دانشکده علوم ریاضی

گروه علوم کامپیوتر

گزارش فنی درس سمینار

شناسایی امضاهای جهشی در تومورهای سرطانی فردی با استفاده از رویکردهای تجزیه محور

توسط:

بهار مهدوی

40152521337

استاد درس:

دکتر منصور رزقی آهق

20 اسفند 1402

به نام خداوند جان آفرین حکیم سخن در زبان آفرین

چکیده

توالی‌یابی کل ژنوم، جامعه ژنومیک سرطان را وارد قلمرو جدیدی کرده است. به لطف قدرت محض ارائه شده توسط هزاران جهش موجود در هر بیمار سرطانی، ما قادر به تشخیص الگوهای ژنریک جهش‌ها، به نام "امضاهای جهشی"، که در طول تومورزایی به وجود می آیند، شده ایم. تجمع جهش‌های پیکری در یک ژنوم نتیجه فعالیت یک یا چند فرآیند جهش‌زایی است که هر کدام اثر خود را به جا می‌گذارند. مطالعه این امضاهای جهشی، دارای پتانسیل مهمی برای درک بیشتر ما از علل و تکامل سرطان است و می‌تواند بینش‌های جدیدی در مورد علل سرطان‌های فردی و پیشگیری و درمان سرطان ارائه دهد. تجزیه و تحلیل امضاهای جهشی یک رویکرد قدرتمند برای درک فرآیندهای جهش‌زای عوامل درون‌زا و برون‌زا است که تکامل ژنوم سرطان را شکل داده است. برای ارزیابی امضاهای جهشی فعال در یک ژنوم سرطانی، ابتدا باید فعالیت‌های آن‌ها را با تخمین تعداد جهش‌های حک شده توسط هر امضا، کمی‌سازی کرد. این زمینه‌ای است که در جهت‌های محاسباتی، تجربی و بالینی در حال توسعه و گسترش است. در این بررسی، توجه بر روی مدل‌های ریاضی و تکنیک‌های محاسباتی متمرکز است که باعث پیشرفت‌های اخیر در این زمینه شده‌اند.

واژه‌های کلیدی

جهش زایی، امضاهای جهشی، ژنومیک سرطان، مدل‌سازی ریاضی، روش‌های محاسباتی

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست رابطه‌ها

عنوان **صفحه**

فصل اول

مقدمه

1.1 مقدمه

پیشرفت‌های اخیر در فن‌آوری‌های توالی‌یابی DNA با کارایی بالا[1]، مطالعاتی را امکان‌پذیر کرده است که هزاران ژنوم یا اگزوم سرطان کامل[2] را بررسی می‌کند. توالی‌یابی کل ژنوم[3]، جامعه ژنومیک سرطان را وارد قلمرو جدیدی کرده است. سرطان یک بیماری ژنومی است که در آن تکثیر کلونال[4] کنترل نشده توسط تغییرات ژنومی در سلول‌های پیکری[5] شروع و تقویت می‌شود (Stratton, Campbell et al. 2009). علیرغم این واقعیت که یک ژنوم سرطان ممکن است بین ده‌ها تا میلیون‌ها جهش پیکری را حمل کند (Alexandrov, Nik-Zainal et al. 2013, Vogelstein, Papadopoulos et al. 2013)، تنها زیرمجموعه کوچکی از این جهش‌ها که جهش‌های "محرک[6]" نامیده می‌شوند، باعث گسترش نئوپلاستیک[7] می‌شوند (Beerenwinkel, Antal et al. 2007, Stratton, Campbell et al. 2009). عموماً اعتقاد بر این است که مابقی جهش‌ها

[1] high-throughput DNA sequencing technologies
[2] whole cancer genomes or exomes
[3] whole-genome sequencing
[4] clonal proliferation
[5] somatic cells
[6] driver
[7] neoplastic expansion

تحت عنوان جهش‌های "مسافری[۸]"، مزیت انتخابی در فرآیندهای دخیل در جهش‌زایی ایجاد نمی‌کنند (Attolini and Michor 2009, Yates and Campbell 2012).

تجمع جهش‌های محرک در ژنوم سلول پیکری نتیجه یک یا چند فرآیند جهش‌زایی است که به طور مداوم یا متناوب در طول عمر ارگانیسم عمل می‌کند (Alexandrov and Stratton 2014). چنین فرآیندهای جهش‌زا، شامل آسیب DNA [۹] توسط ژنوتوکسین عوامل برون زا[۱۰] یا درون زا[۱۱]، همانندسازی معیوب DNA، درج عناصر قابل انتقال[۱۲]، نقص در مکانیسم‌های ترمیم DNA [۱۳] و ویرایش آنزیمی DNA و غیره است (Roberts and Gordenin 2014). بسیاری از این فرآیندها الگوی مشخصی از جهش‌ها را در ژنوم نشان می‌دهند که به عنوان "امضای جهشی[۱۴]" شناخته می‌شود (Pfeifer 2010, Alexandrov, Nik-Zainal et al. 2013). بنابراین، خلاصه تغییرات سوماتیکی در ژنوم سرطان، سابقه‌ای از اثر جهش‌زایی ترکیبی از مخلوط خاصی از فرآیندهای ایجاد کننده آن را تشکیل می‌دهد (Alexandrov, Nik-Zainal et al. 2013). علاوه بر این، از آنجا که بیشتر جهش‌ها مسافر هستند، تا حد زیادی فراتر از تأثیر انتخاب تطبیقی[۱۵] هستند (Rubin and Green 2009).

در تلاش برای شناسایی امضاهای جهشی در یک مجموعه داده، می‌توان یک رویکرد "جهانی[۱۶]" اتخاذ کرد، که در آن امضاهای همه سرطان‌ها، صرف نظر از نوع بافت، جمع شده و میانگین گرفته می‌شود تا مجموعه‌ای از امضاهای اجماع[۱۷] به دست آید (Alexandrov, Nik-Zainal et al. 2013, Alexandrov, Kim et al. 2020). این رویکرد پیش‌فرض می‌گیرد که نمونه‌های بیشتر قدرت بیشتری برای تشخیص امضاهای جدید فراهم می‌کنند. با این حال، یک آنالیز جامع، همچنین با نادیده گرفتن ویژگی‌های امضای خاص بافتی احتمالی که منعکس کننده زیست‌شناسی خاص اندامی است، فرض می‌کند که امضاها در تمام بافت‌ها یکسان هستند و این اخیراً به عنوان یک احتمال برجسته شده است (Degasperi, Amarante et al. 2020). در واقع، تعداد نمونه‌ها به ازای هر نوع

[8] passenger
[9] DNA damage
[10] exogenous
[11] endogenous
[12] insertion of transposable elements
[13] defects in DNA repair mechanisms
[14] mutational signature
[15] adaptive selection
[16] global
[17] consensus signatures

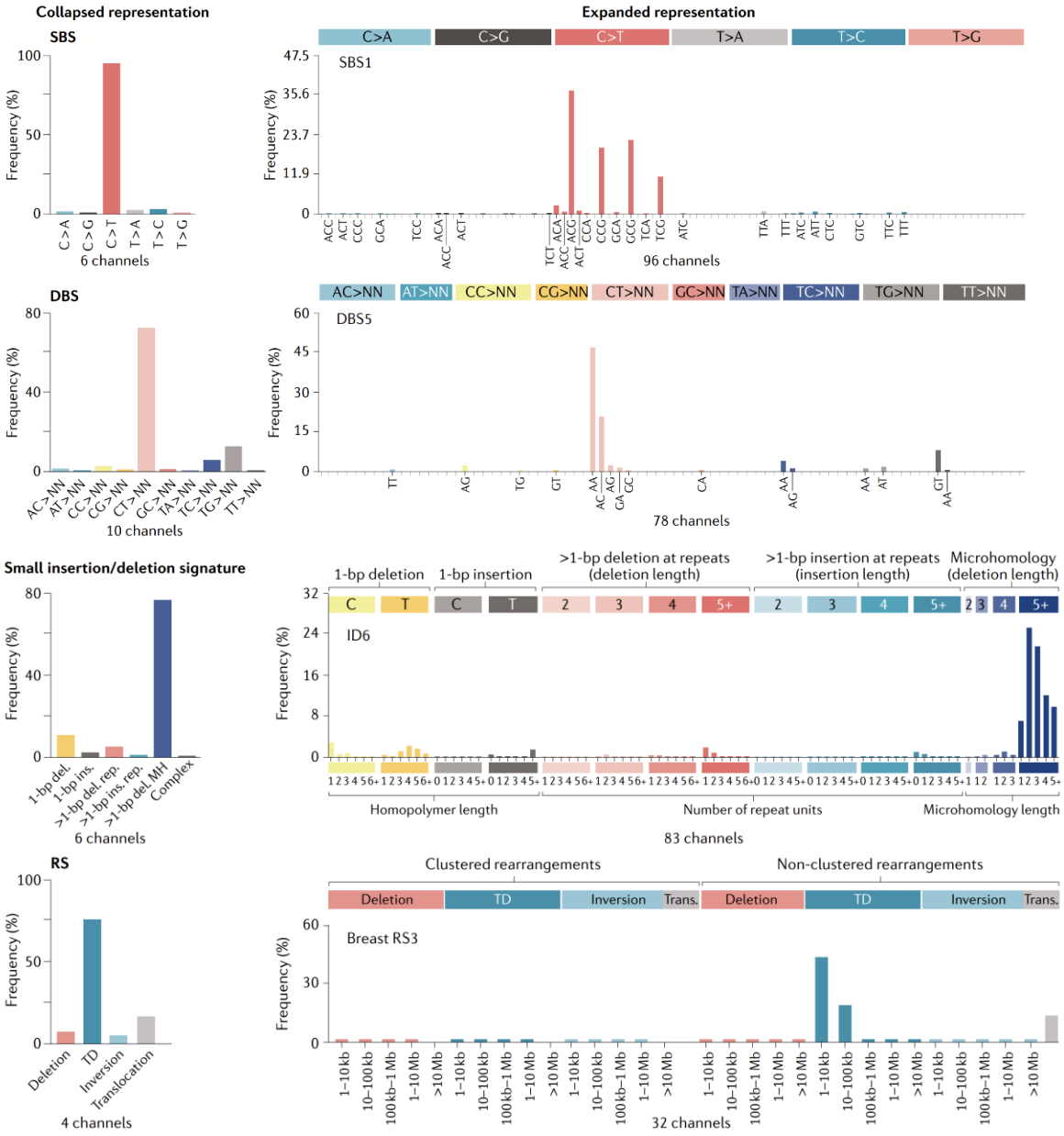
تومور در آنالیزهای گذشته نامتعادل بوده است، در نتیجه در امضای انواع بافت‌های خاص تأثیرگذارتر بوده و به این ترتیب سوگیری بالقوه‌ای[18] را معرفی می‌کند (Alexandrov, Nik-Zainal et al. 2013, Alexandrov, Kim et al. 2020). در مقابل، یک رویکرد "محلی[19]" استخراج امضا را در انواع بافت‌های فردی محدود می‌کند و متعاقباً امضاهای استخراج شده به صورت محلی را بین اندام‌های مختلف مقایسه می‌کند (Degasperi, Amarante et al. 2020). این اجازه می‌دهد تا تغییرات طبیعی[20] در بین بافت‌های مختلف ظاهر شود. در اینجا، ما از اصطلاحات "جهانی" و "محلی" هنگام بحث در مورد امضاها استفاده می‌کنیم.

