

بخش اول: چکیده (Abstract)

ترجمه

چکیده: شبکه تحقیقاتی اطلس ژنوم سرطان (TCGA) مجموعه‌ای بزرگ از مشخصات بالینی و مولکولی بیش از ۱۰,۰۰۰ بیمار سرطانی را در ۳۳ نوع تومور مختلف منتشر کرده است TCGA. با استفاده از این مجموعه، بیش از ۲۰ مقاله نشانگر (Marker Papers) منتشر کرده که جزئیات تغییرات ژنومی و اپی‌ژنومی مرتبط با این انواع تومور را شرح می‌دهند. اگرچه اكتشافات مهمی توسط شبکه TCGA انجام شده، اما فرصت‌هایی برای اجرای روش‌های نوین، روش‌سازی مسیرهای بیولوژیکی جدید و کشف نشانگرهای تشخیصی همچنان وجود دارد. با این حال، استخراج داده‌ها از TCGA چالش‌های بیوانفورماتیکی متعددی را به همراه دارد، مانند بازیابی داده‌ها و یکپارچه‌سازی آن‌ها با داده‌های بالینی و سایر انواع داده‌های مولکولی (مانند RNA و متیلاسیون). ما یک بسته نرم‌افزاری R/Bioconductor را برای مقابله با این چالش‌ها توسعه دادیم. این بسته راه حل‌های بیوانفورماتیکی را با استفاده از یک گردش کار هدایت‌شده ارائه می‌دهد تا به کاربران اجازه دهد داده‌های TCGA را جستجو (Query)، دانلود (Download) و یادگیری (Bioconductor) کنند.

و تحلیل‌های یکپارچه‌سازی (Integrative Analyses) را بر روی آن‌ها انجام دهند. ما از چهار نوع تومور مختلف TCGA (کلیه، مغز، پستان و روده بزرگ) به عنوان مثال استفاده می‌کنیم تا مواردی از قابلیت بازتوانی (Reproducibility)، تحلیل یکپارچه‌سازی و استفاده از بسته‌های مختلف Bioconductor برای پیشبرد و تسريع اكتشافات جدید را نشان دهیم.

نکات کلیدی برای یادگیری

- داده‌های TCGA: بیش از ۱۰,۰۰۰ نمونه از ۳۳ نوع سرطان با داده‌های بالینی و مولکولی در دسترس عموم است.
- مشکل اصلی: استخراج و استفاده از این داده‌ها دشوار است، به ویژه در زمینه بازیابی داده‌ها و یکپارچه‌سازی انواع مختلف داده‌های مولکولی (مثل RNA) با داده‌های بالینی.
- راه حل: بسته نرم‌افزاری TCGAbiolinks در محیط R/Bioconductor توسعه داده شده است.
- عملکرد: به کاربران این امکان را می‌دهد که داده‌ها را جستجو، دانلود و تحلیل‌های یکپارچه‌سازی TCGAbiolinks انجام دهند.
- هدف: ارائه روش‌هایی برای بازتوانی (Reproducibility) نتایج TCGA و انجام تحلیل‌های پیشرفته و یکپارچه شده.

بخش دوم: مقدمه (Introduction)

ترجمه

پاراگراف‌های اول و دوم (Introduction): سرطان یکی از علل اصلی مرگ و میر در سراسر جهان است. پروژه اطلس ژنوم سرطان (TCGA) در سال ۲۰۰۶ با هدف جمع‌آوری و تحلیل داده‌های بالینی و مولکولی بیش از ۳۳ نوع تومور (با نمونه‌برداری از

حدود ۵۰۰ مورد برای هر نوع تومور) آغاز شد و تا به امروز جامع‌ترین مخزن داده‌های مولکولی و بالینی سرطان انسانی را تولید کرده است. تومورهای مورد بررسی توسط TCGA طیف گستردگی را شامل می‌شوند: از جامد تا هماتولوژیک، و از تهاجم کم تا بسیار تهاجمی (از نظر بقا)، و از خوش‌خیم تا متاستاتیک. برای هر مورد سرطان، RNA و پروتئین استخراج شد و سپس پروفایل‌های ژنومی، ترانسکریپتومی، اپی‌ژنومی و پروتئومیکس (اخیراً) با استفاده از مجموعه‌ای متنوع از پلتفرم‌های "امیکس'" (Omics) انجام شد. کنسرسیوم TCGA به چندین گروه کاری تقسیم شده است که هر کدام مسئول تولید، جمع‌آوری یا تحلیل داده‌ها هستند.

