**00\_Covfilecreator.R**  
ساخت جدول هم‌متغیر استاندارد (**CombinedCovariates.csv**) با پاکسازی/نرمال‌سازی مقادیر، تشخیص **Platform** و ثبت خانه ی خالی به‌جای 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 کوارییت‌ها (Fisher/Chi-square با مونت‌کارلو + Spearman برای متغیرهای ترتیبی)، همراه با **level\_counts** و خروجی‌های CSV و **MANIFEST**. 📊

**03\_geo\_combat\_umap\_O.R**  
ComBat برای زیرمجموعه ی **Microarray** با حذف نمونه‌های بافت نامشخص (فقط Normal/Tumor) و ترسیم **UMAP** قبل/بعد؛ ذخیره ی **GEO\_ComBat\_matrix.csv** و جدول **Batch×Tissue**. 🗺️

**04\_normalize\_and\_merge.R**  
ComBat فقط روی بچ‌های چندنمونه‌ای (singleton‌ها دست‌نخورده)، حفظ سطح **Unknown** در بافت، و UMAP/PCA در صورت کفایت نمونه؛ خروجی: **GEO\_ComBat\_matrix.csv** و **GEO\_UMAP\_4panel.png**. 🔧

**05\_all\_combat.R**  
ComBat سراسری با تفکیک **multi vs singleton** و **stratify-by-tissue**؛ مدل انعطاف‌پذیر (Sex/Stage/Age) با حذف تدریجی در تعارض؛ خروجی: **All\_ComBat\_matrix.csv** و **ALL\_UMAP\_4panel.png**. 🌐

**09\_qc\_outliers\_scaling.R**  
QC پیش/پس: نمودار چگالی و باکس، **PCA/UMAP** رنگ‌شده با Tissue، و شناسایی آوت‌لایرها با **RPCA + IQR**؛ گزارش **Outliers\_report.csv** و فهرست نمونه‌های retained/removed. 🧪

**04\_edge.R**  
تحلیل **DGE** با limma برای **Tumor vs Normal** (سراسری و درون‌گروهی)، تولید **Volcano** و **Heatmap** و ذخیره ی نتایج CSV. 🔬

**06\_gsva.R**  
امتیازدهی مسیرها با **GSVA** (H, KEGG, REACTOME, GOBP)، انجام limma روی امتیازها (Tumor−Normal)، ولکانو/هیت‌مپ و آماده‌سازی خروجی‌های شبکه و تلفیق. 🧬

**12\_wgcna\_modules.R**  
ساخت شبکه ی **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 ی همپوشانی یال‌ها (**Jaccard**)؛ خروجی یال‌ها به‌ازای هر روش/گروه. 🔁

**11\_networks.R**  
بازسازی شبکه‌های **ARACNE** به‌تفکیک زیرگروه‌های دموگرافیک/پاتولوژیک با **residualization** مبتنی‌بر کووارییت‌های بقاییِ معنی‌دار؛ خروجی یال‌ها، درجات، و گراف پیش‌نمایش. 🧠

**12\_0network.R**  
نسخه ی ساده و سریع شبکه‌سازی **MI/ARACNE** بدون residualization برای baseline و sanity-check؛ بازکدگذاری سبک کوارییت‌ها و گراف‌های پیش‌نمایش. ⚡

**13\_diffcoexp.R**  
**DGCA** برای شناسایی هم‌بیانی تفاضلی بین **Tumor** و **Normal** (Spearman + FDR)، استخراج زوج‌های معنادار و ماژول‌های **DiffCoExp**. 🔄

**14\_integration\_network\_dge\_gsva.R**  
یکپارچه‌سازی شواهد: **هاب‌های شبکه (ARACNE)** × **ژن‌های DE** و همبستگی **ME×GSVA**؛ خروجی Hub-Drivers و هیت‌مپ مسیرهای شاخص. 🔗

**15\_survival.R**  
تحلیل بقا کلاسیک: **Kaplan–Meier** برحسب 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 ی **Risk score**، **timeROC** و **KM** بر اساس ترتایل ریسک. 📈

**17\_external\_validation.R**  
اعتبارسنجی خارجی: بازتخمین **MEs** با رنگ‌های آموزش، اعمال ضرایب **CoxLASSO**، و گزارش **AUC زمان‌مند** و نمرات ریسک در کُهورت مستقل. ✅

**00\_Covfilecreator.R**  
این اسکریپت یک جدول هم‌متغیر استاندارد می‌سازد تا برای تمام نمونه‌ها در کل پروژه قابل استفاده باشد. خروجی دقیقا 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  
Demo.xlsx شامل سطرهای توصیفی ویژگی‌ها و ستون‌های نمونه‌ها  
RNA\_seq.xlsx در صورت وجود برای استخراج فهرست نمونه‌های رنا  
Micro.xlsx در صورت وجود برای استخراج فهرست نمونه‌های میکرواری  
Full\_matrix.csv در صورت وجود برای تعیین ترتیب نمونه‌ها بر اساس سرستون‌های ماتریس بیان  
وجود Demo.xlsx اجباری است و در صورت نبودن اسکریپت متوقف می‌شود

**وابستگی کتابخانه‌ها**

readxl برای خواندن اکسل  
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  
خواندن جدول Demo و یافتن سطرهای مرتبط با هر ویژگی  
پاکسازی و نرمال‌سازی مقادیر رسته‌ای و عددی  
پر کردن خلأها 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 دیتافریم 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 اسامی سطر در Demo

