می‌ماند

ساخت جدول خروجی و ذخیره

ساخت دیتافریم با ستون‌های دقیق و ترتیب ثابت
نوشتن به CombinedCovariates.csv با گزینه na برابر رشته خالی
چاپ پیام تایید با تعداد ردیف و ستون و مسیر ذخیره

کنترل‌های خطأ و توقف امن

وجود stopifnot با Demo.xlsx کنترل می‌شود
اگر Full_matrix وجود داشته باشد اما ساختار سرستون ناسازگار باشد اجرا متوقف می‌شود
اگر هیچ نمونه‌ای از مجموع فایل‌ها استخراج نشود اجرا متوقف می‌شود
در صورت نیافتتن یک سطر در Demo برداری از مقادیر مفقود تولید می‌شود تا در مراحل بعد مدیریت شود

کیفیت داده و مدیریت تعارض

از TNM تنها در صورت M1 به IV ارتقا می‌باید
فلگ‌های مرحله تنها زمانی پذیرفته می‌شود که تنها یک فلگ روشن باشد
در پر کردن Tissue و Sex نضاد فلگ‌ها موجب عدم جایگذاری می‌شود تا از خطأ جلوگیری گردد

بر اساس انطباق دقیق نام نمونه‌ها تعیین می‌شود Platform

در صورت ناهمنامی بین فایل‌ها خانه خالی ثبت می‌شود

سنجه‌های پیشنهادی برای وارسی خروجی

یکتایی ستون Sample

تعداد خانه‌های خالی در ستون‌های Stage و Tissue

سازگاری مسیری که نشان می‌دهد با حضور نمونه در فایل‌های RNA یا Micro

بازبینی دستی چند نمونه برای LaurenCode و LocationCode و Age_Code

محدودیت‌ها و فرض‌ها

نها به چهار سطح کلی نگاشت می‌شود و حروف زیر مرحله کنار گذاشته می‌شود Stage
از مولفه‌های T و N برای استنتاج مرحله کلی استفاده نمی‌شود تا از خطای بالینی جلوگیری شود
به سه سطح گسسته نگاشت می‌شود که برای تحلیل‌های رسته‌ای مناسب است Age_Code

02_Pvalue.R

هدف

ارزیابی آماری ارتباط میان هم متغیرهای مطالعه و تولید پی ولیو برای تمام جفت های ممکن با رویکرد مقاوم در برابر مقادیر مفقود و سطوح نامتوازن

نتایج به صورت فایل های سی اس وی برای استفاده مستقیم در کنترل کیفیت طراحی مدل و گزارش تحلیلی ذخیره می شوند

ورودی ها

مسیر پایه درایو E پوشش GC زیر خط 29 زیر خط 4 mainDATAFRAME

به عنوان جدول هم متغیرها CombinedCovariates.csv

وجود ستون Sample در فایل هم متغیرها الزامی است

رشته خالی و رشته NA به عنوان مقدار مفقود تفسیر می شوند

خروجی ها

پوشش Results_Covariate_Pvals در مسیر پایه

شامل شمارش سطوح هر متغیر همراه با سطح Missing level_counts.csv

شامل نتایج آزمون برای تمام نمونه ها pairwise_pvalues_All.csv

در صورت وجود ستون **Tissue** و کفايت حجم نمونه

برای زیرمجموعه تومور بدون متغیر **Tissue** pairwise_pvalues_TumorOnly.csv

برای زیرمجموعه نرمال بدون متغیر **Tissue** pairwise_pvalues_NormalOnly.csv

فهرست توصیفی فایل های خروجی MANIFEST.csv

وابستگی ها

دیپلایر و رید آر برای پردازش و نوشتمن جدول ها

استرینگ آر برای کار با رشته ها

آماده سازی و استانداردسازی اولیه

ستون Stage به چهار سطح استاندارد I و II و III و IV نگاشت می شود

حروف زیر مرحله مانند A و B و C حذف می شوند

هر نگاشت غیر از چهار سطح استاندارد به مقدار مفقود تبدیل می شود

ستون های Age_Code و LocationCode و LaurenCode در صورت وجود به عدد صحیح مفقود تبدیل می شوند با سرکوب اخطار تبدیل

فهرست متغیرهای هدف شامل **Tissue** و **Sex** و **Stage** و **Age_Code** و **LocationCode** است

صرفهایی که در فایل ورودی حاضر باشند برای آزمون استفاده می شوند

چاپ خلاصه ورودی

تعداد نمونه ها و نام ستون های قابل آزمون چاپ می شود

برای هر متغیر توزیع سطوح محاسبه و تابیه سطح پر تکرار نمایش داده می شود

سطح Missing به صورت سطح صریح گزارش می شود

این خلاصه به فایل level_counts.csv نیز ذخیره می شود

موتور انتخاب آزمون و منطق مقاوم

تمام آزمون ها با نگه داشتن سطح Missing به عنوان یک سطح مجزا انجام می شود تا حذف سطر به دلیل نبود داده رخ ندهد
برای هر جفت متغیر یک جدول توافقی ساخته می شود
سطرهای و ستون هایی که جمع آن ها صفر است حذف می شوند

انتخاب آزمون

اگر جدول دو در دو باشد آزمون دقیق فیشر اجرا می شود
در جدول های بزرگ تر آزمون کای دو اجرا می شود
اگر شمارش های مورد انتظار در هر سلول کمتر از پنج باشد یا گزینه شبیه سازی فعال باشد از شبیه سازی مونت کارلو با نه هزار و نهصد
و نود و نه تکرار برای محاسبه پی ولیو استفاده می شود
در صورت بروز خطای دو برچسب ChiSqFailed و مقادیر تهی بازگردانده می شود

آزمون روند یکنواخت برای متغیرهای ترتیبی

برای جفت هایی که هر دو متغیر ماهیت ترتیبی دارند یک پی ولیوی مکمل با اسپیرمن گزارش می شود
متغیرهای ترتیبی عبارتند از Age_Code LaurenCode و LocationCode و Stage و

Stage به نگاشت عددی یک تا چهار تبدیل می شود
برای سایر کدها تبدیل مستقیم به عدد انجام می شود
اگر شمار مشاهدات غیرمفقود کمتر از سه باشد پی ولیوی روند گزارش نمی شود