یکی دیگر از ویژگی‌های مهم "کانال‌هایی[21]" است که جهش‌ها را در امضاهای جایگزین[22]، امضاهای درج یا حذف[23] (ایندل) یا (ID) و امضاهای بازآرایی[24] (RS) طبقه‌بندی می‌کنند. از نظر تاریخی، جایگزین‌های تک باز، با ترکیب زمینه‌های توالی طرفین هر جایگزین احتمالی طبقه‌بندی می‌شدند، که منجر به یک الگوی 96 کانالی برای SBSs شد (Nik-Zainal, Alexandrov et al. 2012, Koh, Zou et al. 2020) (**شکل 1**). امضاهای جایگزین دو بازی[25] (DBS) توسط 78 کانال تعریف می‌شوند (Kucab, Zou et al. 2019, Alexandrov, Kim et al. 2020). طبقه‌بندی کانال‌های ایندل و بازآرایی با توجه به اینکه جدیدتر هستند و هنوز به طور گسترده مورد استفاده قرار نگرفته‌اند، در ادامه با جزئیات بیشتری توضیح داده شده‌اند.

شکل 1، تحولات مفهومی و تجسمی امضاهای جهشی را نشان می‌دهد. با توان کافی، روش ارجح برای ارائه امضاهای جهشی جایگزینی تک باز (SBS؛ به عنوان مثال، SBS1) از طریق روش 96-کانالی است. همچنین امکان گسترش روش به 1536 کانال (نشان داده نشده) یا کاهش آن به تنها شش کانال وجود دارد. امضاهای جایگزینی دو بازی (DBS ها؛ به عنوان مثال، DBS5) را می توان با 78 کانال آگنوستیک رشته[26] یا زمانی که بار کم است، با ده موتیف دوتایی[27] تعریف کرد.

---

[18] potential bias
[19] local
[20] natural variation
[21] channels
[22] substitution signatures
[23] insertion or deletion (indel) signatures (IDs)
[24] rearrangement signatures (RSs)
[25] Double-base substitution signatures (DBSs)
[26] strand-agnostic
[27] duplet motifs



شکل 1. تحولات مفهومی و تجسمی امضاهای جهشی

امضاهای درج یا حذف (ایندل) کوچک (کمتر از 100 جفت باز) (به عنوان مثال، امضای ایندل 6 (ID6)) به طور گسترده بر اساس نوع (یعنی درج[۲۸]، حذف[۲۹] یا پیچیده[۳۰]) و - زمانی که تک باز هستند - به عنوان C یا T، و با توجه به طول دنباله تکراری مونونوکلئوتیدی که در آن رخ می‌دهند، طبقه‌بندی می‌شوند. ایندل های طولانی‌تر بر اساس اینکه آیا در تکرارها رخ می‌دهند یا دارای میکروهمولوژی[۳۱] در اتصالات ایندل هستند، طبقه‌بندی می‌شوند. با قدرت کافی، اندازه‌های موتیف و نوکلئوتیدهای تحت تأثیر نیز می‌توانند در نظر گرفته شوند. امضاهای بازآرایی (RS ها؛ برای مثال RS3 سینه) را می‌توان بر اساس چهار نوع بازآرایی و نحوه خوشه‌بندی آنها به صورت منطقه‌ای و با در نظر گرفتن بیشتر اندازه قطعه بازآرایی شده دسته‌بندی کرد. در مواردی که بار جهش کم است (مثلاً در سیستم‌های تجربی) ممکن است ارائه‌های جمع‌شده ضروری باشند. حذف (.del)، درج (.ins)، میکروهومولوژی (MH)، تکرار (.rep)، تکرار پشت سرهم[۳۲] (TD). جابجایی[۳۳] (.trans) (Nik-Zainal, Davies et al. 2016, Alexandrov, Kim et al. 2020)

1.1.1 امضاهای جایگزینی بازی

یک لیست در حال رشد از SBS ها و DBS های گزارش شده وجود دارد (**شکل 1**). همچنین تعداد فزاینده‌ای از تکنیک‌های استنتاجی برای شناسایی آن‌ها وجود دارد (Alexandrov, Nik-Zainal et al. 2013, Rosenthal, McGranahan et al. 2016, Blokzijl, Janssen et al. 2018, Huang, Wojtowicz et al. 2018, Cartolano, Abedpour et al. 2020, Degasperi, Amarante et al. 2020, Fantini, Vidimar et al. 2020). صرف‌نظر از الگوریتم‌های مورد استفاده برای شناسایی امضا، امضاهای رایج معمولاً در اکثر گروه‌های مورد بررسی، به‌طور مداوم قابل شناسایی هستند، برای مثال SBS1 (Alexandrov, Nik-Zainal et al. 2013)، ناشی از دآمیناسیون 5-متیل سیتوزین[۳۴]. به همین ترتیب، امضاهای مرتبط با مواجهه‌های محیطی[۳۵] که تمایل دارند بلافاصله مشهود شوند (به عنوان مثال، SBS7 و DBS1 مرتبط با UV) (Alexandrov, Kim et al. 2020). کمبود در مسیرهای خاص ترمیم DNA باعث جهش‌زایی مشخصی

[28] insertion
[29] deletion
[30] complex
[31] microhomology
[32] tandem duplication
[33] translocation
[34] deamination of 5-methylcytosine
[35] environmental exposures

می‌شود مانند SBS26 و SBS44 (Alexandrov, Kim et al. 2020) مرتبط با کمبود ترمیم ناهماهنگی[۳۶] (MMR)؛ و برخی از امضاهای درون‌زا بسیار متمایز و به راحتی قابل تشخیص هستند، از قبیل آنزیم ویرایش کننده mRNA آپولیپوپروتئین B [۳۷]، آنزیم کاتالیزوری شبه پلی پپتیدی (APOBEC) [۳۸] مرتبط با SBS2 و SBS13 (Alexandrov, Nik-Zainal et al. 2013, Petljak, Alexandrov et al. 2019, Alexandrov, Kim et al. 2020). فرآیندهای جهشی نادری که در فرکانس‌های جمعیتی پایین وجود دارند ممکن است استخراج چالش‌برانگیزتری داشته باشد و تنها در صورتی خود را نشان دهند که در گروه خاص مورد بررسی (مثلاً امضاهای مرتبط با درمان[۳۹]) وجود داشته باشند. هدف ما در این بررسی تمرکز بر اصول راهنما و نه بحث درباره همه امضاها به صورت جداگانه است. این اطلاعات از منابع آنلاین مختلف، مانند COSMIC یا Signal قابل دسترسی است، جایی که ممکن است محتوای دقیق در طول زمان تغییر کند. یک مجموعه مرجع که اغلب در تجزیه و تحلیل استفاده می‌شود، مجموعه COSMIC v2 از 30 امضا است. مجموعه COSMIC v3.1 که در سال 2020 منتشر شد، تعداد کل امضاها را به 49 رساند (Alexandrov, Kim et al. 2020). چندین امضا به عنوان امضاهای مرتبط با درمان ذکر شد (به عنوان مثال، SBS31 و SBS35، مرتبط با پلاتین[۴۰]؛ SBS90 منتسب به دووکارمایسین[۴۱])، و برخی از امضاها به طور قابل توجهی در مجموعه COSMIC v3.1 نسبت به مجموعه COSMIC v2 تغییر یافت (به عنوان مثال، SBS1 و SBS16). اینکه آیا این اصلاحات از نظر بیولوژیکی درست هستند یا صرفاً یک نتیجه ریاضی هستند، چندان روشن نیست و منتظر تأیید مستقل تجربی است. نکته مهم این است که با افزایش تعداد امضاهای مرجع، استفاده و تفسیر دقیق آن‌ها با مشکلاتی همراه است.

1.1.2 امضاهای ایندل

در مقایسه با جایگزین‌ها، ایندل‌های کوچک (کمتر از 100 جفت باز) به دلیل دشواری تاریخی در دست‌یابی به داده‌های ایندل با کیفیت بالا مورد، بررسی قرار نمی‌گیرند. با این وجود، ایندل‌ها در سرطان‌ها رایج هستند و در حدود 10 درصد فراوانی که در آن جایگزینی رخ می‌دهد، و جایگاه ژنومی و ترکیب توالی آن‌ها غیر تصادفی است، اتفاق می‌افتد (Nik-Zainal, Alexandrov et al. 2012, Alexandrov, Nik-Zainal et al. 2013). بنابراین، ایندل‌ها همچنین می‌توانند به عنوان امضاهای بینشی بیولوژیکی ارائه شوند. همیشه نمی‌توان ایندل‌ها را با

[36] mismatch repair
[37] apolipoprotein B mRNA editing enzyme
[38] catalytic polypeptide-like (APOBEC)
[39] treatment-associated signatures
[40] platinum
[41] duocarmycin

یک مختصات تعریف شده با همان دقت جایگزینی‌ها در نظر گرفت، زیرا تعیین موقعیت جهش حذف شده یا درج شده در یک مسیر تکراری پلی‌نوکلئوتیدی غیرممکن است. به این ترتیب، ایندل‌ها بر اساس نوع آن‌ها (حذف، درج یا پیچیده)، اندازه و اینکه آیا ویژگی‌هایی در اتصالات ایندل وجود دارد که می‌توانند زیربنای بیولوژیکی را آشکار کنند، به سادگی طبقه‌بندی شده‌اند (Nik-Zainal, Alexandrov et al. 2012). به عنوان مثال، ایندل‌های 1 جفت باز که در مسیرهای تکراری اتفاق می‌افتند معمولاً از لغزش رشته[۴۲] در طول تکثیر[۴۳] ایجاد می‌شوند، در حالی که ایندل‌هایی که یک توالی میکروهومولوگ مشترک با توالی کناری دارند، تصور می‌شود که اسکارهای ترمیم ناقص شکستگی‌های دو رشته‌ای[۴۴] توسط فرآیندهای جایگزین اتصال-انتهایی[۴۵] هستند (Helleday, Eshtad et al. 2014). این طبقه‌بندی ساده اساس شناسایی سرطان‌های دارای کمبود MMR و کمبود نوترکیبی هموLوگ[۴۶] (HRD) را تشکیل داد (Nik-Zainal, Alexandrov et al. 2012, Davies, Morganella et al. 2017).