**find\_row**

گرفتن بردار نام ویژگی‌ها 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 نرمال مانند adjacent normal یا non tumor به Normal و هر متن حاوی سرطان به Tumor. در غیر این صورت مقدار مفقود

**clean\_sex**

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

**clean\_stage\_overall**

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

**clean\_T و clean\_N و clean\_M**

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

**clean\_location\_code**

نگاشت متن محل تومور به کد سه مقداری  
یک برای ناحیه پروگزیمال 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  
یک برای diffuse  
دو برای intestinal  
سه برای mixed یا indeterminate  
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 شصت  
اگر متن عددی باشد تبدیل می‌شود و اگر کد نوشتاری مانند u40 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**

تعبیر 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 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 از sex\_male و sex\_female  
پر کردن Tissue از tissue\_tumor و tissue\_normal

**which\_flag**

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

**منطق انتخاب نمونه‌ها**

اگر Full\_matrix.csv موجود باشد سرستون‌های آن به عنوان ترتیب نهایی نمونه‌ها استفاده می‌شود و ستون اول را به عنوان ژن کنار می‌گذارد  
اگر موجود نباشد سرستون‌های RNA\_seq.xlsx و Micro.xlsx خوانده می‌شود و مجموعه یکتا ساخته می‌شود. ترتیب برابر است با نمونه‌های رنا به همان ترتیب و سپس نمونه‌های میکرواری که در رنا نبوده‌اند  
اگر هیچ نمونه‌ای یافت نشود اجرای اسکریپت متوقف می‌شود

**خواندن Demo و بازیابی سطرهای ویژگی**

Demo.xlsx به دیتافریم 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  
T و N و M  
LocationCode  
LaurenCode  
Age\_Code  
Age  
فلگ‌های اختیاری برای پر کردن خلأ  
sex\_male  
sex\_female  
stage\_i تا stage\_iv  
tissue\_tumor  
tissue\_normal  
در پایان هر بردار با align به ترتیب نهایی نمونه‌ها هم‌تراز می‌شود

**پاکسازی و نرمال‌سازی اولیه**

برای Tissue و Sex و Stage نسخه تمیز شده تولید می‌شود  
Tissue\_c  
Sex\_c  
Stage\_c

**پر کردن خلأ با فلگ‌ها**

اگر Sex\_c برای یک نمونه مفقود باشد و sex\_male درست و sex\_female نادرست باشد مقدار Male گذاشته می‌شود و برعکس برای Female  
اگر Tissue\_c مفقود باشد از فلگ‌های تومور و نرمال استفاده می‌شود

**استنتاج Stage 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 Stage\_c تخصیص می‌یابد

**استنتاج ایمن Stage از TNM**

اگر Stage\_c همچنان مفقود باشد و مقدار M برابر M1 باشد مرحله IV تعیین می‌شود  
این قاعده بالینی سازگار است زیرا متاستاز دوردست به معنی مرحله چهار است

**قاعده ویژه برای نمونه‌های نرمال**

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

**کدگذاری محل تومور و رده لورن و سن**

LocationCode\_c با تابع تمیزکننده به عدد یک تا سه نگاشت می‌شود  
LaurenCode\_c به عدد یک تا سه  
Age\_Code با اولویت بردار Age\_Code خام و در صورت نبودن از Age استخراج می‌شود

**ادغام مقادیر تمیز با خام**

برای هر ستون  
Tissue و Sex و Stage با merge\_chr  
LocationCode و LaurenCode و Age\_Code با merge\_code  
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 RNA\_seq.xlsx و Micro.xlsx بدون داده خوانده می‌شود  
اگر نام نمونه در RNA دیده شود Platform برابر RNAseq  
اگر در Micro دیده شود Platform برابر Microarray  
اگر در هیچ کدام دیده نشود مقدار خالی ثبت می‌شود

**ساخت دیتافریم نهایی**

ستون‌ها دقیقا 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  
CombinedCovariates.csv در پوشه پایه  
گزینه 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**

Sample نام یکتا برای هر نمونه  
Platform یکی از دو مقدار RNAseq یا Microarray یا خالی  
Tissue تنها دو مقدار Normal یا Tumor یا خالی  
Sex تنها دو مقدار Male یا Female یا خالی  
Stage فقط یکی از I یا II یا III یا IV برای نمونه‌های توموری. برای نمونه‌های نرمال خالی  
LocationCode کد سه مقداری ناحیه در معده. یک پروگزیمال. دو تنه. سه دیستال  
LaurenCode کد سه مقداری رده لورن. یک منتشر. دو روده‌ای. سه مختلط  
Age\_Code 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 خطا 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 سطری یافت نشود برداری از مقادیر مفقود برگردانده می‌شود که در مراحل بعدی مدیریت می‌شود

**نکات کیفیت داده و تعارض‌ها**

Stage 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 دقیقا 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  
Platform 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 Tissue و Stage  
همسویی کامل نام نمونه‌ها با ماتریس بیان و با فایل‌های RNA و Micro  
بازبینی چند مورد نمونه به صورت دستی برای صحت نگاشت‌های Location و Lauren 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 عمدتا 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  
Tissue برابر Tumor یا Normal  
Sex برابر Male یا Female  
Stage برای Tumor یک مقدار مانند II و برای Normal خالی  
LocationCode یک یا دو یا سه  
LaurenCode یک یا دو یا سه  
Age\_Code یک یا دو یا سه  
در خروجی CombinedCovariates.csv دقیقا همین ستون‌ها با همین مقادیر و خانه‌های خالی به جای مقادیر مفقود دیده می‌شود