تابع run_pair_tests و جریان اجرا

انتخاب متغیرها

ابتدا فقط متغیرهایی که در داده حاضر هستند و حداقل دو سطح واقعی دارند نگه داشته می شوند
سطح Missing در این بررسی به عنوان یک سطح مستقل شمرده می شود تا متغیر تک سطح تلقی نشود تنها اگر تمام مقادیر واقعاً یک
مقدار و بدون Missing باشند متغیر ثابت محسوب می شود
تولید همه جفت ها

از ترکیب دو تایی متغیرهای منتخب تمام جفت ها ساخته می شود
اجرای آزمون

برای هر جفت ابتدا آزمون استقلال با منطق فیشر یا کای دو انجام می شود سپس پی ولیوی روند در صورت ترتیبی بودن هر دو متغیر
محاسبه می شود

خروجی هر ردیف شامل Var1 و Var2 و DF و Statistic و Method و PValue و N و Trend_P است
تصحیح چندگانه

ستون adj.P.Val با روش بنجامینی هوچبرگ محاسبه می شود
مرتب سازی

نتایج بر اساس adj.P.Val و سپس PValue به صورت صعودی مرتب می شوند
نوشتمن فایل

نتایج هر برچسب در فایلی با الگوی pairwise_pvalues برچسب نقطه CSV ذخیره می شود
برچسب All برای تمام نمونه ها
برچسب TumorOnly برای زیرمجموعه تومور
برچسب NormalOnly برای زیرمجموعه نرمال

لایه های تحلیل

تمام نمونه ها

اگر تعداد ستون های قابل آزمون کمتر از دو باشد تحلیل متوقف می شود

زیرمجموعه های بافت

اگر ستون **Tissue** حاضر باشد

نمونه های بافت **Normal** و **Tumor** جدا می شوند

برای هر زیرمجموعه در صورتی که حداقل سه نمونه وجود داشته باشد آزمون های جفتی روی همان متغیرها به جز **Tissue** اجرا می شود

اگر شمار نمونه ها کمتر از سه باشد پیام مناسب چاپ و آن لایه رد می شود

تولید مانیفست خروجی

شامل نام فایل ها و شرح هر فایل در پوشه خروجی ذخیره می شود تا رديابی نتایج ساده شود

نکات کیفیت داده و توصیه های تفسیر

وجود سطح **Missing** به عنوان سطح مجزا باعث می شود اثر مقادیر مفقود در آزمون لحظه شود و از حذف لیستی جلوگیری گردد با این

حال اگر سهم **Missing** بسیار زیاد باشد نتایج باید با احتیاط تفسیر شوند

برای جداول بزرگ با فراوانی های مورد انتظار پایین استفاده از شبیه سازی مونت کارلو موجب پایداری پی و لیو می شود اما زمان محاسبه را افزایش می دهد

پی و لیوی روند اسپرمن تنها برای جفت های ترتیبی معنا دارد و مکمل آزمون استقلال است عدم معنی داری در یکی از دو سنجه لزومن دیگری را نقض نمی کند

در خروجی ستون **Method** نوع آزمون را مشخص می کند از جمله **FisherExact** دو در دو و **ChiSqSim** به همراه ابعاد جدول

ستون **N** اندازه نمونه موثر پس از حذف سطر و ستون های دارای جمع صفر را نشان می دهد

ستون **Val** **dot** **adj** معیار اصلی برای کنترل خطای نوع یک در چندگانه سنجی است و برای رتبه بندی روابط پیشنهاد می شود

کنترل خطأ و توقف امن

نبود فایل **CombinedCovariates.csv** باعث توقف با پیام شفاف می شود

نبود ستون **Sample** باعث توقف می شود

اگر تعداد متغیرهای قابل آزمون کمتر از دو باشد اجرای آزمون ها متوقف می شود و خروجی تهی برگردانده می شود

پوشش کاربردی

این اسکریپت برای کشف وابستگی های ساختاری میان هم متغیرهای کلینیکی و نمونه ای طراحی شده است

خروچی های آن می توانند به عنوان ورودی معیار برای کنترل کیفیت سازگاری داده ها انتخاب کواریتی ها در مدل های رگرسیونی و

تنظیم لایه بندی های بعدی به کار روند

03_geo_combat_umap_O.R

هدف

اصلاح بچ اثر در زیرمجموعه میکرواری با استفاده از روش کامبت و ترسیم نگاشت امبدینگ یکنواخت برای آشکارسازی ساختار پنهان
پیش و پس از اصلاح

همه نمونه های بافت با مقدار نامشخص یا سایر حذف می شوند تا اثر بافت تنها شامل نرم ال و تومور باشد ➤

ورودی ها

فایل Full_matrix.csv ماتریس بیان ژن با ژن در سطر و نمونه در ستون

فایل CombinedCovariates.csv جدول هم متغیرها شامل ستون Sample و Platform و Tissue

فایل Demo.xlsx فراداده تکمیلی برای استخراج سری یا موسسه و کشور

وجود هر سه فایل الزامی است و در صورت نبود هر کدام اجرا متوقف می شود

خروجی ها

فایل GEO batchXtissue table.csv جدول توافقی دسته و بافت پس از همه پالایش ها

فایل GEO ComBat matrix.csv ماتریس بیان زیرمجموعه میکرواری پس از کامبت

فایل GEO UMAP 4panel.png تصویر چهار پنلی مقایسه پیش و پس از اصلاح بر حسب دسته و بر حسب بافت

پیام ثبت شده در خروجی استاندارد برای مسیر فایل ها و وضعیت اجرا

بسته های مورد استفاده

دیپلایبر برای داده کاوی

اس وی ای برای اجرای کامبت

ماتریکس استنس برای واریانس سطري

یومپ برای کاهش بعد

جی جی پلات دو برای ترسیم

رید ایکسل برای خواندن فراداده

پیچ ورک برای چیدمان پنل ها

رید آر برای نوشتن جدول ها

انتخاب زیرمجموعه میکرواری

ستون Platform از CombinedCovariates به کار می رود

نمونه هایی که مقدار آن برابر Microarray است نگه داشته می شوند

از تکیه بر ستون Study صرف نظر شده که پیش تر منجر به خطأ می شد

اگر هیچ نمونه میکرواری یافت نشود اجرا با پیام خوانا متوقف می شود

بارگذاری و هم راستاسازی

ماتریس بیان از Full_matrix.csv خوانده می شود

جدول هم متغیر از CombinedCovariates.csv خوانده می شود

نام ستون های ماتریس بیان باید دقیقا با ستون Sample در جدول هم متغیر برابر باشد

در غیر این صورت اجرا متوقف می شود

زیرمجموعه `geo` و `expr geo cov` بر اساس میکرواری ساخته می شود
حداقل سه ستون لازم است تا یومپ اجرا شود