برای استخراج ID ها، آنالیز جهانی 2780 سرطان از انواع بافت‌های متعدد بر روی ایندل‌های طبقه‌بندی شده بر اساس مجموعه‌ای از 83 کانال انجام شد (Alexandrov, Kim et al. 2020)، که بسط داده شده‌ی طبقه‌بندی ایندل قبلی بود (Nik-Zainal, Alexandrov et al. 2012, Alexandrov, Nik-Zainal et al. 2013)؛ برای مثال، ایندل‌های تک بازی بر اساس طول عددی مسیر تکراری که در آن رخ داده‌اند، طبقه‌بندی شدند (**شکل 1**). هفده ID گزارش شد (Alexandrov, Kim et al. 2020). ID6 نشان دهنده یک امضای حذف با واسطه میکروهومولوژی[۴۷] است که در سرطان‌های جهش یافته با BRCA1 و جهش یافته با BRCA2 که قبلاً توضیح داده شد مشاهده می‌شود (Nik-Zainal, Alexandrov et al. 2012). ID1، ID2 و ID7هایی با واسطه تکراری بودند که در تومورهای دارای جهش در دومین‌های تصحیح‌کننده[۴۸] کمبود POLE یا POLD1 و/یا MMR بسیار بالا بودند. ناشی از لغزش همانندسازی[۴۹]، ID1 و ID2 A یا ایندل های T در مسیرهای پلی طویل[۵۰] (dA:dT) نیز در اکثر نمونه‌ها، از جمله بافت‌های طبیعی یافت شدند (Lee-Six, Olafsson et al. 2019).

[42] strand slippage
[43] replication
[44] double-strand breaks
[45] end-joining processes
[46] homologous recombination deficiency (HRD)
[47] microhomology
[48] proofreading domains
[49] replication slippage
[50] long poly

ID3 با سیگار کشیدن و ID13 با قرار گرفتن در معرض اشعه ماوراء بنفش مرتبط بود. اعتقاد بر این بود که ID8 ردپای[۵۱] اتصال-انتهایی غیرهمولوگ[۵۲] براساس میکروهمولوژی 1 جفت باز یا عدم وجود میکروهومولوژی در اتصالات ایندل است (Alexandrov, Kim et al. 2020). زیرمجموعه‌ای از تومورهای ID8 نیز دارای ID17 بودند که طبق گزارش‌ها با جهش سوماتیک TOP2A K743N مرتبط بود (Alexandrov, Kim et al. 2020). ID1، ID2، ID5 و ID8 با سن بیمار در هنگام تشخیص همبستگی پیدا کردند، که نشان دهنده مکانیسم مبتنی بر همانندسازی است (Alexandrov, Kim et al. 2020). علل ID 9 باقی مانده ناشناخته است. چندین کانال پیشنهادی در سراسر گروه حاوی اطلاعات مفید نبودند، و بنابراین بهینه‌سازی بیشتر مورد نیاز است. روش‌های جایگزین برای طبقه‌بندی ایندل‌ها آزمایش نشده‌اند و ممکن است بینش‌های بیولوژیکی را نشان دهند که توسط این رویکرد دستگیر نشده‌اند.

1.1.3 امضاهای بازآرایی

یکی دیگر از دسته‌های مهم جهش‌های سوماتیک، تغییرات یا بازآرایی‌های ساختاری است که ممکن است تکه‌های نسبتاً بزرگی از مواد کروموزومی را در هر جهتی[۵۳] در مقیاس کیلوباز[۵۴] تا مگاباز[۵۵] حذف، دو نسخه[۵۶] و/یا دوباره سوار کند[۵۷]. با استفاده از یک چارچوب فاکتورسازی ماتریس غیر منفی[۵۸] (Alexandrov, Nik-Zainal et al. 2013)، 32 کانال طبقه‌بندی برای RSهای فرضی استخراج شده از آنالیز محلی 560 WGS سرطان سینه پیشنهاد شد (Nik-Zainal, Davies et al. 2016). کانال‌ها نحوه خوشه‌بندی منطقه‌ای[۵۹] نقاط شکست بازآرایی[۶۰]، نوع بازآرایی (به عنوان مثال، حذف، تکراری پشت سر هم[۶۱] (TD)، وارونگی[۶۲] یا جابه‌جایی[۶۳]) و اندازه بازآرایی (**شکل 1**) را در نظر گرفتند. سه مورد از شش RS شناسایی شده، با تومور HRD همبستگی داشتند:

---

[51] footprint
[52] non-homologous
[53] orientation
[54] kilobase
[55] megabase
[56] duplicate
[57] reassemble
[58] non-negative matrix factorization framework
[59] regionally clustered
[60] rearrangement breakpoints
[61] tandem duplication (TD)
[62] inversion
[63] translocation

سرطان‌های دارای جهش BRCA1 و فاقد جهش BRCA2 تعداد بالایی از TDهای کوچک RS3 (کمتر از 10 کیلوباز) را نشان دادند، در حالی که سرطان‌های دارای جهش BRCA1 یا BRCA2 تعداد قابل توجهی حذف RS5 (کمتر از 10 کیلوباز) را نشان دادند. علت TDهای طویل RS1 (بیش از 100 کیلوباز) که با HRD نیز مرتبط است، شناخته نشده است (Nik-Zainal, Davies et al. 2016). تعداد RSها اخیراً به 15 افزایش یافته است (Degasperi, Amarante et al. 2020). طرح طبقه بندی 32 کانالی (Nik-Zainal, Davies et al. 2016) نیز برای گزارش امضاها در گروه‌های سرطان کبد و تخمدان استفاده شده است (Letouzé, Shinde et al. 2017, Hillman, Chisholm et al. 2018).

اخیراً، با استفاده از یک فرآیند دیریکله سلسله مراتبی[۶۴]، گروه کاری تغییرات ساختاری آنالیز تمام سرطان ژنوم کامل[۶۵] (PCAWG) RS 16 را در یک آنالیز جهانی از حدود WGS 2559 سرطان اولیه[۶۶] درگیر در حدود 150000 تغییرات ساختاری گزارش کرده است (Li, Roberts et al. 2020). علاوه بر کلاس‌های بازآرایی مرسوم، نویسندگان مطالعه پیکربندی‌های تغییرات ساختاری ترکیبی از جمله local n-jumps و chromoplexy را با 45 کانال ترکیب کردند (Li, Roberts et al. 2020). دو تا از رایج‌ترین کلاس‌های تغییرات ساختاری، حذف‌ها و TDها، بیشتر بر اساس اندازه، دومین‌های زمان‌بندی همانندسازی[۶۷] و وقوع در سایت‌های شکننده[۶۸] تقسیم شدند. بنابراین سیستم طبقه بندی اولیه بسیار پیچیده بود.

سه امضا با حذف‌های کوچک، متوسط و بزرگ از آنالیز پدیدار شدند (Li, Roberts et al. 2020). یک امضای حذف کوچک که عمدتاً شامل حذف‌های کمتر از 10 کیلوباز و وارونگی‌های متقابل[۶۹] کمتر از 100 کیلوباز است، شبیه حذف‌های RS5 است که در سرطان‌های جهش یافته با BRCA1 یا جهش یافته با BRCA2 مشاهده می‌شود (Nik-Zainal, Davies et al. 2016). یک امضای حذف بزرگ (10 کیلوباز تا 3 مگاباز) یادآور شکل پیچیده‌ای از RS2 سینه بود، در حالی که مکانیسم امضای حذف در اندازه متوسط ناشناخته بود (Li, Roberts et al. 2020). نویسندگان مطالعه فرض کردند که فعالیت تغییر الگو[۷۰] ممکن است امضاهای پیچیده آن‌ها را توضیح دهد، اگرچه این در انتظار تأیید خارجی است.

---

[64] hierarchical Dirichlet process
[65] Pan-Cancer Analysis of Whole Genomes (PCAWG)
[66] primary cancers
[67] replication timing domains
[68] fragile sites
[69] reciprocal inversions
[70] template switching activity

پنج امضای TD شناسایی شدند که بر اساس اندازه و زمان همانندسازی متمایز شدند (Li, Roberts et al. 2020). هر دو امضای TD کوچک با همانندسازی زود هنگام[۷۱] و همانندسازی دیر هنگام[۷۲] (کمتر از 55 کیلوباز) (Li, Roberts et al. 2020) همانطور که قبلاً گزارش شده بود با غیرفعال سازی[۷۳] BRCA1 مرتبط بودند (Nik-Zainal, Davies et al. 2016, Hillman, Chisholm et al. 2018). امضای TD کوچک زود هنگام (Li, Roberts et al. 2020) میکروهومولوژی را در اتصالات نقطه شکست[۷۴] به نمایش می‌گذارد، و همچنین درج‌های الگویی[۷۵] را نشان می‌دهد که گمان می‌رود ردپای فعالیت اتصال انتهایی با واسطه (POLQ) DNA پلیمراز-θ باشد (Ceccaldi, Liu et al. 2015, Mateos-Gomez, Gong et al. 2015, Kamp, Van Schendel et al. 2020). یک امضای کوچک TD همانندسازی با تاخیر به عنوان غنی شده[۷۶] با جهش‌های ژن FANC گزارش شد (Li, Roberts et al. 2020). به طور کلی، امضاهای کوچک TD اغلب با از دست دادن BRCA1 در سرطان سینه و تخمدان همراه است، اگرچه این ارتباط در سرطان‌های کبد، ریه و دهانه رحم مشاهده نمی‌شود (Bayard, Meunier et al. 2018, Li, Roberts et al. 2020). فرآیندهای جهش‌زا متمایز می‌توانند به طور قابل قبولی بر روی فنوتیپ‌های[۷۷] TD مشابه در بافت‌های مختلف همگرا شوند[۷۸].