**01\_Covfilecreator.R**  
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

**ورودی‌ها**

مسیر پایه درایو E پوشه GC زیرخط 29 زیرخط 4 زیرخط mainDATAFRAME  
Demo.xlsx شامل ردیف ویژگی‌ها و ستون نمونه‌ها  
RNA\_seq.xlsx در صورت وجود برای فهرست نمونه‌های رنا  
Micro.xlsx در صورت وجود برای فهرست نمونه‌های میکرواری  
Full\_matrix.csv در صورت وجود برای تحمیل ترتیب نمونه‌ها بر اساس سرستون‌های ماتریس بیان  
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**

CombinedCovariates.csv با ستون‌های دقیق که در بخش هدف آمده است  
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 RNA و Micro  
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 خلأ با استفاده از فلگ‌های یک به یک در صورت وجود  
استنتاج مرحله کلی از فلگ‌ها و در صورت نیاز از مولفه M در TNM  
قواعد دامنه شامل خالی کردن Stage برای نمونه‌های Normal  
تشخیص Platform برای هر نمونه بر اساس حضور در فایل‌های RNA یا Micro  
ساخت جدول خروجی و نوشتن 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  
مصرف برای تطبیق نام ردیف‌ها در Demo

**find\_row**

دریافت 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 adjacent normal و non tumor به Normal  
عبارات حاوی carcinoma یا cancer یا tumor به Tumor  
غیر از این مقادیر مفقود

**clean\_sex**

تبدیل m و male به Male و f و female 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\_T و clean\_N و clean\_M**

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 تنه 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 متداول مانند u40 یا 40-60 یا 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 به درست و صفر و false و no و رشته خالی به نادرست  
سایر موارد مفقود  
این تابع برای فلگ‌ها در پر کردن خلأ استفاده می‌شود

**align**

هم‌ترازی یک بردار بر اساس ترتیب نمونه‌ها  
این کار تضمین می‌کند تمام بردارهای ویژگی دقیقا 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 sex\_male و sex\_female و برای Tissue از tissue\_tumor و tissue\_normal

**which\_flag**

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

**منطق انتخاب نمونه‌ها**

اگر Full\_matrix.csv وجود داشته باشد  
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 RNA\_seq.xlsx و Micro.xlsx برداشته می‌شود  
ترکیب یکتا ساخته می‌شود  
ابتدا نمونه‌های موجود در RNA و سپس نمونه‌های Micro که در RNA نیستند  
اگر هیچ نمونه‌ای یافت نشود اجرا متوقف می‌شود

**استخراج ویژگی‌ها از Demo**

Demo.xlsx به دیتافریم تبدیل می‌شود  
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 خلأ  
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 خلأ با فلگ‌ها**

اگر Sex\_c مفقود باشد و فلگ‌ها تضاد نداشته باشند مقدار Male یا Female از فلگ‌ها تعیین می‌شود  
اگر Tissue\_c مفقود باشد و فلگ‌ها تضاد نداشته باشند مقدار Tumor یا Normal از فلگ‌ها تعیین می‌شود

**استنتاج مرحله از فلگ و از TNM**

اگر Stage\_c مفقود باشد و دقیقا یکی از فلگ‌های مرحله فعال باشد همان سطح به Stage\_c اختصاص می‌یابد  
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 عمدا خالی می‌ماند  
هدف جلوگیری از نسبت دادن مرحله به نمونه‌های غیرتوموری است

**نگاشت کدها**

LocationCode\_c از متن به کد عددی یک تا سه  
LaurenCode\_c از متن به کد عددی یک تا سه  
Age\_Code با اولویت بردار Age\_Code خام و در نبود آن از Age استخراج می‌شود

**ادغام نهایی مقادیر**

ستون‌های متنی Tissue و Sex و Stage با merge\_chr  
ستون‌های کدی LocationCode و LaurenCode و Age\_Code با merge\_code  
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 RNA\_seq.xlsx و Micro.xlsx بدون بارگذاری داده خوانده می‌شود  
اگر نمونه در RNA دیده شود مقدار Platform برابر RNAseq  
اگر نمونه در Micro دیده شود مقدار Platform برابر Microarray  
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 خطا و توقف امن**

وجود Demo.xlsx با stopifnot 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 نیافتن یک سطر در Demo برداری از مقادیر مفقود تولید می‌شود تا در مراحل بعد مدیریت شود

**کیفیت داده و مدیریت تعارض**

Stage از TNM تنها در صورت M1 به IV ارتقا می‌یابد  
فلگ‌های مرحله تنها زمانی پذیرفته می‌شود که تنها یک فلگ روشن باشد  
در پر کردن Sex و Tissue تضاد فلگ‌ها موجب عدم جایگذاری می‌شود تا از خطا جلوگیری گردد  
Platform بر اساس انطباق دقیق نام نمونه‌ها تعیین می‌شود  
در صورت ناهمنامی بین فایل‌ها خانه خالی ثبت می‌شود

**سنجه‌های پیشنهادی برای وارسی خروجی**

یکتایی ستون Sample  
تعداد خانه‌های خالی در ستون‌های Tissue و Stage  
سازگاری مسیری که Platform نشان می‌دهد با حضور نمونه در فایل‌های RNA یا Micro  
بازبینی دستی چند نمونه برای LocationCode و LaurenCode و Age\_Code