استخراج ویژگی های دسته بندی از **Demo**

توابع کمکی برای نرمال سازی کلیدها حروف کوچک و حذف فاصله های اضافی
یافتن ردیف های مرتبط با سازمان و کشور و سری از دمو با مجموعه ای از کلیدواژه های متراffد
اگر ردیفی یافت نشود یک `\N` با مقادیر تهی ساخته می شود
`\N` های سازمان و کشور و سری سپس بر اساس شناسه نمونه های میکرواری هم راستا می شوند

ساخت دسته یا بتجه برای کامبیت

اولویت با سری جی اس ای است

اگر سری معتبر و دارای حداقل دو مقدار یکتا باشد از آن به عنوان دسته استفاده می شود
در غیر این صورت اگر حداقل یکی از سازمان یا کشور در دسترس باشد از برهم کنش سازمان ضربدر کشور به عنوان دسته بهره گرفته می شود

در غیر این صورت اگر پلتفرم بیش از یک سطح داشته باشد همان پلتفرم به عنوان دسته استفاده می شود
در نهایت اگر هیچ کدام مهیا نبود یک دسته واحد با برچسب جی ای او ساخته می شود
این راهبرد تطبیقی سبب بهبود مدل نویزی دسته ها در داده های ناهمگون می شود

پالایش های پایه پیش از کامبیت

حذف ژن های با واریانس صفر با استفاده از واریانس سط्रی بزرگ تر از صفر
حذف نمونه های بافت نامشخص شامل `unknown` و `other` و `na` و رشته خالی
نگه داری تنها دو کلاس نرمال و تومور برای بافت
بازنیت عامل بافت به دو سطح نرمال و تومور
پس از پالایش بافت اگر شمار نمونه ها کمتر از سه شود یومپ قابل اجرا نیست و اجرا متوقف می شود

حذف دسته های بسیار کوچک پس از پالایش بافت

جدول فراوانی دسته ها محاسبه می شود
دسته هایی که اندازه آنها کوچکتر از دو است حذف می شوند
این کار پایداری برآورد پارامترهای کامبیت را افزایش می دهد
پس از این حذف نیز شرط حداقل سه نمونه دوباره بررسی می شود

گزارش دهی پیش از کامبیت

جدول توافقی دسته ضربدر بافت پس از تمام پالایش ها ساخته و در فایل `GEO batchXtissue table.csv` ذخیره می شود
این جدول به کنترل کیفیت چیدمان نمونه ها در دسته ها کمک می کند

اجرای کامبیت با طراحی تطبیقی

اگر تعداد سطوح دسته بیشتر از یک باشد کامبیت اجرا می شود
مدل طراحی برای تعديل اثر بافت فقط زمانی شامل بافت است که هر دو کلاس نرمال و تومور حضور داشته باشند
اگر تنها یک کلاس بافت موجود باشد مدل تنها شامل عرض از مبدأ است
کامبیت با برآورد میانگین و واریانس اجرا می شود و پیش نویس های پیشین غیرفعال هستند
در صورت وجود تنها یک دسته پیام مناسب چاپ و ماتریس بدون تغییر عبور داده می شود

پیکربندی و اجرای یومپ

بذر تصادفی روی عدد چهل و دو ثابت می شود تا تکرار پذیری تضمین شود
همسایه ها تعداد حداقل پانزده یا یک کمتر از شمار نمونه هر کدام کوچکتر
فاصله مین دیست برابر دو دهم

یومپ پیش از کامبت روی ماتریس بیان بدون اصلاح اجرا می شود
برای یومپ پس از کامبت ابتدا سلکشن ژنی بر اساس بیشترین واریانس انجام می شود
تعداد ژن های منتخب حداقل دو و حداقل سه هزار با توجه به تعداد ژن های با واریانس مثبت
یومپ پس از کامبت روی ترانهاده ماتریس منتخب اجرا می شود

این دو اجرا امکان مقایسه مستقیم ساختار دسته و بافت پیش و پس از اصلاح را فراهم می کند

آماده سازی داده برای ترسیم چهار پنل

چهار چارچوب داده ساخته می شود
پیش از اصلاح رنگ بر اساس دسته
پس از اصلاح رنگ بر اساس دسته
پیش از اصلاح رنگ بر اساس بافت
پس از اصلاح رنگ بر اساس بافت

برای دسته ها یک پالت رنگی ثابت تعریف می شود با طول کافی برای پوشش همه سطوح
برای بافت رنگ سبز برای نرمال و قرمز برای تومور تعیین می شود

این همسان سازی رنگ تفسیر بصری را ساده می کند

توابع ترسیم و چیدمان

تابع plot panel یک نمودار پراکنش با نقطه های نیمه شفاف می سازد
در صورت عاملی بودن رنگ از مقیاس رنگ دستی استفاده می شود و در غیر این صورت از گرادیان پیوسته
قالب بنده تم به صورت تم بی دابلیو با شبکه روشن و عنوان بولد در مرکز است
چهار پنل با چیدمان دو در دو توسط پچ ورک کنار هم قرار می گیرند
عنوان کلی تصویر زیرمجموعه میکروواری یومپ پیش و پس است
زیرنویس بسته به اجرای کامبت متن مناسب را نشان می دهد
تصویر در مسیر پایه با نام GEO UMAP 4panel.png ووضوح سیصد نقطه در اینچ و اندازه دوازده در ده اینچ ذخیره می شود