گروه‌های کاری تحلیل (AWGs) توسط اعضای جامعه علمی تشکیل شده‌اند تا تحلیل داده‌ها را برای هر نوع تومور یا سرطان‌های سیستمیک و اخیراً برای تحلیل پان‌سرطان (Pan-cancer) رهبری کنند. اعضای AWG داده‌های موجود در دسترس عموم را از طریق پورتال داده‌های TCGA دانلود و تحلیل می‌کنند. این یافته‌ها به توسعه نشانگرهای زیستی تشخیصی و پیش‌آگهی بالینی و همچنین بازتعریف طبقه‌بندی‌های قبلی تومورها منجر شده است. با وجود فراوانی و دسترسی به داده‌های TCGA، هنوز چالش‌های بزرگی برای بیوانفورماتیکدان‌ها، پزشکان و زیست‌شناسان مولکولی وجود دارد که علاوه‌مند به استفاده از این داده‌ها هستند.

✿ نکات کلیدی برای یادگیری

- TCGA: از سال ۲۰۰۶ شروع شد و جامع‌ترین مخزن داده‌های سرطان انسانی شامل بیش از ۳۳ نوع تومور را ارائه می‌کند.
- انواع داده‌های تولیدی: ژنومیکس، ترانسکریپتومیکس، اپی‌ژنومیکس و پروتئومیکس.
- گروه‌های کاری تحلیل (AWGs): متخصصان تحلیل که مسئول تولید گزارش‌های رسمی و مقالات نشانگر TCGA هستند.
- نتیجه تأثیرگذاری TCGA: توسعه نشانگرهای زیستی بالینی و بازتعریف طبقه‌بندی تومورها.
- مشکل با وجود داده‌ها: با وجود دسترسی بالا، هنوز چالش‌های عمدی برای محققان جهت مهار داده‌های TCGA و پیشبرد تحقیقاتشان وجود دارد.

بخش سوم: چالش‌ها و نیاز به ابزار جدید

ترجمه

پاراگراف‌های سوم، چهارم و پنجم (چالش‌ها): یکی از گروه‌های محققان که با چالش روبه‌رو هستند، تحلیلگرانی هستند که به دنبال بازتولید برخی از یافته‌های اصلی TCGA و ادغام روش‌های نوین در مراحل پیش‌پردازش، پردازش و فیلتر کردن (مانند نرم‌افزاری و انتخاب ویژگی) هستند. با این حال، داده‌ها و آرشیوهای TCGA دائمًا در حال تغییر هستند، چه به دلیل تولید داده‌های جدید و چه به دلیل ابطال داده‌های کم‌کیفیت یا اشتباہ.

برای همگام شدن با این ساختار پویا، سرویس وب مرکز هماهنگی داده (DCCWS) برای دسترسی به پایگاه داده TCGA در دسترس قرار گرفت. علاوه بر این، متدولوژی‌های تحلیلی TCGA معمولاً به صورت اسناد Sweave R یا اسکریپت‌های داخلی R ارائه می‌شوند که استفاده از آن‌ها برای بسیاری دشوار است. بسیاری از مطالعات، از جمله مقالات نشانگر TCGA، جداول تکمیلی خود را در وبسایتها خارجی، فایل‌های PDF یا جداول اکسل منتشر می‌کنند که تلاش برای بازتولید یافته‌ها یا یکپارچه‌سازی آن‌ها با داده‌های شخصی را چالش برانگیζتر می‌کند.

ابزارهای موجود و کمبودهای آن‌ها:

اخیراً چندین ابزار برای بازیابی داده‌های TCGA در دسترس قرار گرفته‌اند (مانند TCGA-Assembler، CGDS-R، TCGA-Assembler و cBioPortal و RTCGAToolbox). این ابزارها به سه دسته تقسیم می‌شوند: ۱. دانلود داده‌ها (مانند TCGA-Assembler)، ۲. تحلیل و یکپارچه‌سازی داده‌ها (مانند canEnvolve)، و ۳. دانلود و تحلیل داده‌ها (مانند RTCGAToolbox).

با وجود این بسته‌های نرم‌افزاری، هیچ‌کدام از آن‌ها تحلیل‌های یکپارچه‌سازی را با استفاده از متدولوژی‌های طراحی شده توسط AWG‌های TCGA (مانند شناسایی ژن‌های خاموش شده اپیژنتیک یا شناسایی تعداد کبی‌های عملکردی) انجام نمی‌دهند. همچنین، این ابزارها قادر به دانلود داده‌های آرشیو شده (نسخه‌های قدیمی) نیستند، که برای تحلیل مجدد مطالعات قبلی TCGA بسیار حیاتی است. نکته مهم دیگر این است که هیچ‌کدام از ابزارهای موجود داده‌های دانلود شده را به صورت آبجکت 'SummarizedExperiment' فراهم نمی‌کنند، در حالی که این فرمت برای یکپارچه‌سازی کامل و استفاده از سایر بسته‌های Bioconductor ضروری است.