ما هنوز در مراحل اولیه درک نحوه طبقه‌بندی ایندل‌ها و بازآرایی‌ها هستیم. کانال‌های چندگانه[۷۹] کنونی مورد استفاده در استخراج ID و RS امکان مقایسه بین مطالعات را فراهم نمی‌کنند زیبایی چارچوب امضای جهشی فقط در الگوریتم‌های ریاضی آن نهفته نیست زیرا این الگوریتم‌ها اغلب به سادگی رویکردهای تجزیه ماتریسی هستند، بلکه به نحوه طبقه‌بندی جهش‌ها قبل از فاکتورسازی[۸۰] هم مربوط است. تعداد بیش از حد کانال‌های فاقد اطلاعات مهم قدرت تشخیص امضا را کاهش می‌دهد (**شکل 1**). برعکس، کانال‌هایی که خیلی کم هستند ممکن است احتمال تشخیص بیولوژیکی جدید را کاهش دهند. از طرفی کانال‌هایی که بیش از حد پیچیده هستند، قابلیت استفاده را کاهش می‌دهند و احتمالاً به امضاهای مختلط[۸۱] منجر می‌شوند و تفسیر را به چالشی غیرضروری تبدیل

[71] early-replicating
[72] late-replicating
[73] inactivation
[74] breakpoint junctions
[75] templated insertions
[76] enriched
[77] phenotypes
[78] converge
[79] multifarious channels
[80] decomposition
[81] mixed signatures

می‌کنند. بعلاوه، مرتبه‌های بزرگی کمتر ایندل‌ها و بازآرایی‌ها نسبت به جایگزینی‌های تک باز وجود دارد. بنابراین، سایر مسائل بالقوه مرتبط با قدرت ممکن است هنوز خود را نمایش دهند.

1.1.4 امضای تعداد کپی[82]

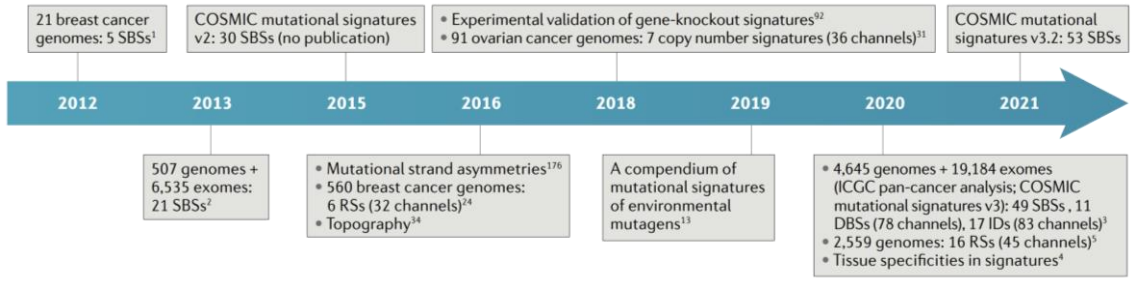
فرآیندهای جهش گسسته[۸۳] می‌تواند منجر به سود[۸۴] و زیان[۸۵] DNA شود (یعنی تغییرات تعداد کپی در سرطان‌ها). تا به امروز، تعداد کمی امضای تعداد کپی در آنالیزهای محلی سرطان تخمدان، پروستات و بافت نرم با استفاده از روش‌های مختلف گزارش شده است (Macintyre, Goranova et al. 2018, Steele, Tarabichi et al. 2019, Wang, Li et al. 2021). طبقه‌بندی ویژگی‌های تعداد کپی قبل از استخراج، از ویژگی‌های مبتنی بر توزیع استفاده می‌کرد و پیچیده و خاص هرگروهی بود (Macintyre, Goranova et al. 2018, Wang, Li et al. 2021). تعداد کپی را می‌توان از توالی‌یابی کم عمق[۸۶] کم گذر[۸۷] یا داده‌های ریزآرایه[۸۸] استنباط کرد و ممکن است روشی ارزان‌تر برای طبقه‌بندی تومور و پیش‌بینی نتیجه بیماری باشد. با این حال، امضاهای تعداد کپی وضوح محدودی دارند، زیرا تغییرات ژنومی را در مقیاس کروموزومی و نه در مقیاس نوکلئوتیدی گزارش می‌کنند، و بنابراین دقت ارائه شده توسط جایگزینی وفنوتیپ‌های ایندل را نخواهند داشت.

1.2 مروری بر پیشینه پژوهش

مفهوم امضاهای جهشی در سال 2012 به دنبال اثبات این موضوع معرفی شد که تجزیه و تحلیل همه جهش‌های جایگزینی[۸۹] در مجموعه‌ای از توالی کامل ژنومی (WGS) سرطان سینه می‌تواند الگوهای ثابت جهش‌زایی در سراسر تومورها که در طول تومورزایی[۹۰] به وجود می آیند را نشان دهد (Nik-Zainal, Alexandrov et al. 2012)

[82] copy number
[83] discrete mutational processes
[84] gains
[85] losses
[86] shallow
[87] low-pass
[88] microarray data
[89] substitution mutations
[90] tumorigenesis

(**شکل 2**). این الگوها نشان‌های فیزیولوژیکی[۹۱] آسیب DNA و فرآیندهای ترمیم بودند که در طول تومورزایی[۹۲] رخ داده بود و می‌توانست تومورهای BRCA1-null و BRCA2-null را از سرطان‌های پراکنده سینه[۹۳] متمایز کند. متعاقباً، یک مطالعه برجسته این اصل را بر روی حدود WGS 500 و حدود 6500 توالی کامل-اگزومی تومورها[۹۴] در 30 نوع سرطان اعمال کرد و 21 امضاهای جهشی جایگزینی تک بازی[۹۵] (SBSs) را نشان داد (Alexandrov, Nik-Zainal et al. 2013). اخیرا، تجزیه و تحلیل به روز شده حدود WGS 4600 و حدود 19000 نمونه توالی کامل- اگزومی تعداد SBS های شناخته شده را به 49 افزایش داد (Alexandrov, Kim et al. 2020). پیچیدگی بیشتر، از جمله ویژگی‌های بافتی احتمالی برای برخی از امضاهای جهشی، نیز نشان داده شده است (Degasperi, Amarante et al. 2020). **شکل 2** مطالعات تجربی اثبات مفهوم کلیدی با تمرکز بر نمونه‌های انسانی را نیز نشان داده است (Nik-Zainal, Alexandrov et al. 2012, Haradhvala, Polak et al. 2016, Morganella, Alexandrov et al. 2016, Nik-Zainal, Davies et al. 2016, Macintyre, Goranova et al. 2018, Zou, Owusu et al. 2018, Kucab, Zou et al. 2019, Li, Roberts et al. 2020).



شکل 2. گزارش زمانی از چگونگی تکامل مفاهیم در زمینه امضاهای جهشی در طول زمان

اگرچه امضاهای جهشی یک مفهوم نسبتاً جدید در بیولوژی سرطان هستند، اولین توصیفات انحرافات ژنومی[۹۶] ناشی از یک فرآیند خاص به اوایل قرن بیستم باز می‌گردد، زمانی که اشعه ایکس برای ایجاد شکستگی

[91] physiological imprints
[92] tumorigenesis
[93] sporadic breast cancers
[94] whole-exome-sequenced tumours
[95] single-base substitution mutational signatures
[96] genomic aberrations

کروموزوم در سلول‌های تحت تابش یافت شد (MULLER 1932, Bauer, Demerec et al. 1938, Sax 1938). الگوهای جهش دقیق‌تری در دهه 1960 گزارش شد، به ویژه اتصال عرضی[۹۷] بازهای پیریمیدین مجاور[۹۸] (CC، CT، TC، TT) که به دلیل اشعه ماوراء بنفش، تبدیل سیتوزین[۹۹] به تیمین[۱۰۰] ($C > T$) و سیتوزین-سیتوزین به تیمین- تیمین ($CC > TT$) را در سایت‌های دی پیریمیدین[۱۰۱] ایجاد می‌کند (Howard and Tessman 1964, Setlow and Carrier 1966, Pfeifer, You et al. 2005). سایر پیوندهای علّی بین عوامل جهش‌زا و الگوهای تغییرات سوماتیکی نیز ایجاد شده است، مانند جابجایی‌های گوانین[۱۰۲] به تیمین ($G > T$) ناشی از ترکیب‌های اضافی گوانین که توسط مواد سرطان‌زا[۱۰۳] موجود در دود تنباکو ایجاد می‌شوند (Pfeifer, Denissenko et al. 2002, Govindan, Ding et al. 2012). علاوه بر این، برخی از عوامل شیمی درمانی[۱۰۴] نیز جهش‌زا هستند و ممکن است امضای جهشی خود را در ژنوم سرطان بیماران مبتلا به بدخیمی‌های ثانویه[۱۰۵] نقش کنند (Hunter, Smith et al. 2006, Harris 2013). این مثال‌ها اهمیت مطالعه الگوهای جهش پیکری را برای درک ما از مکانیسم‌های مولکولی نئوپلازی[۱۰۶] نشان می‌دهند، که به طور بالقوه امکان کشف جهش‌زاهای جدید را فراهم می‌کند (Alexandrov, Nik-Zainal et al. 2013, Alexandrov and Stratton 2014, Helleday, Eshtad et al. 2014, Roberts and Gordenin 2014).