همزمان تصویر در خروجی گرافیکی نیز چاپ می شود تا بررسی سریع ممکن شود

ذخیره سازی نتایج نهایی

ماتریس بیان پس از کامبت برای زیرمجموعه میکروواری در فایل GEO ComBat matrix.csv نوشته می شود
این ماتریس فقط شامل نمونه هایی است که در تمام مراحل پالایش باقی مانده اند
پیامی حاوی نام سه فایل خروجی در پایان چاپ می شود تا ردگیری نتایج آسان گردد

نکات کیفیت و ملاحظات تحلیلی

حذف نمونه های بافت نامشخص سبب می شود سیگنال بافتی به شکل تمیزتری در کامبت مدل شود و از القای بایاس نامعلوم جلوگیری شود

حذف دسته های تک نمونه ای پس از پالایش بافت باعث می شود پارامترهای کامبت برای هر دسته از داده کافی بهره ببرند و از بیش برآذش جلوگیری شود

استفاده از متغیر بافت در ماتریس طراحی تنها هنگامی که هر دو کلاس حضور دارند از خطای هم خطی و ناپایداری جلوگیری می کند

انتخاب ژن های با واریانس بالا برای یومپ پس از اصلاح باعث تمرکز بعد کاهی بر ژن های اطلاع بخش می شود و نویز را می کاهد هم‌سویی دقیق نام ستون های ماتریس بیان با شناسه های نمونه در کواریتیت ها یک الزام سختگیرانه است و هر عدم تطبیق بلا فاصله باعث توقف می شود

04_normalize_and_merge.R

هدف

• اصلاح اثر بج برای داده های میکروواری (GEO/GSM-like) با روش ComBat فقط روی بج هایی که بیش از یک نمونه دارند، سپس ترسیم UMAP/PCA در صورت کفاوت نمونه ها. نمونه های تک تایی (singleton) بدون تغییر نگه داشته می شوند.

ورودی ها

• ماتریس بیان (ردیف = ژن، ستون = نمونه).

• کواریتیت ها شامل ستون های Sample و Platform و Tissue.

• فراداده ای تکمیلی برای استخراج Series/GSE و Organization و Country.

خروجی ها

• ماتریس بیان برای زیرمجموعه Microarray از اصلاح.

• جدول توافقی (Batch×Tissue با حفظ سطح). GEO_batchXtissue_table.csv —

• چهار پنل UMAP پیش/پس بر حسب Batch و Tissue در صورت ≤ 3 نمونه.

• پیام های وضعیت در خروجی استاندارد.

بسته ها /وابستگی ها

readr , patchwork , readxl , viridis , ggplot2 , umap , matrixStats , sva , • dplyr

منطق و مراحل

☒ انتخاب میکروواری: keep ← (Platform == 'Microarray');

☒ استخراج دسته: اولویت با Series/GSE در Demo.xlsx ; در غیر این صورت سپس Platform و Organization×Country در نهایت یک دسته ی تک.

☒ پالایش پایه: حذف ژن های با واریانس صفر؛ ساخت عامل بافت سسطوحی (Normal/Tumor/Unknown) بدون حذف Unknown.

☒ تفکیک دسته ها: تقسیم به بج های چند نمونه ای vs تک نمونه ای؛ اجرای ComBat فقط روی چند نمونه ای ها.

☒ طراحی ComBat: اگر در چند نمونه ای ها هر دو کلاس بافت (Unknown) بدون حضور داشت، tissue را در مدل می گنجاند؛ و گرنه مدل فقط عرض از مبدأ.

ترکیب مجدد: اتصال ستون‌های اصلاح شده ی چندنمونه‌ای با تکنمونه‌ای های دست‌نخورده و بازگردانی ترتیب اصلی ستون‌ها.

MAP U محافظه کار: فقط اگر $n \geq 3$ ؛ پیش از اصلاح روی کل ژن‌ها؛ پس از اصلاح روی «تا ۳۰۰۰ ژن با بیشترین واریانس.»

Tissue (Normal/ Tumor/ Unknown). ترسیم ۴ پنل: رنگ‌های Batch و رنگ‌های ثابت برای viridis .

ذخیره: نوشتمن ماتریس اصلاح شده و جدول BatchxTissue و تصویر UMAP.

نکات و ملاحظات

ها اصلاح نمی‌شوند تا از برآوردهای ناپایدار پارامترهای ComBat جلوگیری شود.

حضور سطح Unknown در Tissue حفظ می‌شود تا پوشش کامل داده‌ها از دست نرود.

05_all_combat.R

هدف

• اصلاح اثر بج در تمام نمونه‌ها با راهبرد تفکیک چندنمونه‌ای/تکنمونه‌ای و stratify-by-tissue برای اجرای ComBat در هر لایه بافت (Normal/Tumor/Unknown)، سپس مرور UMAP کلی و تولید خروجی‌های جامع .

ورودی‌ها

• ماتریس بیان اولیه. Full_matrix.csv —

• کواریتی‌ها (Sample/Platform/...) با افزودن ستون‌های مکمل از Demo.xlsx در صورت فقدان.

• استخراج Series/Organization/Country و نیز Sex/Stage/Age/Site از Demo.xlsx —

خروجی‌ها

• ماتریس بیان پس از اصلاح سراسری. All_CoMBat_matrix.csv —

• کواریتی‌ها با ستون‌های تکمیل شده/نرمال شده. All_Covariates_augmented.csv —

• قبیل/بعد بر حسب Study ALL_UMAP_4panel.png — UMAP

• پیام‌های وضعیت و ساخت دایرکتوری. Results_all

بسته‌ها / وابستگی‌ها

readr, patchwork, viridis, ggplot2, umap, matrixStats, sva, readxl, dplyr

منطق و مراحل

• غنی‌سازی کواریتی‌ها از Demo.xlsx (Sex/Stage/Age/Site) و هم‌استاسازی با نمونه‌ها، نرمال‌سازی Stage I-IV به Platform استخراج از Study در صورت نبود.

• ساخت batch_all: برای میکرواری، ترجیح ORGCTY: SERIES: یا Study. برای بقیه.

 **تفکیک به ComBat**: اجرای stratify-by-tissue ، multi vs singleton: فقط روی multi و ساخت مدل انعطاف‌پذیر (Sex/Stage/Age) با حذف تدریجی در صورت هم خطی/ناسازگاری.