✿ نکات کلیدی برای یادگیری

- چالش‌های داده‌ها: تغییر دائمی داده‌ها و آرشیوها و دشواری در بازتولید نتایج به دلیل ارائه متدولوژی‌ها به صورت اسکریپت‌های سفارشی.
- نقاط ضعف ابزارهای قبلی:
 1. عدم قابلیت یکپارچه‌سازی پیشرفته: انجام ندادن تحلیل‌های یکپارچه‌سازی با استفاده از متدولوژی‌های خاص TCGA.
 2. عدم پشتیبانی از نسخه‌های آرشیو: عدم توانایی در دانلود نسخه‌های قدیمی‌تر داده‌ها برای بازتولید مطالعات گذشته.
 3. عدم فرمت استاندارد: داده‌ها را در قالب استاندارد 'SummarizedExperiment' Bioconductor نمی‌دهند، که اتصال به سایر بسته‌های R/Bioconductor را دشوار می‌سازد.

بخش چهارم: معرفی TCGAbiolinks و اهداف آن

ترجمه

پاراگراف ششم (معرفی TCGAbiolinks) : ما در اینجا ابزار نرمافزاری جدیدی به نام TCGAbiolinks را توصیف می کنیم که به جستجو، دانلود، تحلیل و یکپارچهسازی داده های TCGA در قالب یک بسته جمعی Bioconductor کمک می کند . TCGAbiolinks به طور انحصاری در زبان R توسعه یافته و از بسیاری از طراحی های بسته و آجکت های مشخص شده توسط استفاده می کند . پروژه Bioconductor نرمافزاری با کیفیت بالا، مستندات خوب و قابلیت همکاری Bioconductor را تضمین می کند و امکان یکپارچهسازی با صدها بسته دیگر در R را فراهم می سازد (Interoperable).

اهداف چهارگانه TCGAbiolinks عبارتند از:

1. تسهیل بازیابی داده ها از طریق DCCWS (سرвис وب مرکز هماهنگی داده) TCGA.
2. آماده سازی داده ها با استفاده از استراتژی های پیش پردازش مناسب.
3. فراهم کردن ابزاری برای انجام تحلیل های استاندارد و تحلیل های یکپارچه سازی پیشرفته.
4. امکان دانلود یک نسخه خاص از داده ها و در نتیجه، باز تولید آسان نتایج تحقیقات قبلی.

این ابزار برای اکثر داده های TCGA از کلاس 'SummarizedExperiment' Bioconductor استفاده می کند، که امکان یکپارچه سازی آسان با سایر انواع داده ها و روش های آماری در مخزن Bioconductor را می دهد.

نکات کلیدی برای یادگیری

- TCGAbiolinks یک ابزار جدید R/Bioconductor است که به طور خاص برای غلبه بر چالش های قبلی طراحی شده است.
- چهار هدف اصلی:
 - 1. بازیابی آسان داده ها.
 - 2. آماده سازی مناسب داده ها (پیش پردازش).
 - 3. انجام تحلیل های استاندارد و پیشرفته یکپارچه سازی.
 - 4. باز تولید نتایج قبلی با دانلود نسخه های خاص داده ها.

پشتیبانی از Bioconductor : استفاده از کلاس 'SummarizedExperiment' به عنوان استاندارد، برای اطمینان از یکپارچه سازی کامل با صدها بسته تحلیلی دیگر در R.

بخش پنجم: ساختار بسته نرم افزاری (Materials and Methods)

ترجمه

پاراگراف‌های اول (ساختار TCGAbiolinks: (TCGAbiolinks) تحت مجوز GPLv3 است که در دسترس Bioconductor از توابعی تشکیل شده است که می‌توانند در سه سطح اصلی گروه‌بندی شوند:

1. داده (Data) : مدیریت بازیابی و جستجوی داده‌ها.

2. تحلیل (Analysis) : انجام تحلیل‌های آماری و بیوانفورماتیکی.

3. تصویرسازی (Visualization) : ارائه بصری نتایج تحلیل.