الگوهای جهش‌های متعدد در ژنوم‌های سرطانی معمولاً روی یکدیگر قرار می‌گیرند و داده‌ها را غیرقابل درک می‌کنند. در سال L. Alexandrov 2012، راهی برای حل ریاضی این مسئله ارائه کرد و نشان داد که الگوهای جهش از جهش‌زاهای فردی یافت شده در یک تومور را می‌توان با استفاده از یک رویکرد ریاضی به نام جداسازی منبع کور[۱۰۷] از یکدیگر متمایز کرد (Nik-Zainal, Alexandrov et al. 2012). در سال 2013، تیم وی اولین چارچوب محاسباتی را برای رمزگشایی امضاهای جهشی از داده‌های ژنومیک سرطان منتشر کردند

[97] crosslinking
[98] adjacent pyrimidine bases
[99] cytosine
[100] thymine
[101] dipyrimidine sites
[102] guanine
[103] carcinogens
[104] chemotherapeutic
[105] secondary malignancies
[106] neoplasia
[107] blind source separation

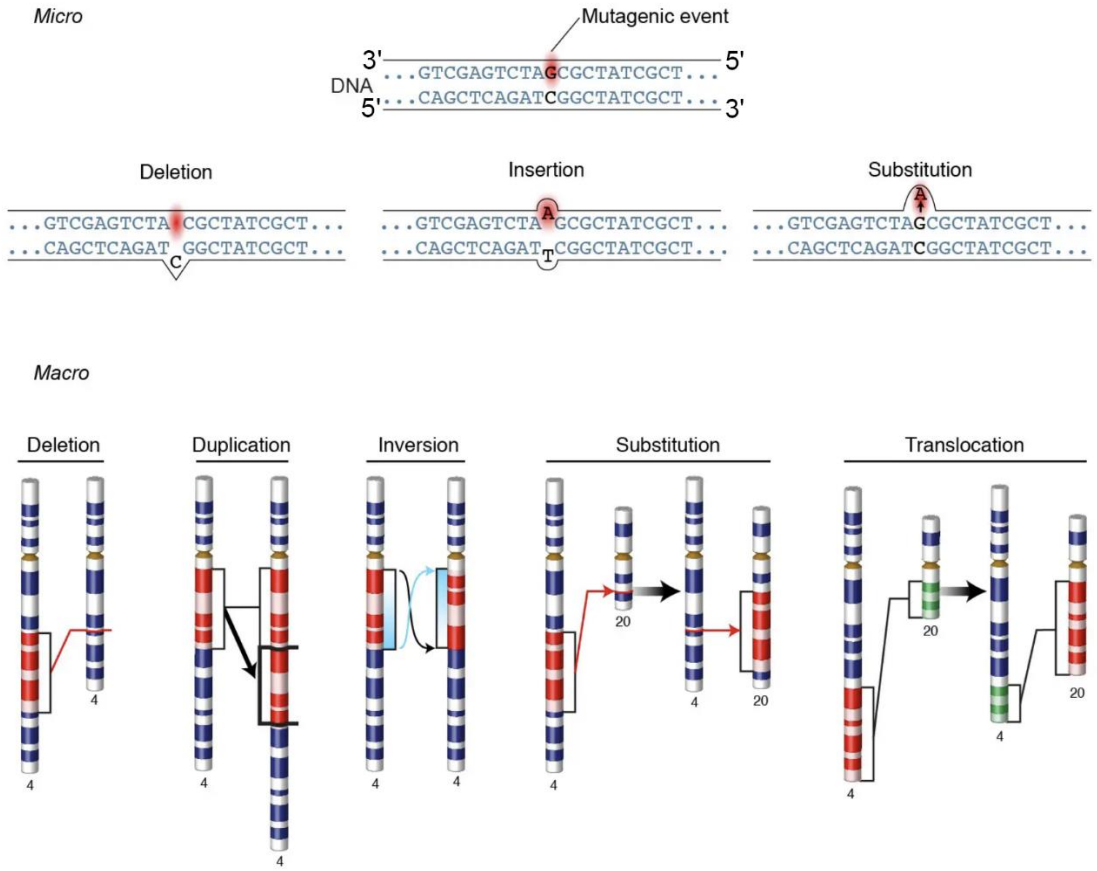
(Alexandrov, Nik-Zainal et al. 2013). متعاقباً، آن‌ها این چارچوب را برای بیش از هفت هزار ژنوم سرطانی به کار بردند و اولین نقشه جامع از امضاهای جهشی در سرطان انسان را ایجاد کردند (Alexandrov, Nik-Zainal et al. 2013). در حال حاضر، بیش از صد امضای جهش یافته در فهرست سرطان انسان شناسایی شده است (Alexandrov, Kim et al. 2020, Degasperi, Amarante et al. 2020, Degasperi, Zou et al. 2022, Islam, Díaz-Gay et al. 2022, Ledford 2022).

1.3 بیان مسئله

امضای جهشی را می‌توان از نظر ریاضی به عنوان رابطه‌ای بین یک فرآیند جهش‌زا (معلوم یا ناشناخته) و مجموعه‌ای از انواع جهش پیکری تعریف کرد. بسیاری از دسته‌های تغییرات ژنومی **(شکل 3)** می‌توانند به‌عنوان ویژگی‌های[۱۰۸] یک امضای جهشی عمل کنند، از جمله جایگزین‌های تک بازی[۱۰۹] (SBSs) یا دوگانه (DBS)[۱۱۰]، درج‌ها و حذف‌های کوچک (ایندل‌ها)[۱۱۱]، تغییرات تعداد کپی[۱۱۲]، بازآرایی‌های ساختاری[۱۱۳]، رویدادهای ادغام عناصر قابل انتقال[۱۱۴]، هایپرجهش موضعی (کاتائگیس)[۱۱۵] و تغییرات اپی ژنتیکی[۱۱۶]. با توجه به اینکه بیشتر مطالعات تا به امروز بر روی جایگزینی‌های تک بازی متمرکز شده است، در عمل تنها تعداد محدودی از ویژگی‌ها را می‌توان در انتزاع ریاضی[۱۱۷] یک امضای جهشی گنجاند. با این حال، امضاهای مبتنی بر ایندل (Morganella, Alexandrov et al. 2016, Nik-Zainal, Davies et al. 2016) یا انواع ساختاری[۱۱۸] (Morganella, Alexandrov et al. 2016, Nik-Zainal, Davies et al. 2016, Secrier, Li et al. 2016) و تغییرات تعداد کپی (Díaz-Gay, Vangara et al. 2023) نیز توصیف شده است. اگرچه ما هنوز در مراحل اولیه درک نحوه طبقه‌بندی ایندل‌ها، بازآرایی‌ها و تعداد کپی‌ها هستیم و مدل‌سازی دقیق آن‌ها

[108] features
[109] single-base substitutions
[110] single-nucleotide or dinucleotide substitutions
[111] small insertions and deletions (indels)
[112] copy number changes
[113] structural rearrangements
[114] transposable element integration events
[115] localized hypermutation (kataegis)
[116] epigenetic changes
[117] mathematical abstraction
[118] structural variants

چالش برانگیزتر است. بنابراین آن‌ها هنوز به طور گسترده مورد استفاده قرار نگرفته‌اند. علاوه بر این، برخی از امضاهای جایگزین به طور مداوم با ویژگی‌هایی مانند افزایش تعداد ایندل‌ها یا بازآرایی یک کلاس خاص، رویدادهای کاتائگیس، یا سوگیری‌ها[۱۱۹] در رشته رونویسی[۱۲۰] که در آن جهش‌ها رخ می‌دهد، مرتبط هستند (Nik-Zainal, Alexandrov et al. 2012, Alexandrov, Nik-Zainal et al. 2013, Alexandrov, Jones et al. 2015, Schulze, Imbeaud et al. 2015, Nik-Zainal, Davies et al. 2016). بنابراین، در نظر گرفتن چنین ویژگی‌هایی به عنوان محدودیت‌های بیولوژیکی برای شناسایی امضاها مفید است، حتی اگر مدل‌سازی دقیق آن‌ها چالش برانگیزتر باشد.



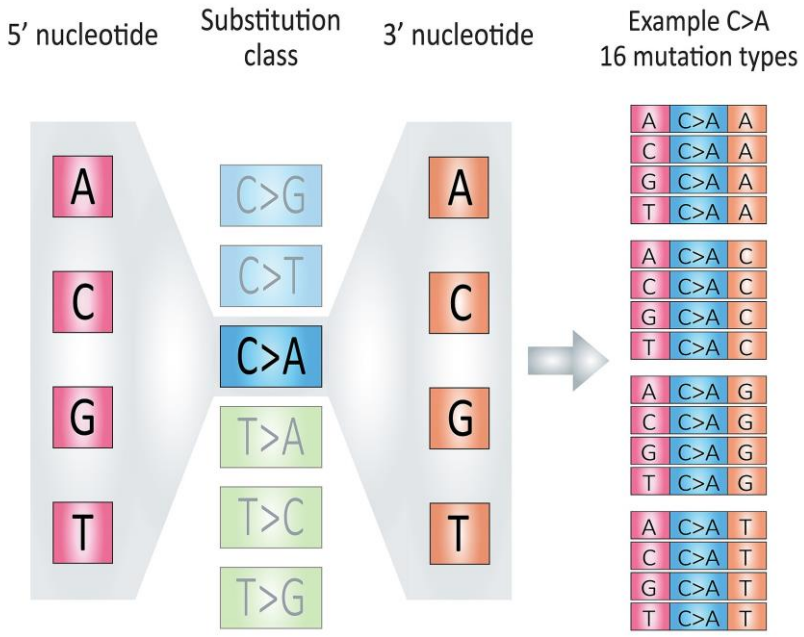
شکل 3. دسته‌های کلی تغییرات ژنومی و انواع مختلف جهش‌ها

[119] biases

[120] transcriptional strand

یکی از معروف‌ترین روش‌های مدلسازی جایگزین‌های تک بازی به این صورت است که، شش کلاس جایگزینی پایه وجود دارد: C>A، C>G، C>T، T>A، T>C، T>G. جایگزینی G>T معادل جایگزینی C>A در نظر گرفته می‌شود زیرا تشخیص اینکه در ابتدا جایگزینی در کدام رشته DNA (رو به جلو[۱۲۱] یا معکوس[۱۲۲]) رخ داده است، امکان‌پذیر نیست. بنابراین هر دو جایگزین C>A و G>T به عنوان بخشی از کلاس "C>A" محاسبه می‌شوند. به دلیل مشابه جهش‌های G>C، G>A، A>T، A>G و A>C به ترتیب به عنوان بخشی از کلاس‌های "C>G"، "C>T"، "T>A"، "T>C" و "T>G" محاسبه می‌شوند.