 **بازترکیب**: اتصال ستون‌های اصلاح‌شده ی multi با singleton‌های دستنخورده و حفظ ترتیب.

 **UMAP کلی**: چهار پنل {Before/After} × {Study,Tissue} با ۳۰۰۰ ژن پروواریانس.

 **ذخیره**: ماتریس نهایی + کوارییت‌های بهروزشده.

نکات و ملاحظات

 **طراحی مدل ComBat**: به صورت پویا ستون‌های مشکل‌زا را حذف می‌کند تا اجرا شکست نخورد.

 **بافت در مدل ComBat**: مستقیماً استفاده نمی‌شود چون stratify-by-tissue انجام شده است.

qc_outliers_scaling.R_06

هدف

• کنترل کیفیت پس از ComBat (و در صورت موجود بودن، قبل از آن) با ترسیم چگالی/باکس، نگاشتهای PCA و UMAP و شناسایی آوت‌لایرها با RPCA و قوانین IQR.

ورودی‌ها

• الزامی All_CoBat_matrix.csv —

• الزامی All_Covariates_augmented.csv —

• اختیاری برای مقایسه pre vs post Full_matrix.csv —

خروجی‌ها

QC_box_post.png — .QC_box_pre.png ، • QC_density_pre_post.png نمودارهای چگالی/باکس.

• PCA_pre_post.png .UMAP_pre_post.png یا نسخه‌های «post» تنها.

• جدول پرچم‌گذاری آوت‌لایرها. Outliers_report.csv —

removed_samples.txt — . retained_samples.txt فهرست نمونه‌های نگهداشته/حذف شده.

• پیام نهایی با شمار نمونه‌های نگهداری شده.

بسته‌ها / وابستگی‌ها

patchwork , rrcov , umap , ggplot2 , matrixStats , readr ، • dplyr

منطق و مراحل

.safe_log1p برای ایمن‌سازی لگاریتم در صورت مقادیر ≥ -1 .

نمودارها: چگالی/باکس برای pre/post ; PCA/UMAP رنگشده برحسب Tissue.

تشخیص آوتلایر \leftrightarrow RPCA (PcaHubert) : فاصله ی مقاوم + قوانین IQR روی اندازه ی کتابخانه و میانه ی سیگنال .

گزارش : ادغام پرچم‌های RPCA و IQR در Any_Outlier ، و نوشتن خروجی‌ها .

نکات و ملاحظات

اگر Full_matrix.csv در دسترس نباشد، مقایسه ی pre/post محدود به post خواهد بود .

آستانه‌های IQR با ضریب ۳ به صورت پیش‌فرض تنظیم شده‌اند .

edge.R_07

هدف

• تحلیل دیفرانسیل بیان (DGE) با limma برای مقایسه ی Tumor vs Normal به صورت سراسری و همچنین درون‌گروهی با طراحی مقاوم (per-demographic) .

ورودی‌ها

• ماتریس بیان پس از اصلاح All_CoMBat_matrix.csv —

• کواریتیت‌های نهایی All_Covariates_augmented.csv —

خروجی‌ها

• نتایج سراسری DGE_global_Tumor_vs_Normal.csv —

• ولکانوی سراسری Volcano_global_Tumor_vs_Normal.png —

• هیتمپ ۵۰ زن برتر بر اساس $|t|$ در صورت کفایت Heatmap_global_Top50.png —

• نتایج درون‌گروهی برای per_demo/.csv — Sex/Stage/Studу/Platform.

بسته‌ها / وابستگی‌ها

circlize ,matrixStats ,readr ,dplyr ,ComplexHeatmap ,EnhancedVolcano , ggplot2 ,• limma

منطق و مراحل

ساخت طراحی Tissue2 (Normal/Tumor/Unknown) + کواریتیت‌های آگاه به اطلاع (Study/Platform/Sex/Stage/AgeZ).

برآش eBayes → topTable . با ماتریس کنترast ImFit → contrasts.fit

ترسیم Volcano : با آستانه ی $p=0.05$ و $\log2(1.5) \approx FC$ در صورت وجود .

تحلیل درون‌گروهی: تکرار همان چارچوب در هر سطح متغیر (با شرط وجود هر دو کلاس بافت و $n \geq 6$) .

نکات و ملاحظات

در Tissue نگه داشته می شود ولی کنتراست صرفاً بین Normal و Tumor تعریف شده است.

طراحی فقط کواریتیت های اطلاع بخش را اضافه می کند تا از بیش برآذش جلوگیری شود.

gsva.R_08

هدف

• امتیازدهی مسیرهای زیستی با GSVA (روی بیان ژنی) پس از ComBat با نگاشت شناسه ها به SYMBOL و سپس انجام limma بر امتیازها برای کنتراست (Tumor vs Normal) سراسری و درون گروهی . (همچنین تهیه ی خروجی های آماده برای شبکه .

ورودی ها

• ماتریس بیان اصلاح شده . All_CoMBat_matrix.csv —

• کواریتیت ها . All_Covariates_augmented.csv —

خروجی ها

• ماتریس امتیاز برای مجموعه های GSVA_scores_.csv —

• نتایج limma GSVA_DGE__Tumor_vs_Normal.csv — روی امتیازها .

Heatmap_GSVA__Top30.png — . • Volcano_GSVA__Tumor_vs_Normal.png نمودارها .

• تجمعیت نتایج مجموعه ها . GSVA_TvN_allCollections_results.csv —

• فهرست امتیازها (در صورت وجود) . GSVA_scores_list.rds —

• خروجی های آماده (Network prep در انتهای اسکریپت) .

بسته ها /وابستگی ها

.readr .matrixStats .dplyr .circlize .ComplexHeatmap .ggplot2 .limma .msigdbr .• GSVA EnhancedVolcano .org.Hs.eg.db .AnnotationDbi

منطق و مراحل

تشخیص نوع شناسه (SYMBOL/ENSEMBL/ENTREZ) ID و نگاشت به SYMBOL ؛ میانگین گیری روی تکراری ها .

واکشی مجموعه ژنی با msigdbr با چندین مسیر جایگزین برای سازگاری؛ فیلتر مجموعه ها با اندازه ≥ 5 ژن .