این بسته روش‌های متعددی را برای تحلیل پلتفرم‌های آزمایشگاهی فردی (مانند تحلیل بیان افتراقی یا شناسایی مناطق متیله‌شده افتراقی) و روش‌هایی برای تصویرسازی (مانند نمودارهای بقا، نمودارهای آتشفسانی و نمودارهای ستاره‌ای) فراهم می‌کند. علاوه بر این، TCGAbiolinks تحلیل عمیق یکپارچه‌سازی چندین پلتفرم (مانند تعداد کپی و بیان، یا بیان و متیلاسیون DNA) را ارائه می‌دهد.

سطح 1: توابع داده (Data Functions) این سطح بازیابی و جستجوی داده‌های TCGA را مدیریت می‌کند و به سه تابع اصلی تقسیم می‌شود:

- TCGAquery : به کاربر اجازه می‌دهد تا داده‌های جدید و آرشیو شده را از پورتال داده TCGA جستجو کند و نمونه‌هایی را برای دانلود انتخاب کند TCGA . داده‌هایی از بیش از ۲۴ نوع سرطان و ۶ نوع داده مولکولی (mRNA, miRNA, پروتئین، متیلاسیون و اگزوم) او ۳ نوع گزارش بالینی را ارائه می‌کند.
- TCGAdownload : تابعی که داده‌های شرح داده شده در لیست نمونه‌های ارائه شده توسط TCGAquery را دانلود می‌کند . این تابع به کاربر اجازه می‌دهد تا داده‌ها را بر اساس سطح (۱، ۲ یا ۳) یا نمونه‌های مشترک بین دو یا سه پلتفرم مختلف انتخاب کند.

- TCGAprepare : تابعی که داده‌های آزمایش‌های سطح سه را می‌خواند و آن‌ها را برای تحلیل‌های بعدی آماده‌سازی می‌کند . به طور خاص، آبجکت‌ها در قالب 'SummarizedExperiment' خلاصه می‌شوند تا یکپارچه‌سازی آسانی با سایر بسته‌های Bioconductor داشته باشند.

نکات کلیدی برای یادگیری

- سه سطح اصلی TCGAbiolinks : داده (Data) ، تحلیل (Analysis) ، و تصویرسازی (Visualization).

قابلیت‌های تحلیل و تصویرسازی: تحلیل‌های مانند بیان افتراقی (Differential Expression) و مناطق متیله شده (Differentially Methylated Regions) و تصویرسازی‌هایی مانند نمودارهای بقا و نمودار ستاره‌ای (Starburst Plot).

• توابع داده (Data Functions)

- برای جستجو و انتخاب نمونه‌ها (شامل داده‌های آرشیو شده). TCGAquery
- برای دانلود داده‌ها بر اساس لیست‌های انتخابی و سطوح داده (۱، ۲ یا ۳). TCGAdownload
- برای آماده‌سازی داده‌ها، به خصوص تبدیل داده‌های سطح سه به آبجکت استاندارد TCGAprepare . 'SummarizedExperiment'

بخش ششم: توابع تحلیل و تصویرسازی (Analysis and Visualization)

ترجمه

سطح ۲: توابع تحلیل (Analysis Functions) توابع تحلیل برای تجزیه و تحلیل داده‌های TCGA از طریق روش‌های متداول و نوین طراحی شده‌اند. تحلیل‌های بعدی (Downstream) می‌توانند به تحلیل نظارت‌شده (Supervised) (مانند تحلیل بیان افتراقی و تحلیل غنی‌سازی) یا تحلیل نظارت‌نشده (Unsupervised) (مانند خوشبندی و تحلیل بقا) تقسیم شوند.

• TCGAanalyze Normalization: به کاربران اجازه می‌دهد تا رونوشت‌های mRNA و miRNA را با استفاده از EDASeq نرمال‌سازی کنند.

• TCGAanalyze DEA: به کاربر اجازه می‌دهد تا بیان یا مناطق افتراقی (Differential Expression or DEGs) را بین دو گروه شناسایی کند. این تابع از بسته edgeR برای تشخیص ژن‌های بیان افتراقی (Fold Change) و اهمیت آماری (FDR) فیلتر کرد.

• TCGAanalyze DMR: برای شناسایی مناطق متیله شده افتراقی (DMRs) بین دو گروه استفاده می‌شود.

• TCGAanalyze Clustering: به کاربر اجازه می‌دهد تا تحلیل خوشبندی سلسله مراتبی را انجام دهد.