گرفتن اطلاعات از جفت بازهای مجاور '5 و '3 منجر به ایجاد 96 نوع جهش ممکن می‌شود (مانند A[C>A]A، A[C>A]T، و غیره) (**شکل 4**). کاتالوگ جهش یک تومور با دسته‌بندی هر جهش نوکلئوتیدی منفرد[۱۲۳] (SNV) در یکی از 96 نوع جهش و شمارش تعداد کل جایگزینی‌ها برای هر یک از این 96 نوع جهش ایجاد می‌شود.



شکل 4. دسته‌بندی 96 نوع جهش جایگزینی تک بازی

[121] forward

[122] reverse

[123] single nucleotide variant (SNV)

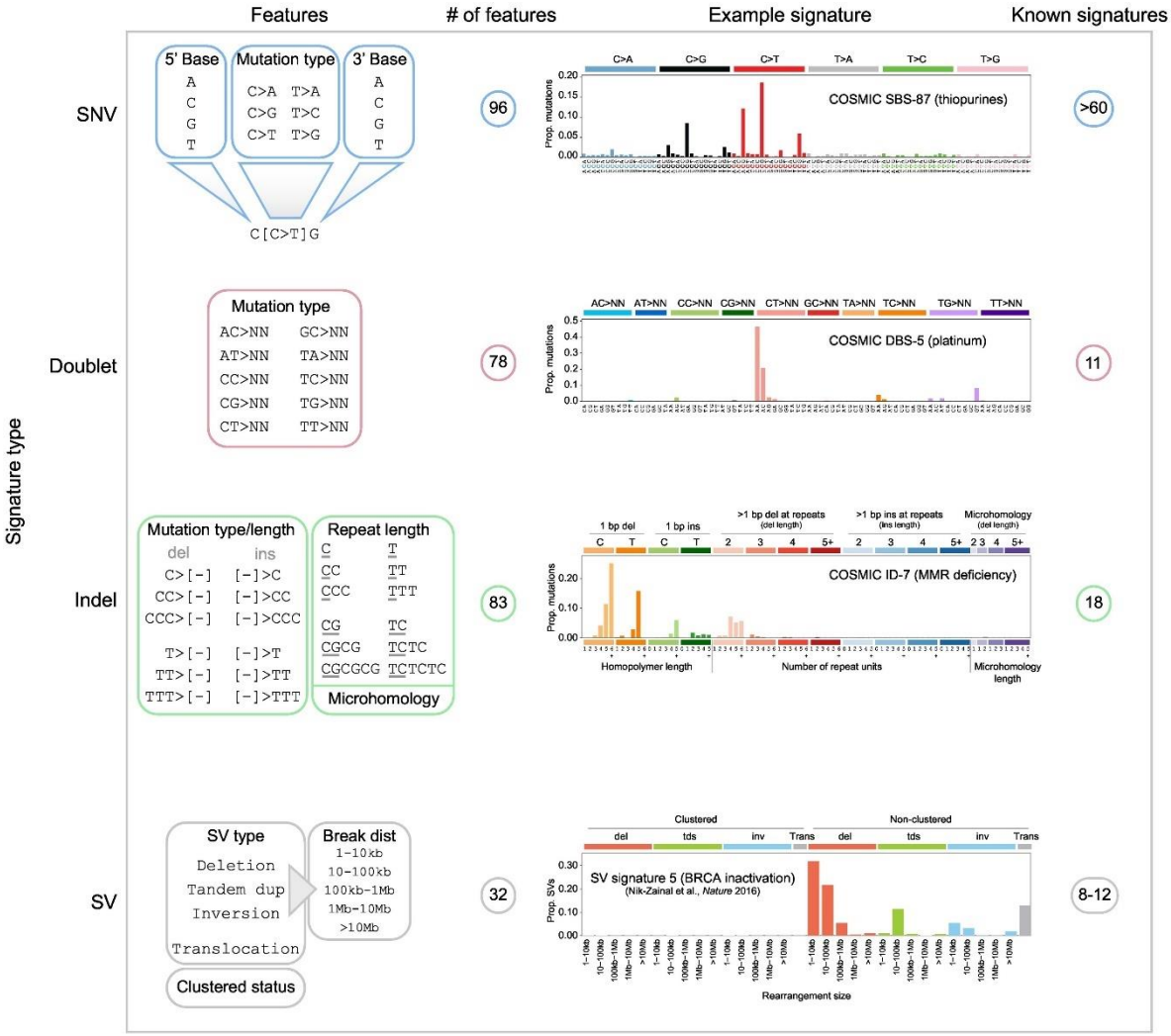
در **شکل 4** با در نظر گرفتن باز مجاور '5 (A, C, G, T)، 6 کلاس جایگزینی (C>A, C>G, C>T, T>A, T>C, T>G) و باز مجاور '3 (A، C، G، T) دسته‌بندی 96 (4 x 6 x 4 = 96) نوع جهش ایجاد می‌گردد که 16 نوع جهش کلاس جایگزینی C>A به عنوان مثال نشان داده شده است.

راه‌های علمی جدید برای شناسایی و آنالیز انحرافات ژنومی، از جمله استخراج امضاهای جهش از مجموعه‌ای از جهش‌های پیکری، بررسی شده است. این کاتالوگ‌هایی[۱۲۴] از امضاها را تولید کرده است که در انواع نئوپلازی‌های انسانی عمل می‌کنند (Alexandrov, Nik-Zainal et al. 2013, Alexandrov, Jones et al. 2015, Schulze, Imbeaud et al. 2015, Nik-Zainal, Davies et al. 2016).. هنگامی که کاتالوگ یا ماتریس جهشی (به عنوان مثال تعداد 96 نوع جهش برای جایگزینی‌های تک بازی) یک تومور به دست آید، دو رویکرد برای رمزگشایی مشارکت امضاهای جهشی مختلف در چشم انداز ژنومی تومور دنبال می‌شود. ابتدا کاتالوگ جهشی تومور با کاتالوگ جهشی مرجع یا مجموعه داده مرجع امضاهای جهشی مقایسه می‌گردد. سپس مدل‌سازی امضاهای جهشی می‌تواند با استفاده از روش‌های تجزیه محور مانند فاکتورسازی ماتریس غیرمنفی[۱۲۵] (NMF) برای شناسایی فرآیندهای جهش جدید بالقوه صورت گیرد (Alexandrov, Nik-Zainal et al. 2013). شناسایی سهم امضاهای جهشی متنوع در سرطان‌زایی بینشی در مورد بیولوژی تومور فراهم می‌کند و می‌تواند فرصت‌هایی را برای درمان هدفمند ارائه دهد.

تا سال 2021 و در نسخه 3.2 دیتابیس COSMIC[۱۲۶]، بیش از 60 امضای جهشی SBS در سراسر سرطان‌ها بر اساس 96 نوع ممکن که زمینه سه نوکلئوتیدی SNV را در نظر می‌گیرند، شناسایی و گنجانده شده است (Alexandrov, Kim et al. 2020). 11 DBS و 18 امضای ایندل نیز شناسایی شده‌اند (Alexandrov, Kim et al. 2020). در نهایت، 8-12 امضاهای SV بر اساس الگوهای شماره کپی مجاور[۱۲۷]، جهت گیری نقطه شکست[۱۲۸]، وجود SV های خوشه‌ای، و زمینه توالی نزدیک[۱۲۹] شناسایی شده‌اند (Papaemmanuil, Rapado et al. 2014, Nik-Zainal, Davies et al. 2016, Davies,

[124] catalogues
[125] non-negative matrix factorization (NMF)
[126] Catalogue of Somatic Mutations in Cancer (COSMIC) database
[127] adjacent copy number patterns
[128] breakpoint orientation
[129] nearby sequence context

(Glodzik et al. 2017, Hoang, Cornish et al. 2019) (**شکل5**). که البته در این چند سال اخیر این تعداد به طور منظم به روز شده است.



شکل 5. ویژگی‌های انواع تک نوکلئوتیدی (SNV یا SBS)، دوگانه (DBS)، درج-حذف (indel)، و تغییرات ساختاری (SV)

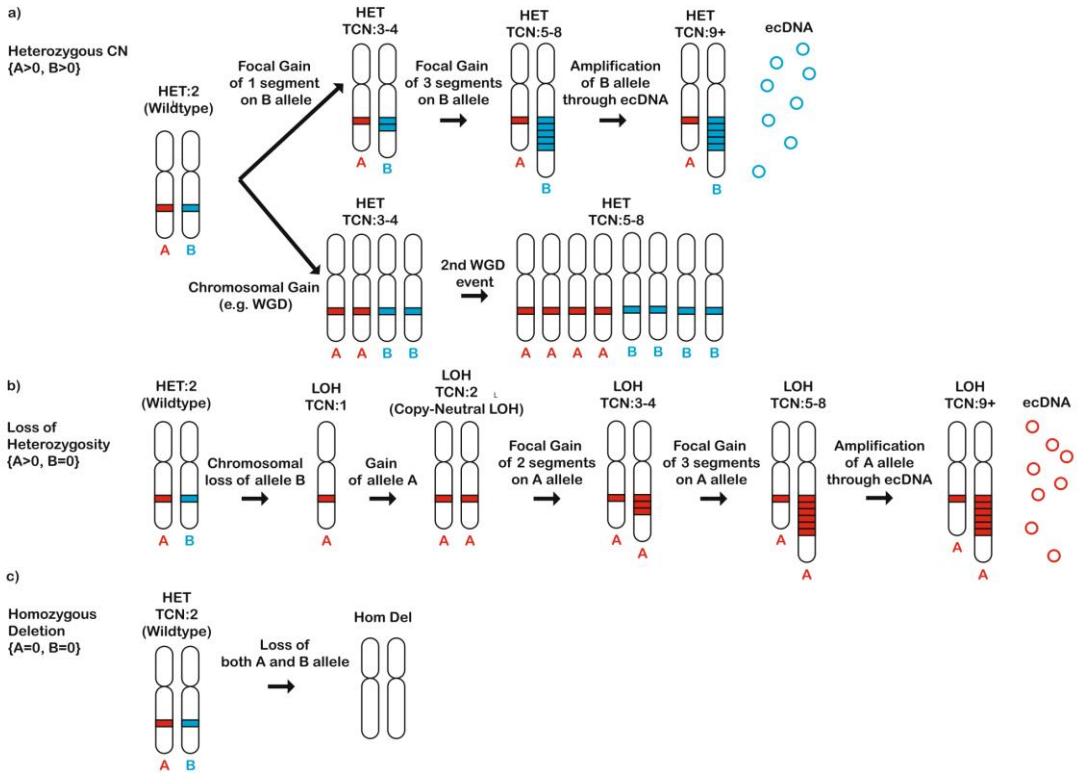
در **شکل 5** ستون اول ویژگی‌های ژنومی هر نوع امضا را نشان می‌دهد. برای SNV ها، زمینه سه نوکلئوتیدی، شامل بازهای'5 و '3 محل جهش یافته، همراه با نوع واریانت (به عنوان مثال، C>T) در نظر گرفته می‌شود. برای دوگانه‌ها، فقط آلل‌های مرجع و تغییر یافته در نظر گرفته می‌شوند (N نشان دهنده هر بازی است). برای