اجرای (GSVA fallback) یا z-mean مبتنی بر GSVAParam) در صورت عدم دسترسی به .

روی امتیازها: طراحی مقاوم مشابه DGE ژنی؛ کنتراست Tumor-Normal ؛ ذخیره و رسم . limma

تحلیل درون گروهی Sex/Stage/Stdy/Platform: با شروط حضور هر دو کلاس بافت و $n \geq 6$.

خروجی‌های با واریانس بالا برای SYMBOL/all/Tumor/Normal + فایل‌های کواریتی Network prep: پاکسازی شده.

نکات و ملاحظات

در نبود مجموعه‌ی معتبر، اسکریپت با پیام خطأ متوقف می‌شود تا کاربر مجموعه‌ها را اصلاح کند. ✓

نگاشت ENSEMBL با حذف پسوند نسخه انجام می‌شود. ✓

wgcna_modules.R_09

هدف

• تشکیل شبکه‌ی همبیانی وزندار (WGCNA) روی ژن‌های پُرواریانس، استخراج مدول‌ها، ایگن‌مدول‌ها، و ارتباط آن‌ها با صفات (Tissue/Sex/Stage/...) به علاوه‌ی تحلیل بقا در صورت موجود بودن ✎.

ورودی‌ها

• ماتریس بیان All_CoBat_matrix.csv —

• کواریتی‌ها All_Covariates_augmented.csv —

خروجی‌ها

• تخصیص ژن → مدول WGCNA_ModuleColors.csv —

• ایگن‌مدول‌ها WGCNA_ModuleEigengenes.csv —

• دندروگرام + رنگ مدول‌ها WGCNA_Dendro.png —

• Module_Trait_heatmap.png — . • Module_Trait_correlations.csv — همبستگی مدول-صفت.

• اتصال درون‌مدولی IntramodularConnectivity.csv —

• در صورت وجود داده‌ی بقا) Module_Cox_univariate.csv — (

بس腾ه‌ها / وابستگی‌ها

• WGCNA pheatmap , survival , patchwork , ggplot2 , readr , dplyr

منطق و مراحل

• انتخاب ژن‌ها goodSamplesGenes: بر اساس واریانس؛ غربال topN=10000 ✎

• انتخاب توان pickSoftThreshold: با fallback به 6 در نبود تخمین معتبر؛ ذخیره‌ی نمودارهای SFT

• ساخت شبکه networkType='signed': با blockwiseModules 0.25 حداقل اندازه‌ی مدول 30، ادغام در ارتفاع

• ارتباط با صفات: محاسبه‌ی همبستگی ایگن‌مدول‌ها با Traits عددی/باينري؛ p-value دانشجویی ✎

تحليل بقا: در صورت وجود ستون های زمان/رخداد، محاسبه HR تک متغیره برای هر مدول. ❤️

نکات و ملاحظات

به صورت عددی سازی شده (Stage → 1..4) جهت همبستگی استفاده می شوند. ✓ Traits

برای پایداری، نوع شبکه و TOM به صورت 'signed' تنظیم شده است. ✓

network_variants.R_10

هدف

- ساخت چند نوع شبکه ی ژنی بر اساس معیارهای اطلاعاتی/همبستگی (ARACNE/CLR/MRNET/PCIT) در دو گروه Tumor و Normal و گزارش همپوشانی یال‌ها.

ورودی‌ها

• ماتریس بیان اصلاح شده. All_Combat_matrix.csv —

• کواریتی‌ها. All_Covariates_augmented.csv —

خروجی‌ها

• یال‌های شبکه برای هر گروه. pcit_.csv — , mrnet_.csv , clr_.csv . • aracne_.csv

• جدول همپوشانی (Jaccard/Inter/Union) بین روش‌ها. edge_overlap_jaccard.csv —

• پیام نهایی «Network variants built.»

بسته‌ها / وابستگی‌ها

pcit , pROC , pheatmap , purrr , matrixStats , readr , dplyr , igraph , infotheo . • minet

منطق و مراحل

انتخاب ژن topN=2000: بر اساس واریانس. ⚡

گروه‌بندی Tumor و Normal بر اساس tissue ؛ نیاز به $n \geq 10$ در هر گروه. 👤

شبکه‌های minet: گرسیسته‌سازی با ARACNE/CLR/MRNET. infotheo (nbins=3) → MI → MI

PCIT: همبستگی اسپیرمن و آزمون استقلال جزئی برای تصمیم‌گیری یال‌ها.

نوشتن یال‌ها: خروجی هر روش/گروه؛ سپس محاسبه Jaccard و آمار همپوشانی. ⚪

نکات و ملاحظات

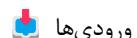
وزن یال در minet بر اساس MI و در PCIT بر اساس $|\rho|$ است. ✓

در نبود نمونه ی کافی در هر گروه، آن گروه رد می‌شود. ✓

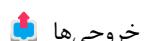
11_networks.R

هدف 

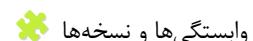
بازسازی شبکه‌های هم‌کنش ژنی مبتنی بر اطلاعات متقابل (Mutual Information) با ARACNE به صورت لایه‌ای در زیرگروه‌های جمعیت‌شناختی/پاتولوژیک، به همراه با قیمانده‌گیری (residualization) مشروط بر کوواریتی‌های بقامحور معنی‌دار (Cox) برای کنترل مخدوش‌گرها.

ورودی‌ها 

ماتریس بیان پس از بکارچه‌سازی/حذف اثر سری‌ها (All_CoBat_matrix.csv) : ژن‌های نمونه‌ها
کوواریتی‌ها (All_Covariates_augmented.csv) : ستون Sample همتراز با ستون‌های ماتریس بیان

خروجی‌ها 

نت‌لیست هر گروه : Network_<Group>_edges.csv (gene1,gene2,weight)
درجه ی رأس‌ها : Network_<Group>_node_degrees.csv
پیش‌نمایش گراف (Gephi) برای مرورگرهای گراف / Network_<Group>_graph.png + GraphML
برای Cytoscape)

وابستگی‌ها و نسخه‌ها 

ARACNE و minet/infotheo برای تخمین MI
limma::removeBatchEffect برای با قیمانده‌گیری چندمتغیره
readr/dplyr برای IO/Dاده
igraph برای گراف،
منطق و خطوط کلی الگوریتم

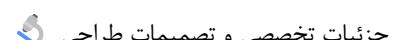
پالایش پیوستگی : انتخاب topN_genes با بیشترین واریانس (rowVars) جهت کاهش نویز و پیچیدگی محاسباتی.
تعریف زیرگروه‌ها (Sex/Stage/LocationCode/LaurenCode/AgeCode/Tissue) و فیلتر نمونه‌ها با حداقل n (min_samples=10).