سطح ۳: توابع تصویرسازی (Visualization Functions) بخش تصویرسازی به کاربر این امکان را می‌دهد که نتایج تولید شده توسط بخش تحلیل را با استفاده از نقشه‌های حرارتی (Heatmap)، نمودارهای خوشبندی، نمودارهای آتشفشنای (Volcano Plot)، تحلیل غنی‌سازی مسیرها و تحلیل مؤلفه‌های اصلی (PCA) به تصویر بکشد. علاوه بر این، روش‌هایی برای تولید نمودار ستاره‌ای (Starburst Plot) ارائه می‌شود که نتایج متیلاسیون DNA و بیان ژن TCGA را یکپارچه می‌کند.

نکات کلیدی برای یادگیری

- تحلیل‌های پشتیبانی شده: نرم‌افزاری، تحلیل بیان افتراقی (DEA)، شناسایی مناطق متشابه شده افتراقی (DMR) و خوشبندی.
- الگوریتم‌های استفاده شده: استفاده از بسته‌های شناخته شده Bioconductor مانند edgeR برای تحلیل بیان افتراقی.
- قابلیت‌های تصویرسازی: تولید نقشه‌های حرارتی، نمودار آتشفسانی، تحلیل PCA، و مهم‌تر از همه، نمودار ستاره‌ای (Starburst Plot) که برای نمایش نتایج تحلیل یکپارچه‌سازی متیلاسیون و بیان ژن طراحی شده است.
- تحلیل یکپارچه‌سازی TCGAbiolinks: برای تحلیل یکپارچه‌سازی داده‌ها (مانند متیلاسیون DNA و بیان ژن) طراحی شده است.

بخش هفتم: نتیجه‌گیری (Conclusion)

ترجمه

پاراگراف آخر (نتیجه‌گیری): ما در اینجا ابزار جدیدی به نام TCGAbiolinks را معرفی کردیم که به صورت رایگان در پروژه Bioconductor در دسترس است. چندینتابع مفید را برای جستجو، دانلود و آماده‌سازی نمونه‌های TCGA برای تحلیل داده‌ها فراهم می‌کند. این توابع فرصتی را برای کاربران نهایی فراهم می‌کنند تا داده‌های TCGA را به آسانی و بدون نیاز به پیمایش در پortal‌های مختلف داده جمع‌آوری کنند.

اگرچه ابزارهای زیادی برای دانلود داده‌های TCGA وجود دارند، اما قبل‌آمیخته کدام قادر به دانلود داده‌های آرشیو شده یا زیرگروه‌های بالینی منتشرشده نبودند، که این امر برای بازتولید نتایج TCGA حیاتی است. همچنین، تا قبل از TCGAbiolinks، هیچ ابزار موجودی وجود نداشت که بتواند بیان ژن و متیلاسیون DNA یا تعداد کبی و بیان ژن را به طور کامل و در قالب یک بسته استاندارد و قابل بازتولید یکپارچه کند.

ما نشان دادیم که با استفاده از TCGAbiolinks می‌توانیم نتایج مطالعات قبلی ارائه شده توسط شبکه تحقیقاتی TCGA را بازتولید کنیم. همچنین، با گنجاندن اطلاعات زیرگروه بالینی و مولکولی موجود، کاربران اکنون می‌توانند نشانگرهای زیستی برای زیرگروه‌های توموری خاص را بر اساس همبستگی با بقا شناسایی کنند TCGAbiolinks. یک مجموعه جامع از خطوط لوله (Pipelines) را فراهم می‌کند و به کاربران امکان بازتولید و انجام تحلیل‌های یکپارچه‌سازی داده‌های TCGA را می‌دهد. به عنوان یک بسته Bioconductor، ابزار ما می‌تواند داده‌های دانلود شده TCGA را برای یکپارچه‌سازی با بسته‌های موجود آماده کند و دسترسی به انبوهی از تحلیل‌های آماری را که اکنون در حال بررسی هستند، برای کاربران نهایی فراهم کند.

نکات کلیدی برای یادگیری

- تفاوت کلیدی TCGAbiolinks : قابلیت دانلود داده‌های آرشیو شده و زیرگروه‌های بالینی منتشرشده و انجام تحلیل یکپارچه‌سازی کامل بیان ژن با متیلاسیون/DNA تعداد کمی.
- تأیید قابلیت: این بسته توانایی بازتولید نتایج منتشرشده توسط شبکه TCGA را دارد.
- نقش در آینده TCGAbiolinks : یک مجموعه جامع از ابزارها را برای بازتولید و تحلیل یکپارچه‌سازی فراهم می‌کند و با استاندارد Bioconductors ، دروازه‌ای به روی مجموعه‌ای غنی از ابزارهای آماری برای تحقیقات ژنومیک باز می‌کند.