**محدودیت‌ها و فرض‌ها**

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

**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 پوشه GC زیرخط 29 زیرخط 4 زیرخط mainDATAFRAME  
CombinedCovariates.csv به عنوان جدول هم متغیرها  
وجود ستون Sample در فایل هم متغیرها الزامی است  
رشته خالی و رشته NA به عنوان مقدار مفقود تفسیر می شوند

خروجی ها  
پوشه Results\_Covariate\_Pvals در مسیر پایه  
level\_counts.csv شامل شمارش سطوح هر متغیر همراه با سطح Missing  
pairwise\_pvalues\_All.csv شامل نتایج آزمون برای تمام نمونه ها  
در صورت وجود ستون Tissue و کفایت حجم نمونه  
pairwise\_pvalues\_TumorOnly.csv برای زیرمجموعه تومور بدون متغیر Tissue  
pairwise\_pvalues\_NormalOnly.csv برای زیرمجموعه نرمال بدون متغیر Tissue  
MANIFEST.csv فهرست توصیفی فایل های خروجی

وابستگی ها  
دیپلایر و رید آر برای پردازش و نوشتن 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 زیرمرحله مانند A و B و C حذف می شوند  
هر نگاشت غیر از چهار سطح استاندارد به مقدار مفقود تبدیل می شود  
ستون های LocationCode و LaurenCode و Age\_Code در صورت وجود به عدد صحیح تبدیل می شوند با سرکوب اخطار تبدیل  
فهرست متغیرهای هدف شامل Tissue و Sex و Stage و LocationCode و LaurenCode و Age\_Code است  
صرفا متغیرهایی 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 تا بیست سطح پرتکرار 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  
سطرها 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 خطا در کای 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 Stage و LocationCode و LaurenCode و Age\_Code  
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 واقعا 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 Method و Statistic و DF و P.Value 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 P.Value به صورت صعودی مرتب می شوند  
نوشتن 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 Tumor و Normal 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  
MANIFEST.csv شامل نام فایل ها و شرح هر فایل در پوشه خروجی ذخیره می شود تا ردیابی 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 لحاظ 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 ولیو می شود اما زمان محاسبه را افزایش می دهد  
پی ولیوی روند اسپیرمن تنها برای جفت های ترتیبی معنا دارد و مکمل آزمون استقلال است عدم معنی داری در یکی از دو سنجه لزوما دیگری را نقض نمی کند  
در خروجی ستون Method نوع آزمون را مشخص می کند از جمله FisherExact دو در دو و ChiSq 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 adj dot P dot Val معیار اصلی برای کنترل خطای نوع یک در چندگانه سنجی است و برای رتبه بندی روابط پیشنهاد می شود

کنترل خطا 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  
خروجی های آن می توانند به عنوان ورودی معیار برای کنترل کیفیت سازگاری داده ها انتخاب کوارییت 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 نرمال 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**  
دیپلایر برای داده کاوی  
اس وی ای برای اجرای کامبت  
ماتریکس استتس برای واریانس سطری  
یومپ برای کاهش بعد  
جی جی پلات دو برای ترسیم  
رید ایکسل برای خواندن فراداده  
پچ 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 خطا می شد  
اگر هیچ نمونه میکرواری یافت نشود اجرا با پیام خوانا متوقف می شود

**بارگذاری و همراستاسازی**  
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 expr geo و cov geo 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 مترادف  
اگر ردیفی یافت نشود یک向ه با مقادیر تهی ساخته می شود  
向ه های سازمان و کشور و سری سپس بر اساس شناسه نمونه های میکرواری همراستا می شوند

**ساخت دسته یا بتچ برای کامبت**  
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 کامبت**  
حذف ژن های با واریانس صفر با استفاده از واریانس سطری 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  
بازتعریف 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 کامبت را افزایش می دهد  
پس از این حذف نیز شرط حداقل سه نمونه دوباره بررسی می شود

**گزارش دهی پیش از کامبت**  
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 کامبت با طراحی تطبیقی**  
اگر تعداد سطوح دسته بیشتر از یک باشد کامبت اجرا می شود  
مدل طراحی برای تعدیل 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  
کامبت 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 تکرارپذیری 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 حداکثر 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 بندی 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 میکرواری 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 و وضوح 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 میکرواری 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 برازش 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 بعدکاهی بر ژن های اطلاع بخش می شود و نویز را می کاهد  
همسویی دقیق نام ستون های ماتریس بیان با شناسه های نمونه در کوارییت ها یک الزام سختگیرانه است و هر عدم تطبیق بلافاصله باعث توقف می شود ✅

**04\_normalize\_and\_merge.R**

**هدف**

• اصلاح اثر بچ برای داده‌های میکرواری (GEO/GSM‑like) با روش ComBat فقط روی بچ‌هایی که بیش از یک نمونه دارند، سپس ترسیم UMAP/PCA در صورت کفایت نمونه‌ها. نمونه‌های تک‌تایی (singleton) بدون تغییر نگه داشته می‌شوند. 🎯

**ورودی‌ها**

• Full\_matrix.csv — ماتریس بیان (ردیف=ژن، ستون=نمونه).

• CombinedCovariates.csv — کوارییت‌ها شامل ستون‌های Sample و Platform و Tissue.

• Demo.xlsx — فراداده ی تکمیلی برای استخراج Series/GSE، Organization و Country.

**خروجی‌ها**

• GEO\_ComBat\_matrix.csv — ماتریس بیان برای زیرمجموعه‌ی Microarray پس از اصلاح.