ایندل‌ها، نوع و طول جهش در نظر گرفته می‌شود، جایی که «[−]» دنباله‌های حذف شده (del) یا درج شده (in) را نشان می‌دهد. طول تکرارهای موضعی نیز در نظر گرفته می‌شود (به عنوان مثال، اگر GCCCG به GCCG تبدیل شود، 3 برای 3 C خواهیم داشت.) و همچنین اینکه آیا میکروهومولوژی[۱۳۰] در مجاورت ناحیه حذف شده وجود دارد یا خیر. SV ها به چهار نوع نشان داده شده، کلاسه‌بندی می‌شوند و SV های داخل کروموزومی (تکثیرات پشت سر هم[۱۳۱]، حذف‌ها، و وارونگی‌ها[۱۳۲]) بر اساس فاصله بین دو نقطه شکست (فاصله شکست[۱۳۳]) ساب کلاسه‌بندی می‌شوند. اینکه آیا SV های مشابه در یک منطقه ژنومی خاص خوشه‌بندی شده‌اند نیز در نظر گرفته می‌شود. ستون دوم تعداد ویژگی‌های مورد استفاده برای طبقه‌بندی هر امضا را نشان می‌دهد. به عنوان مثال، امضاهای SNV دارای چهار باز احتمالی $5'$، شش نوع جهش احتمالی، و چهار باز احتمالی $3'$ هستند که 96 ویژگی را ایجاد می‌کند. ستون سوم نمونه‌هایی از امضاها را نشان می‌دهد، با محور y که نسبت جهش‌ها در هر ویژگی را نشان می‌دهد. شناسه امضاهای شناسایی شده COSMIC (در صورت وجود) و علت‌شناسی[۱۳۴] امضا را نشان می‌دهد. ستون چهارم تعداد امضاهای گزارش شده در مطالعات منتشر شده یا COSMIC را نشان می‌دهد. مخفف: MMR، بازسازی نابرابر[۱۳۵].

مطابق **شکل 6** طرح طبقه‌بندی شماره کپی شامل 48 کانال متقابل منحصر به فرد است که بر اساس وضعیت هتروزیگوسیتی[۱۳۶]، اندازه قطعه[۱۳۷] و تعداد کل شماره کپی[۱۳۸] (TCN) تقسیم می‌شود.در **قسمت a شکل** نشان می‌دهد که، در وضعیت هتروزیگوسیتی، هر دو آلل[۱۳۹] حفظ می‌شوند و می‌توان یک یا هر دو آلل را تقویت کرد. این تقویت می‌تواند کانونی[۱۴۰] (پانل بالا) باشد یا می‌تواند یک کروموزوم یا حتی کل ژنوم (پانل پایین) را در برگیرد. دسته هتروزیگوت[۱۴۱] بیشتر بر اساس TCN تقسیم می‌شود ($TCN = 1$، $TCN = 2$، $TCN = 3/4$،

[130] microhomology
[131] tandem duplications
[132] inversions
[133] break dist
[134] etiology
[135] mismatch repair
[136] heterozygosity
[137] segment size
[138] total copy number (TCN)
[139] allel
[140] focal
[141] heterozygous

TCN = 5-8 و TCN >= 9). در **قسمت b شکل** در وضعیت از دست دادن هتروزیگوسیتی[۱۴۲] (LOH) ، یکی از آلل‌ها از بین می‌رود. سپس آلل باقیمانده را می‌توان کپی کرد (یعنی LOH خنثی را کپی کرد)، و تحت تقویت بیشتری قرار گرفته و در نتیجه وضعیت‌های تعداد کل شماره کپی بالاتر می‌رود. به دنبال آن، دسته LOH بر اساس TCN(TCN = 1، TCN = 2، TCN = 3/4، TCN = 5-8 و TCN >= 9) تقسیم می‌شود. دسته‌های هتروزیگوت و LOH بیشتر بر اساس اندازه بخش تقسیم می‌شوند: 0 تا 100 کیلوبایت، 100 کیلوبایت - 1 مگابایت، 1 مگابایت - 10 مگابایت، 10 مگابایت - 40 مگابایت، > 40 مگابایت. تکثیرهای LOH یا هتروزیگوت سطح بالا (به عنوان مثال TCN = 5-8 یا TCN >= 9) را می‌توان روی DNA خارج کروموزومی[۱۴۳] (به صورت دایره‌های قرمز نشان داده شده) و همچنین روی کروموزوم‌های خطی انجام داد. **قسمت c شکل** نشان می‌دهد که حذف‌های هموزیگوت منجر به از دست دادن هر دو آلل می‌شود و بر اساس اندازه بخش حذف شده تقسیم می‌گردد: 0 - 100kb ، 1 Mb - kb و > Mb1



شکل 6. نمای طرح طبقه‌بندی شماره کپی

[142] loss of heterozygosity (LOH)

[143] extrachromosomal DNA

آنالیز امضاهای جهشی یک رویکرد قدرتمند برای درک فرآیندهای جهش‌زایی است که تکامل ژنوم سرطان را شکل داده است. برای ارزیابی امضاهای جهشی فعال در یک ژنوم سرطانی، ابتدا باید فعالیت‌های آن‌ها را با تخمین تعداد جهش‌های حک شده[۱۴۴] توسط هر امضا، کمی‌سازی کرد. مطالعه این امضاهای جهشی دارای پتانسیل مهمی برای درک بیشتر ما از علل سرطان‌های فردی است و می‌تواند بینش‌های جدیدی در رابطه با پیشگیری و درمان سرطان ارائه دهد. امروزه، آنالیزهای امضای جهشی به یک جزء استاندارد مطالعات ژنومی تبدیل شده‌اند، زیرا می‌توانند منابع محیطی[۱۴۵] و درون‌زای جهش‌زایی را در هر تومور نشان دهند. در واقع، این رشته نوپا در جهت‌های محاسباتی، تجربی و بالینی در حال توسعه و گسترش است و به سمت استفاده از روش بالینی معنادار، آگاهی‌رسانی، تلاش‌هایی جهت پیشگیری سرطان، شناسایی پتانسیل‌های سرطانی ناشناخته (Grolleman, De Voer et al. 2019, Georgeson, Pope et al. 2021)، طبقه‌بندی بیماری‌زایی جهش‌های رده زایشی[۱۴۶] (Georgeson, Harrison et al. 2022)، هدایت روش‌های تشخیصی، پیش‌بینی پاسخ به درمان برای مدیریت بالینی بیماران سرطانی (Davies, Glodzik et al. 2017)، شناسایی حساسیت به درمان‌های ضد سرطان (Levatić, Salvadores et al. 2022) و مداخلات شخصی‌سازی شده سرطان[۱۴۷] پیش می‌رود (Harris 2013, Poon, McPherson et al. 2014, Alexandrov, Nik-Zainal et al. 2015, Poon, Huang et al. 2015, Fox, Salk et al. 2016, Li, Wu et al. 2016, Secrier, Li et al. 2016, Levatić, Salvadores et al. 2022).

در حالی که توسعه روش‌هایی برای کشف امضاهای جهشی به موفقیت قابل توجهی دست یافته است و این‌ها روندهای مثبتی هستند، این هنوز یک زمینه نوظهور است که ناشی از پیشرفت‌های تحلیلی و تکنولوژیکی اخیر است، جا دارد که بپرسیم آیا محدودیت‌هایی برای این حوزه‌ای که به طور قابل توجهی در حال گسترش است وجود دارد یا خیر. علیرغم پیشرفت‌ها در این زمینه و در حالی که همانطور که تعداد فزاینده‌ای از امضاهای کلاس‌های مختلف جهشی گزارش می‌شود (Alexandrov, Kim et al. 2020, Degasperi, Amarante et al. 2020, Li, Roberts et al. 2020)، در تلاش برای رمزگشایی علل، همبستگی‌هایی بین آن‌ها و عوامل مختلفی مانند سن و مواجهه با رفلاکس اسید یا درمان‌های دارویی مشخص شده است (Alexandrov, Jones et al. 2015, Secrier, Li et al. 2016, Pich, Muiños et al. 2019). با این حال، منشأ بسیاری از امضاها همچنان مبهم است. علاوه بر این، در حالی که تجزیه و تحلیل‌های قبلی امضاهای منفرد مثل تابش اشعه

[144] imprinted mutations
[145] environmental
[146] germline variants
[147] personalized cancer interventions

ماوراء بنفش (یعنی SBS7) (Alexandrov, Nik-Zainal et al. 2013) را گزارش می‌کردند، مطالعات و تجزیه و تحلیل‌های جدیدتر، نسخه‌های متعدد و متفاوتی از این امضاها را نسبت به موارد قبلی گزارش کردند (یعنی SBS7a، SBS7b، SBS7c و SBS7d) (Alexandrov, Kim et al. 2020)، که جامعه را به این سوال سوق می‌دهد که آیا برخی از یافته‌ها منعکس کننده زیست شناسی هستند یا صرفاً نتایج انتزاعی ریاضی‌اند و به طور کلی این یافته‌ها تا چه میزان دقیق بوده و بدون تأیید تجربی قابل اعتماد و استناد هستند (Koh, Degasperi et al. 2021). بنابراین، تلاش برای تأیید تجربی این نتایج ریاضی انتزاعی ضروری است. کاربران غیرمتخصص امضاهای جهشی به بینش در مورد مسائل عملی و هشدارها در استفاده از چارچوب‌های تجزیه و تحلیل امضا نیاز دارند. برای پزشکان، استفاده از امضاهای جهشی قابل اعتماد برای طبقه‌بندی بالینی بسیار مهم است.