غربال کوواریتی‌های بقا (Surv) : تک متغیره بر روی dd = coxph(Surv ~ Sex/Stage/LocationCode/LaurenCode/Age) sig_covars. FDR<0.05 ⇒

برای هر گروه، با قیمانده‌گیری بیان ژن نسبت به \{ گروه \} با limma::removeBatchEffect sig_covars طرح طراحی ماتریسی با کدگذاری موهومی.

گسسته‌سازی نمونه‌محور → (nbins=3) ماتریس MI (build.mim, mi.empirical) برای حذف وابستگی‌های غیرمستقیم.

استخراج یال‌ها + MI(upper-triangle) وزن CSV/PNG/GraphML = ، محاسبه درجه، خروجی.

جزئیات تخصصی و تصمیمات طراحی 

اطلاعات متقابل نسبت به همبستگی خطی، وابستگی‌های غیرخطی را نیز پوشش می‌دهد؛ ARACNE با اصل حذف مسیرهای ترو-مدیاتور، یال‌های کاذب را می‌زداید.

قبل از MI از تورش ناشی از سن/مرحله/جنس و ... می‌کاهد؛ در این اسکریپت باقیمانده‌گیری سمت زن‌ها انجام می‌شود (ستون محور).

Discretization با nbins=3 تعادلی میان سوگیری و واریانس برقرار می‌کند؛ برای داده‌های بسیار بزرگ می‌توان nbins را افزایش داد.

حداقل اندازه‌ی گروه (min_samples=10) برای پایداری MI لازم است؛ در غیر این صورت برآوردهای MI مستعد نوسان‌اند. مقیاس‌بندی/نرمال‌سازی پیش‌نیاز: خروجی پس از ComBat، به صورت مقیاس یکسان، و حذف batch‌اثرات ماکروسکوپیک. پارامترها و تنظیمات

2000: topN_genes بهتر است با توجه به n نمونه و بودجه‌ی محاسباتی تنظیم شود.

3: mi_bins برای سطوح نویز متغیر، 3-5 آزمایش شود.

10: min_samples برای گروه‌های کوچک‌تر، شبکه ساخته نمی‌شود.

200: preview_edges جهت رسم سریع.

توابع/بلوک‌های کلیدی

build_cov_mm: ساخت ماتریس طراحی از کوواریتی‌های معنی‌دار بقایی (کدگذاری عامل‌ها).
infer_group_var: جلوگیری از over-adjustment با حذف کوواریت خود گروه.
build_mi_net: گسسته‌سازی، ARACNE، MI، استخراج یال‌ها و تولید پیش‌نمايش.

کیفیت‌سنجی و تفسیر

توزیع درجات (degree) برای شناسایی هاب‌ها؛ heavy-tail توزیع نشانگر ساختار مقیاس‌آزاد.
بررسی پایداری با bootstrapping (پیشنهاد توسعه) یا تغییر nbins/topN_genes و ارزیابی همپوشانی یال‌ها.

12_0network.R

هدف

ساخت نسخه‌ی ساده و سریع از شبکه‌های MI/ARACNE بدون باقیمانده‌گیری و بدون غربالگری بقا برای ایجاد baseline و امكان مقایسه با پیاده‌سازی پیشرفته‌ی 11_networks.R.

ورودی‌ها/خروجی‌ها

Expr: All_Combat_matrix.csv
Cov: All_Covariates_augmented.csv → بازکدگذاری سریع
LocationCode/LaurenCode/StageClean/AgeGroup
: Network_<Group>_edges.csv + Network_<Group>_graph.png + خروجی‌ها Covariates_recoded_for_networks.csv

روش

بازکدگذاری ساده محل‌الارن و گروه سنی به کدهای regex با ۳..۱ (⚠️ حساس به نویسه‌گذاری).
انتخاب ۲۰۰۰ ژن با واریانس بالا، گسسته‌سازی nbins=3، محاسبه MI و ARACNE.
تشکیل گروهها (Tissue/Lauren/Age/Clean/Stage/Sex) با آستانه $n \geq 10$ و ساخت شبکه.

کاربرد و محدودیت‌ها

💡 مناسب برای پیش‌نمایش سریع و sanity-check.
⚠️ عدم کنترل مخدوش‌گرها \Rightarrow برای تحلیل نهایی به networks.R ۱۱ تکیه کنید.

13_diffcoexp.R

🎯 هدف

تحلیل همیانی تفاضلی (DGCA) میان وضعیت‌های Tumor و Normal با هدف استخراج زوج‌زن‌هایی که همبستگی آن‌ها بین دو وضعیت تغییر معنی‌دار دارد و تشکیل ماژول‌های DiffCoExp به صورت مولفه‌های همبستگی.

↗️ روش‌شناسی

کاهش بعد: انتخاب ۱۲۰۰ ژن پر تغییر (rowVars).
تعريف کلاس‌ها classes $\in \{\text{Tumor}, \text{Normal}\}$: و فیلتر نمونه‌های Unknown.
BH \Rightarrow q.adjusted. Spearman و اصلاح چندگانه. dddcorAll با همبستگی DCM استخراج زوج‌های معنادار ($q < 0.05$) و ساخت گراف بدون جهت؛ مولفه‌های همبند \Rightarrow برچسب.

📤 خروجی‌ها

DGCA_all_pairs.csv تمام زوج‌ها
DGCA_sig_pairs_q05.csv زوج‌های معنادار
DiffCoEx_modules.csv انجمان/ماژول‌های تفاضلی

🌐 نکات تفسیری

زوج‌هایی که علامت/قدر همبستگی‌شان بین حالت‌ها واگرایت می‌تواند نشانه‌ی rewiring در توموروژن باشد.
ماژول‌های DCM کاندید مسیرهای مختلط شده‌اند؛ ترکیب با GSVA/DGE توصیه می‌شود.