• GEO\_batchXtissue\_table.csv — جدول توافقی Batch×Tissue (با حفظ سطح Unknown).

• GEO\_UMAP\_4panel.png — چهار پنل UMAP پیش/پس برحسب Batch و Tissue (در صورت ≥3 نمونه).

• پیام‌های وضعیت در خروجی استاندارد.

**بسته‌ها / وابستگی‌ها**

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

**منطق و مراحل**

📌 انتخاب میکرواری: keep ← (Platform == 'Microarray'); همراستاسازی expr و cov بر اساس Sample.

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

🧹 پالایش پایه: حذف ژن‌های با واریانس صفر؛ ساخت عامل بافت سه‌سطحی (Normal/Tumor/Unknown) بدون حذف Unknown.

🔀 تفکیک دسته‌ها: تقسیم به بچ‌های چندنمونه‌ای vs تک‌نمونه‌ای؛ اجرای ComBat فقط روی چندنمونه‌ای‌ها.

🧪 طراحی ComBat: اگر در چندنمونه‌ای‌ها هر دو کلاس بافت (بدون Unknown) حضور داشت، tissue را در مدل می‌گنجاند؛ وگرنه مدلِ فقط عرض از مبدأ.

🔗 ترکیب مجدد: اتصال ستون‌های اصلاح‌شده ی چندنمونه‌ای با تک‌نمونه‌ای‌های دست‌نخورده و بازگردانی ترتیب اصلی ستون‌ها.

🗺️ UMAP محافظه‌کار: فقط اگر n≥3؛ پیش از اصلاح روی کل ژن‌ها؛ پس از اصلاح روی «تا ۳۰۰۰ ژن با بیشترین واریانس».

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

💾 ذخیره: نوشتن ماتریس اصلاح‌شده و جدول Batch×Tissue و تصویر UMAP.

**نکات و ملاحظات**

✅ Singletonها 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 بچ در تمام نمونه‌ها با راهبرد تفکیک چندنمونه‌ای/تک‌نمونه‌ای و stratify-by-tissue برای اجرای ComBat در هر لایه ی بافت (Normal/Tumor/Unknown)، سپس مرور UMAP کلی و تولید خروجی‌های جامع. 🌐

**ورودی‌ها**

• Full\_matrix.csv — ماتریس بیان اولیه.

• CombinedCovariates.csv — کوارییت‌ها (Sample/Platform/…) با افزودن ستون‌های مکمل از Demo.xlsx در صورت فقدان.

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

**خروجی‌ها**

• All\_ComBat\_matrix.csv — ماتریس بیان پس از اصلاح سراسری.

• All\_Covariates\_augmented.csv — کوارییت‌ها با ستون‌های تکمیل‌شده/نرمال‌شده.

• ALL\_UMAP\_4panel.png — UMAP قبل/بعد برحسب Study و Tissue.

• پیام‌های وضعیت و ساخت دایرکتوری Results\_all.

**بسته‌ها / وابستگی‌ها**

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

**منطق و مراحل**

🧭 غنی‌سازی کوارییت‌ها از Demo.xlsx (Sex/Stage/Age/Site) و همراستاسازی با نمونه‌ها؛ نرمال‌سازی Stage به I–IV؛ استخراج Study از Platform در صورت نبود.

🏷️ ساخت batch\_all: برای میکرواری، ترجیح SERIES: یا ORGCTY:؛ برای بقیه Study.

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

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

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

💾 ذخیره: ماتریس نهایی + کوارییت‌های به‌روزشده.

**نکات و ملاحظات**

✅ طراحی مدل ComBat به‌صورت پویا ستون‌های مشکل‌زا را حذف می‌کند تا اجرا شکست نخورد.

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

——————————————————————————————

**06\_qc\_outliers\_scaling.R**

**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**

• All\_ComBat\_matrix.csv — الزامی.

• All\_Covariates\_augmented.csv — الزامی.

• Full\_matrix.csv — اختیاری برای مقایسه ی pre vs post.

**خروجی‌ها**

• QC\_density\_pre\_post.png ، QC\_box\_pre.png ، QC\_box\_post.png — نمودارهای چگالی/باکس.

• PCA\_pre\_post.png ، UMAP\_pre\_post.png یا نسخه‌های «post» تنها.

• Outliers\_report.csv — جدول پرچم‌گذاری آوت‌لایرها.

• retained\_samples.txt ، removed\_samples.txt — فهرست نمونه‌های نگه‌داشته/حذف‌شده.

• پیام نهایی با شمار نمونه‌های نگه‌داری‌شده.

**بسته‌ها / وابستگی‌ها**

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

**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: 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 به‌صورت پیش‌فرض تنظیم شده‌اند.

——————————————————————————————

**07\_edge.R**

**هدف**

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

**ورودی‌ها**

• All\_ComBat\_matrix.csv — ماتریس بیان پس از اصلاح.

• All\_Covariates\_augmented.csv — کوارییت‌های نهایی.

**خروجی‌ها**

• DGE\_global\_Tumor\_vs\_Normal.csv — نتایج سراسری.

• Volcano\_global\_Tumor\_vs\_Normal.png — ولکانوی سراسری.

• Heatmap\_global\_Top50.png — هیت‌مپ ۵۰ ژن برتر بر اساس |t| (در صورت کفایت).

• per\_demo/.csv — نتایج درون‌گروهی برای Sex/Stage/Study/Platform.

**بسته‌ها / وابستگی‌ها**

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

**منطق و مراحل**

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

🧪 برازش: lmFit → contrasts.fit با ماتریس کنتراست Tumor−Normal → eBayes → topTable.

🌋 ترسیم: Volcano با آستانه ی p=0.05 و FC≈log2(1.5)؛ هیت‌مپ top‑50 در صورت وجود.

👥 تحلیل درون‌گروهی: تکرار همان چارچوب در هر سطحِ متغیر (با شرط وجود هر دو کلاس بافت و n≥6).

**نکات و ملاحظات**

✅ Unknown در Tissue نگه داشته می‌شود ولی کنتراست صرفاً بین Normal و Tumor تعریف شده است.

✅ طراحی فقط کوارییت‌های اطلاع‌بخش را اضافه می‌کند تا از بیش‌برازش جلوگیری شود.

——————————————————————————————

**08\_gsva.R**

**هدف**

• امتیازدهی مسیرهای زیستی با GSVA روی بیان ژنی (پس از ComBat) با نگاشت شناسه‌ها به SYMBOL و سپس انجام limma بر امتیازها برای کنتراست Tumor vs Normal (سراسری و درون‌گروهی). همچنین تهیه ی خروجی‌های آماده برای شبکه. 🧬

**ورودی‌ها**

• All\_ComBat\_matrix.csv — ماتریس بیان اصلاح‌شده.

• All\_Covariates\_augmented.csv — کوارییت‌ها.

**خروجی‌ها**

• GSVA\_scores\_.csv — ماتریس امتیاز برای مجموعه‌های H, C2:KEGG/REACTOME, C5:GOBP.

• GSVA\_DGE\_\_Tumor\_vs\_Normal.csv — نتایج limma روی امتیازها.

• Volcano\_GSVA\_\_Tumor\_vs\_Normal.png ، Heatmap\_GSVA\_\_Top30.png — نمودارها.

• GSVA\_TvN\_allCollections\_results.csv — تجمیع نتایج مجموعه‌ها.

• GSVA\_scores\_list.rds — فهرست امتیازها (در صورت وجود).

• خروجی‌های آماده ی Network prep (در انتهای اسکریپت).

**بسته‌ها / وابستگی‌ها**

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

**منطق و مراحل**

🆔 تشخیص نوع شناسه (SYMBOL/ENSEMBL/ENTREZ) و نگاشت به SYMBOL؛ میانگین‌گیری روی تکراری‌ها.

📚 واکشی مجموعه ژنی با msigdbr با چندین مسیر جایگزین برای سازگاری؛ فیلتر مجموعه‌ها با اندازه ≥۵ ژن.

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

🧪 limma روی امتیازها: طراحیِ مقاوم مشابه DGE ژنی؛ کنتراست Tumor−Normal؛ ذخیره و رسم.

🧩 تحلیل درون‌گروهی: Sex/Stage/Study/Platform با شروطِ حضور هر دو کلاس بافت و n≥6.

🕸️ Network prep: خروجی‌های با واریانس بالا برای all/Tumor/Normal + نسخه ی SYMBOL و فایل‌های کوارییت پاک‌سازی‌شده.

**نکات و ملاحظات**

✅ در نبود مجموعه ی معتبر، اسکریپت با پیام خطا متوقف می‌شود تا کاربر مجموعه‌ها را اصلاح کند.

✅ نگاشت ENSEMBL با حذف پسوند نسخه انجام می‌شود.

——————————————————————————————

**09\_wgcna\_modules.R**

**هدف**

• تشکیل شبکه ی هم‌بیانی وزندار (WGCNA) روی ژن‌های پُرواریانس، استخراج مدول‌ها، ایگن‌مدول‌ها، و ارتباط آن‌ها با صفات (Tissue/Sex/Stage/…) به‌علاوه ی تحلیل بقا در صورت موجود بودن. 🧩

**ورودی‌ها**

• All\_ComBat\_matrix.csv — ماتریس بیان.

• All\_Covariates\_augmented.csv — کوارییت‌ها.

**خروجی‌ها**

• WGCNA\_ModuleColors.csv — تخصیص ژن→مدول.

• WGCNA\_ModuleEigengenes.csv — ایگن‌مدول‌ها.

• WGCNA\_Dendro.png — دندروگرام + رنگ مدول‌ها.

• Module\_Trait\_correlations.csv ، Module\_Trait\_heatmap.png — همبستگی مدول‑صفت.

• IntramodularConnectivity.csv — اتصال درون‌مدولی.

• Module\_Cox\_univariate.csv — (در صورت وجود داده ی بقا).

**بسته‌ها / وابستگی‌ها**

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

**منطق و مراحل**

🎯 انتخاب ژن‌ها: topN=10000 بر اساس واریانس؛ غربال goodSamplesGenes.

⚙️ انتخاب توان: pickSoftThreshold با fallback به 6 در نبود تخمین معتبر؛ ذخیره ی نمودارهای SFT.

🕸️ ساخت شبکه: blockwiseModules با networkType='signed'، حداقل اندازه ی مدول 30، ادغام در ارتفاع 0.25.

🔗 ارتباط با صفات: محاسبه ی همبستگی ایگن‌مدول‌ها با Traits عددی/باینری؛ p‑value دانشجویی.

❤️ تحلیل بقا: در صورت وجود ستون‌های زمان/رخداد، محاسبه ی HR تک‌متغیره برای هر مدول.

**نکات و ملاحظات**

✅ Traits به‌صورت عددی‌سازی‌شده (Stage→1..4) جهت همبستگی استفاده می‌شوند.