14_integration_network_dge_gsVA.R

🎯 هدف

یکپارچه‌سازی شواهد: (۱) هاب/درایورهای شبکه (ARACNE) و (۲) همبستگی ماژول‌های WGCNA با نمرات مسیر GSVA برای ارائه‌ی تصویر چند-لایه از مکانیزم‌ها.

⭐️ گام‌ها

محاسبه مرکزیت‌ها (degree/betweenness/closeness) از گراف Hub-Driver: صد ک = هاب‌ها؛ Tumor_ARACNE.

درجه. همپوشانی با \Rightarrow DGE_global_Tumor_vs_Normal (adj.FDR<0.05)

HubDrivers_Tumor_ARACNE.csv.

R = cor(ME×GSVA: GSVA (H/C2/C5). و امتیاز‌های Module Eigengenes (WGCNA) محاسبه ی، ME×GSVA: مبارگذاری pathway) و تولید ماتریس‌های همبستگی و هیت‌مپ مسیرهای شاخص.

خروجی‌ها

Tumor_ARACNE_centrality.csv, Tumor_ARACNE_hubs.csv, HubDrivers_Tumor_ARACNE.csv
ME_GSVA_correlations.csv + ME_GSVA_heatmap_top30.png

نکات کلیدی

هاب‌های همزمان DE، کاندید درایورهای شبکه‌ای با پشتیبانی بیان تفاضلی هستند.
ارتباط سطح ژن-ماژول را به سطح مسیر ارتقا می‌دهد و برای تفسیر بیولوژیک حیاتی است.

15_survival.R

هدف

تحلیل بقا: استخراج خودکار ستون‌های زمان/وضعیت از Kaplan-Meier گروهی (دموگرافیک) و مدل Cox چندمتغیره ی تنظیم شده؛ اختیاری: بقا بر مبنای نمرات مسیر GSVA.

ورودی‌ها/خروجی‌ها

All_Covariates_augmented.csv + Demo.xlsx
OS_time/OS_event برای نگاشت خودکار .Cox_adjusted_forest.png .Cox_adjusted_coefficients.csv .KM_Shامل_Results_Survival نتایج: پوشش GSVA) KM_pathway_score.png صورت وجود

روش

نمودار کلیدها (lower/trim) برای یافتن سطرهای زمان/وضعیت در Demo.xlsx و نگاشت به نمونه‌ها. KM: مقایسه ی بقای دسته‌ها (Sex/Stage) با آزمون log-rank (pval).

Cox تنظیم شده با CI95% HR و خروجی Sex+Stage+Age+Study+Platform+Tissue؛ اختیاری → GSVA: میانگین چند مسیر (EMT/Angiogenesis) hallmarks → KM. میانه و

ملاحظات

واحد زمان (روز/ماه) در داده‌ها ممکن است متفاوت باشد؛ نمودها/مدل‌ها نسبت به واحد حساس‌اند.
کیفیت نگاشت Demo.xlsx به شدت به یکسانی شناسه ی نمونه‌ها وابسته است.

16_survival_advanced.R

هدف

مدل سازی پیش‌بینی بقا با Cox LASSO روی ویژگی‌های سطح-خلاصه Module Eigengenes و اختیاری هاب‌زن‌ها (، تولید نمره ROC زمان‌مند، و لایه‌بندی KM بر اساس ترتیلهای ریسک).

خطوط اصلی

فیلتر نمونه‌ها به موارد با OS_time/OS_event معتبر؛ هم‌راستاسازی expr↔cov ویژگی‌ها WGCNA MES (اجباری) + اختیاری: ژن‌های هاب (degree) ۹۵٪ با نام‌گذاری HUB_<Gene>. + Risk score = $X\beta$. cv.lambda.minCox LASSO (glmnet) cv. خروجی: ضرایب غیرصفر t ∈ {12,36,60} (بر اساس سه کاته ROC timeROC در) می‌باشد. KM به مقیاس سه کاته ریسک.

نکات و ترفندها

در کد، یک اصلاح انجام شده: نام‌گذاری ستون‌های هاب ابتدا اشتباه بود و سپس با HUB_<Gene> اصلاح شده است. Feature leakage را با استفاده از تنها MEs یا با cross-study validation کنترل کنید. برای تعادل بایاس/واریانس، در n کوچک، ElasticNet با α=1 مناسب‌تر از LASSO پیچیده است.

خروجی‌ها

KM_by_risk_tertiles.png .timeROC_AUC.csv .Risk_scores_train.csv .CoxLASSO_selected_features.csv

17_external_validation.R

هدف

اعتبارسنجی خارجی مدل بقای آموزش‌دیده: بازتخمین Eigengene WGCNA و محاسبه ریسک با ضرایب CoxLASSO؛ گزارش AUC زمان‌مند و نمرات ریسک در گُهورت مستقل.

جريان کار

بارگذاری رنگ‌های ماژولی (WGCNA_ModuleColors.csv) از مجموعه آموزش و محاسبه MEs در اعتبارسنجی با همان رنگ‌ها (moduleEigengenes).

بارگذاری ضرایب انتخابی CoxLASSO از آموزش و تشکیل ماتریس ویژگی X فقط برای فیچرهای موجود. Risk_scores_VALID.csv + timeROC_AUC_VALID.csv. Rیسک = $X\beta$ محاسبه ROC زمان‌مند در 60/36/12 و ذخیره.

پیش‌نیازها/فرض‌ها

تعریف مسیرهای پوششی اعتبارسنجی از طریق GC_VALIDATION_DIR. همسانی شناسه‌های نمونه و ژن بین آموزش و اعتبارسنجی intersect (برای ژن‌ها انجام می‌شود.)

محدودیت و توسعه

برای فیچرهای از نوع HUB، بازتولید ویژگی در گُهورت اعتبارسنجی نیازمند نگاشت ژن‌محور است (در این کد برای سادگی فقط MEs اعتبارسنجی می‌شوند).

توصیه calibration plot: و تصمیم‌منحنی (net benefit) برای تکمیل ارزیابی.