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

——————————————————————————————

**10\_network\_variants.R**

**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**

• All\_ComBat\_matrix.csv — ماتریس بیان اصلاح‌شده.

• All\_Covariates\_augmented.csv — کوارییت‌ها.

**خروجی‌ها**

• aracne\_.csv ، clr\_.csv ، mrnet\_.csv ، pcit\_.csv — یال‌های شبکه برای هر گروه.

• edge\_overlap\_jaccard.csv — جدول همپوشانی (Jaccard/Inter/Union) بین روش‌ها.

• پیام نهایی «Network variants built.»

**بسته‌ها / وابستگی‌ها**

• minet ، infotheo ، igraph ، dplyr ، readr ، matrixStats ، purrr ، pheatmap ، pROC ، pcit

**منطق و مراحل**

🎯 انتخاب ژن: topN=2000 بر اساس واریانس.

👥 گروه‌بندی: Tumor و Normal بر اساس Tissue؛ نیاز به n≥10 در هر گروه.

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

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

🔗 نوشتن یال‌ها: خروجیِ هر روش/گروه؛ سپس محاسبه ی Jaccard و آمار همپوشانی.

**نکات و ملاحظات**

✅ وزن یال در minet بر اساس MI و در PCIT بر اساس |ρ| است.

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

——————————————————————————————

**11\_networks.R**

**هدف 🎯**

بازسازی شبکه‌های هم‌کنش ژنی مبتنی‌بر اطلاعات متقابل (Mutual Information) با ARACNE به‌صورت لایه‌ای در زیرگروه‌های جمعیت‌شناختی/پاتولوژیک، به‌همراه باقیمانده‌گیری (residualization) مشروط بر کووارییت‌های بقا‌محورِ معنی‌دار (Cox) برای کنترل مخدوش‌گرها.

**ورودی‌ها 📥**

ماتریس بیانِ پس از یکپارچه‌سازی/حذف اثر سری‌ها: All\_ComBat\_matrix.csv (ژن‌ها×نمونه‌ها)

کووارییت‌ها: 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 (top-edges): Network\_<Group>\_graph.png + GraphML برای مرورگرهای گراف (Gephi/ Cytoscape)

**وابستگی‌ها و نسخه‌ها 🧩**

minet/infotheo برای تخمین MI و ARACNE

limma::removeBatchEffect برای باقیمانده‌گیری چندمتغیره

igraph برای گراف، readr/dplyr برای IO/داده

**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 کووارییت‌های بقا: dd = coxph(Surv) تک‌متغیره بر روی Sex/Stage/LocationCode/LaurenCode/Age؛ FDR<0.05 ⇒ فهرست sig\_covars.

برای هر گروه، باقیمانده‌گیری بیان ژن نسبت به sig\_covars\{گروه} با limma::removeBatchEffect (طرح طراحی ماتریسی با کدگذاری موهومی).

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

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

**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.

🧪 Residualization قبل از MI از تورش ناشی از سن/مرحله/جنس و ... می‌کاهد؛ در این اسکریپت باقیمانده‌گیری سمت ژن‌ها انجام می‌شود (ستون‌محور).

🧯 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 این‌صورت برآوردهای MI مستعد نوسان‌اند.

🧮 مقیاس‌بندی/نرمال‌سازی پیش‌نیاز: خروجی پس از ComBat، به‌صورت مقیاس یکسان، و حذف batch اثرات ماکروسکوپیک.

**پارامترها و تنظیمات 🔧**

2000**:** **topN\_genes** بهتر است با توجه به n نمونه و بودجه ی 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 کووارییت‌های معنی‌دارِ بقایی (کدگذاری عامل‌ها).

infer\_group\_var: جلوگیری از over-adjustment با حذف کووارییتِ خودِ گروه.

build\_mi\_net: گسسته‌سازی، MI، ARACNE، استخراج یال‌ها و تولید پیش‌نمایش.

**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 نسخه ی ساده و سریع از شبکه‌های MI/ARACNE بدون باقیمانده‌گیری و بدون غربالگری بقا برای ایجاد baseline و امکان مقایسه با پیاده‌سازی پیشرفته ی 11\_networks.R.

**ورودی‌ها/خروجی‌ها 📥📤**

Expr: All\_ComBat\_matrix.csv

Cov: All\_Covariates\_augmented.csv → بازکدگذاری سریع LocationCode/LaurenCode/StageClean/AgeGroup

خروجی‌ها: Network\_<Group>\_edges.csv + Network\_<Group>\_graph.png + Covariates\_recoded\_for\_networks.csv

**روش 🧠**

بازکدگذاری ساده محل/لارن و گروه سنی به کدهای 1..3 با regex های سبک (⚠️ حساس به نویسه‌گذاری).

انتخاب 2000 ژن با واریانس بالا، گسسته‌سازی nbins=3، محاسبه MI و ARACNE.

تشکیل گروه‌ها (Sex/StageClean/Location/Age/Lauren/Tissue) با آستانه n≥10 و ساخت شبکه.

**کاربرد و محدودیت‌ها ⚖️**

💡 مناسب برای پیش‌نمایش سریع و sanity-check.

⚠️ عدم کنترل مخدوش‌گرها ⇒ برای تحلیل نهایی به 11\_networks.R تکیه کنید.

————————————————————————————————————————

**13\_diffcoexp.R**

**هدف 🎯**

تحلیل هم‌بیانی تفاضلی (DGCA) میان وضعیت‌های Tumor و Normal با هدف استخراج زوج‌ژن‌هایی که همبستگی آن‌ها بین دو وضعیت تغییر معنی‌دار دارد و تشکیل ماژول‌های DiffCoExp به‌صورت مولفه‌های همبستگی.

**روش‌شناسی 🔬**

کاهش بعد: انتخاب ≤1200 ژنِ پرتغییر (rowVars).

تعریف کلاس‌ها: classes ∈ {Tumor, Normal} و فیلتر نمونه‌های Unknown.

ddcorAll با همبستگی Spearman و اصلاح چندگانه BH ⇒ q.adjusted.

استخراج زوج‌های معنادار (q<0.05) و ساخت گراف بدون جهت؛ مولفه‌های همبند ⇒ برچسب DCM.

**خروجی‌ها 📤**

DGCA\_all\_pairs.csvتمام زوج‌ها

DGCA\_sig\_pairs\_q05.csv زوج‌های معنادار

DiffCoEx\_modules.csv انجمن/ماژول‌های تفاضلی

**نکات تفسیری 🧭**

زوج‌هایی که علامت/قدر همبستگی‌شان بین حالت‌ها واگراست می‌تواند نشانه ی rewiring در توموروژنز باشد.

ماژول‌های DCM کاندید مسیرهای مختل‌شده‌اند؛ ترکیب با DGE/GSVA توصیه می‌شود.

————————————————————————————————————————

**14\_integration\_network\_dge\_gsva.R**

**هدف 🎯**

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

**گام‌ها 🧭**

Hub-Driver: محاسبه مرکزیت‌ها (degree/betweenness/closeness) از گراف Tumor\_ARACNE؛ هاب‌ها = صدک ۹۵ درجه. همپوشانی با DGE\_global\_Tumor\_vs\_Normal (adj.FDR<0.05) ⇒ HubDrivers\_Tumor\_ARACNE.csv.

ME×GSVA: بارگذاری Module Eigengenes (WGCNA) و امتیازهای GSVA (H/C2/C5). محاسبه ی R = cor(ME, 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.

ME×GSVA 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 Demo.xlsx و اجرای Kaplan–Meier گروهی (دموگرافیک) و مدل Cox چندمتغیره ی تنظیم‌شده؛ اختیاری: بقا بر مبنای نمرات مسیر GSVA.

**ورودی‌ها/خروجی‌ها 📥📤**

All\_Covariates\_augmented.csv + Demo.xlsx برای نگاشت خودکار OS\_time/OS\_event

نتایج: پوشه Results\_Survival شامل KM\_ ، Cox\_adjusted\_coefficients.csv ، Cox\_adjusted\_forest.png ، و (در صورت وجود GSVA) KM\_pathway\_score.png

**روش 🧪**

نرمال‌سازی 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 تنظیم‌شده با Sex+Stage+Age+Study+Platform+Tissue؛ خروجی HR و CI95%.

اختیاری: GSVA → میانگین چند مسیر hallmark (مثلاً EMT/Angiogenesis) → تقسیم میانه و KM.

**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 🎯**

مدل‌سازی پیش‌بینی بقا با Cox LASSO روی ویژگی‌های سطح-خلاصه (Module Eigengenes و اختیاری هاب‌ژن‌ها)، تولید نمره ی ریسک، ارزیابی ROC زمان‌مند، و لایه‌بندی KM بر اساس ترتایل‌های ریسک.

**خطوط اصلی 🧠**

فیلتر نمونه‌ها به موارد با OS\_time/OS\_event معتبر؛ هم‌راستاسازی expr↔cov.

ویژگی‌ها: WGCNA MEs (اجباری) + اختیاری: ژن‌های هاب (۹۵% degree) با نام‌گذاری HUB\_<Gene>.

Cox LASSO (glmnet) با cv.lambda.min؛ خروجی: ضرایب غیرصفر + Risk score = Xβ.

timeROC در t ∈ {12,36,60} (ماه/روز بسته به مقیاس) + KM بر اساس سه‌کاته ی 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، α=1 (LASSO) مناسب‌تر از ElasticNet پیچیده است.

**خروجی‌ها 📤**

CoxLASSO\_selected\_features.csv ، Risk\_scores\_train.csv ، timeROC\_AUC.csv ، KM\_by\_risk\_tertiles.png

————————————————————————————————————————

**17\_external\_validation.R**

**هدف 🎯**

اعتبارسنجی خارجی مدل بقای آموزش‌دیده: بازتخمین Eigengeneها بر مبنای رنگ‌های آموزشی WGCNA و محاسبه ی ریسک با ضرایب CoxLASSO؛ گزارش AUC زمان‌مند و نمرات ریسک در کُهورت مستقل.

**جریان کار 🔄**

بارگذاری رنگ‌های ماژولی (WGCNA\_ModuleColors.csv) از مجموعه ی آموزش و محاسبه ی MEs در اعتبارسنجی با همان رنگ‌ها (moduleEigengenes).

بارگذاری ضرایب انتخابی CoxLASSO از آموزش و تشکیل ماتریس ویژگی X فقط برای فیچرهای موجود.

ریسک = Xβ، محاسبه ROC زمان‌مند در 12/36/60 و ذخیره ی Risk\_scores\_VALID.csv + timeROC\_AUC\_VALID.csv.

**پیش‌نیازها/فرض‌ها ✅**

تعریف مسیرهای پوشه ی 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 کُهورت اعتبارسنجی نیازمند نگاشت ژن‌محور است (در این کد برای سادگی فقط MEs اعتبارسنجی می‌شوند).

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