فصل دوم

روش‌ها

2.1 مواد و روش‌ها

مجموعه داده مورد استفاده در این مطالعه، شامل الگوهای SBS از 2700 ژنوم سرطان قبلا شبیه‌سازی شده از پروژه PCAWG [۱۴۸]، مربوط به 300 تومور از 9 نوع سرطان مختلف، از جمله: کارسینوم سلول انتقالی مثانه[۱۴۹]، آدنوکارسینوم مری[۱۵۰]، آدنوکارسینوم سینه[۱۵۱]، کارسینوم سلول سنگفرشی ریه[۱۵۲]، کارسینوم سلول کلیه[۱۵۳]، آدنوکارسینوم تخمدان [۱۵۴]، استئوسارکوم[۱۵۵]، آدنوکارسینوم دهانه رحم[۱۵۶] و آدنوکارسینوم معده[۱۵۷] است که ژنوم سرطان این نمونه‌ها با استفاده از 21 امضای مرجع COSMIC مختلف شبیه‌سازی شده‌اند (Islam, Díaz-Gay et al. 2022, Díaz-Gay, Vangara et al. 2023).

[148] Pan-Cancer Analysis of Whole Genomes
[149] bladder transitional cell carcinoma
[150] esophageal adenocarcinoma
[151] breast adenocarcinoma
[152] lung squamous cell carcinoma
[153] renal cell carcinoma
[154] ovarian adenocarcinoma
[155] osteosarcoma
[156] cervical adenocarcinoma
[157] stomach adenocarcinoma

حداقل دو رویکرد مجزا برای آنالیز امضاهای جهشی وجود دارد. استخراج de novo یک رویکرد یادگیری ماشینی بدون نظارت[۱۵۸] است که امکان شناسایی الگوهای امضاهای جهش‌یافته شناخته شده و ناشناخته قبلی را فراهم می‌کند (Islam, Díaz-Gay et al. 2022). این نوع آنالیز از آنجایی که به گروه‌های بزرگی شامل معمولاً بیش از 100 نمونه نیاز دارد، عمدتاً برای استخراج امضاهای مرجع استفاده می‌شود. در مقابل، بازسازی امضاهای جهشی یک رویکرد بهینه‌سازی عددی است که با تعیین تعداد جهش‌های منتسب به هر عامل امضا در آن نمونه، امکان تخصیص امضاهای شناخته شده (در بیشتر موارد، مرجع) را به یک نمونه جداگانه فراهم می‌کند. در حالی که بازسازی نمی‌تواند فعالیت‌های امضاهای جهشی ناشناخته قبلی را شناسایی یا کمی کند، این رویکرد به طور گسترده در گروه‌های کوچک و برای نمونه‌های بالینی که ارزیابی‌ها تقریباً به طور انحصاری برای یک بیمار سرطانی انجام می‌شود، به کار می‌رود (Rosenthal, McGranahan et al. 2016).

2.1.1 استخراج de novo امضاهای جهشی

استخراج de novo امضاهای جهشی (Alexandrov, Nik-Zainal et al. 2013) یک رویکرد یادگیری ماشینی بدون نظارت است که در آن یک ماتریس، M، که مربوط به جهش‌های پیکری در مجموعه‌ای از نمونه‌های سرطانی تحت یک طبقه‌بندی جهشی (Bergstrom, Huang et al. 2019) است، با حاصلضرب دو ماتریس با رتبه پایین[۱۵۹]، S و A تقریب زده می‌شود. ماتریس S مجموعه‌ای از امضاهای جهشی را منعکس می‌کند، در حالی که ماتریس A شامل فعالیت‌های امضاها می‌شود. یک فعالیت مربوط به تعداد جهش‌های ایجاد شده توسط یک امضا در نمونه سرطان است (**شکل 7**).

برای به حداکثر رساندن تأثیر تجزیه و تحلیل امضای جهشی، درک دقیق ریاضی و الگوریتم‌های قوی برای افزایش دقت و قابلیت تفسیر آن مورد نیاز است. این زمینه در دهه گذشته شاهد توسعه سریع تکنیک‌های محاسباتی بوده است. ابزارهای محبوبی مانند MuSiCal (Jin, Gulhan et al. 2024)، SigProfilerExtractor (Islam, Díaz-Gay et al. 2022)، SignatureAnalyzer (Kasar, Kim et al. 2015, Taylor-Weiner, Aguet et al. 2019, Alexandrov, Kim et al. 2020)، signature.tools.lib (Degasperi, Amarante et al. 2020, Degasperi, Zou et al. 2022) و سایرین به موفقیت قابل توجهی دست یافته‌اند (Omichessan, Severi et al. 2019). چارچوب‌های محاسباتی برای استخراج de novo امضاهای جهشی توسعه یافته تا به امروز در **جدول 1** نشان داده شده است.

[158] unsupervised

[159] low-rank matrices

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
Emu (Fischer, Illingworth et al. 2013)	matrix	C++	EM	original implementation (Fischer, Illingworth et al. 2013)	no	yes	yes ❊	BIC (Schwarz 1978)	SBS-96	no	no
Maftools (Mayakonda, Lin et al. 2018)	matrix, MAF	R-Bioconductor	NMF	NMF R package (Gaujoux and Seoighe 2010)	no	yes	no	–	SBS-96	SBS-96	1 to 1
MutationalPatterns (Blokzijl, Janssen et al. 2018)	matrix, VCF	R-Bioconductor	NMF	NMF R package (Gaujoux and Seoighe 2010)	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
✦ MuSiCal (Jin, Gulhan et al. 2024)	matrix	Python	mvNMF	original implementation	yes	yes	yes ❊	sparse NNLS	new mutational catalog (Jin, Gulhan et al. 2024)	new catalog (Jin, Gulhan et al. 2024)	1 to many
MutSignatures (Fantini, Vidimar et al. 2020)	matrix, VCF, MAF	R	NMF	Brunet et al. (Brunet, Tamayo et al. 2004)	no	no	no	–	SBS-96	SBS-96	1 to 1
MutSpec (Ardin, Cahais et al. 2016)	matrix, VCF, custom	Galaxy, Perl, R	NMF	NMF R package (Gaujoux and Seoighe 2010)	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
SigFit (Gori and Baez-Ortega 2018)	matrix	R	Bayesian inference	Stan R package (Carpenter, Gelman et al. 2017)	no	yes	yes ❊	Elbow method (Thorndike 1953)	SBS-96	SBS-96, SBS-192	1 to 1
SigMiner (Wang, Li et al. 2021)	matrix, MAF	R	(automatic) Bayesian NMF, (manual) NMF	(automatic) SignatureAnalyzer implementation, (Kasar, Kim et al. 2015) (manual) NMF R package (Gaujoux and Seoighe 2010)	no	yes ❊	yes	ARD (Tan and Févotte 2012)	SBS-96, DBS-78, ID-83	generic	1 to 1
SignatureAnalyzer (Kasar, Kim et al. 2015, Taylor-Weiner, Aguet et al. 2019)	matrix, MAF	R (CPU),)Dempster, Laird et al. 1977(Python (GPU) (Suri and Roy 2017)	Bayesian NMF	original implementation (Kasar, Kim et al. 2015, Taylor-Weiner, Aguet et al. 2019)	yes	no	yes	ARD (Tan and Févotte 2012)	SBS-96, DBS-78, ID-83	SBS-96, DBS-78, ID-83	1 to 1
SignatureToolsLib (Degasperi, Amarante et al. 2020)	matrix, VCF, custom	R	NMF	NMF R package (Gaujoux and Seoighe 2010)	no	yes	no	–	SBS-96, DBS-78, ID-83, SV-32	SBS-96, SV-32, generic	1 to 2
SigneR (Rosales, Drummond et al. 2017)	matrix, VCF	R-Bioconductor, C++	Bayesian NMF	original implementation (Rosales, Drummond et al. 2017)	no	yes	yes ❊	BIC (Schwarz 1978)	SBS-96	SBS-96	no
✦ SigProfilerExtractor (Islam, Díaz-Gay et al. 2022)	matrix, VCF, MAF, custom	Python, R wrapper	NMF	original implementation	yes	yes	yes ❊	NMFk (Nebgen, Vangara et al. 2021)	SBS-96, DBS-78, ID-83, CN-48, others,(Bergstrom, Huang et al. 2019) any	SBS-96, DBS-78, ID-83, CN-48, SV-32, others,(Bergstrom, Huang et al. 2019) generic	1 to many
SigProfiler_PCAWG (Alexandrov, Kim et al. 2020)	matrix, VCF, MAF, custom	Python, MATLAB	NMF	Brunet et al. (Brunet, Tamayo et al. 2004)	no	yes	no	–	SBS-96, DBS-78, ID-83, others,(Bergstro	SBS-96, DBS-78, ID-83	no

									m, Huang et al. 2019) any		
SomaticSignatures (Gehring, Fischer et al. 2015)	matrix, VCF	R-Bioconductor	NMF, PCA	NMF R package (Gaujoux and Seoighe 2010) pcaMethods R package (Stacklies, Redestig et al. 2007)	no	yes	no	–	SBS-96	SBS-96	no
TensorSignatures (Vöhringer, Hoeck et al. 2021)	VCF	Python	NTF	TensorFlow (Abadi, Agarwal et al. 2016)	yes	yes	yes❊	BIC (Schwarz 1978)	tensor	SBS-96 with strand bias	no

جدول 1. مروری بر ابزارهای بیوانفورماتیک توسعه داده شده جهت استخراج de novo امضاهای جهشی.

MAF، فرمت حاشیه‌نویسی جهش[۱۶۰]؛ VCF، فرمت فراخوانی جهش[۱۶۱]؛ EM، الگوریتم به حداکثر رساندن انتظار[۱۶۲]. NMF، فاکتورسازی ماتریس غیر منفی. PCA، تجزیه و تحلیل اجزای اصلی[۱۶۳]. NTF، فاکتورسازی تانسور غیر منفی[۱۶۴]؛ ARD، تعیین ارتباط خودکار[۱۶۵]؛ BIC، معیار اطلاعات بیزی[۱۶۶]. COSMIC، کاتالوگ جهش های جسمی در سرطان. SBS، جایگزین‌های تک بازی. DBS، جایگزین‌های بازی دوگانه؛ ID ، درج‌ها و حذف‌های کوچک؛ CN، شماره کپی؛ SV، جهش‌های ساختاری.

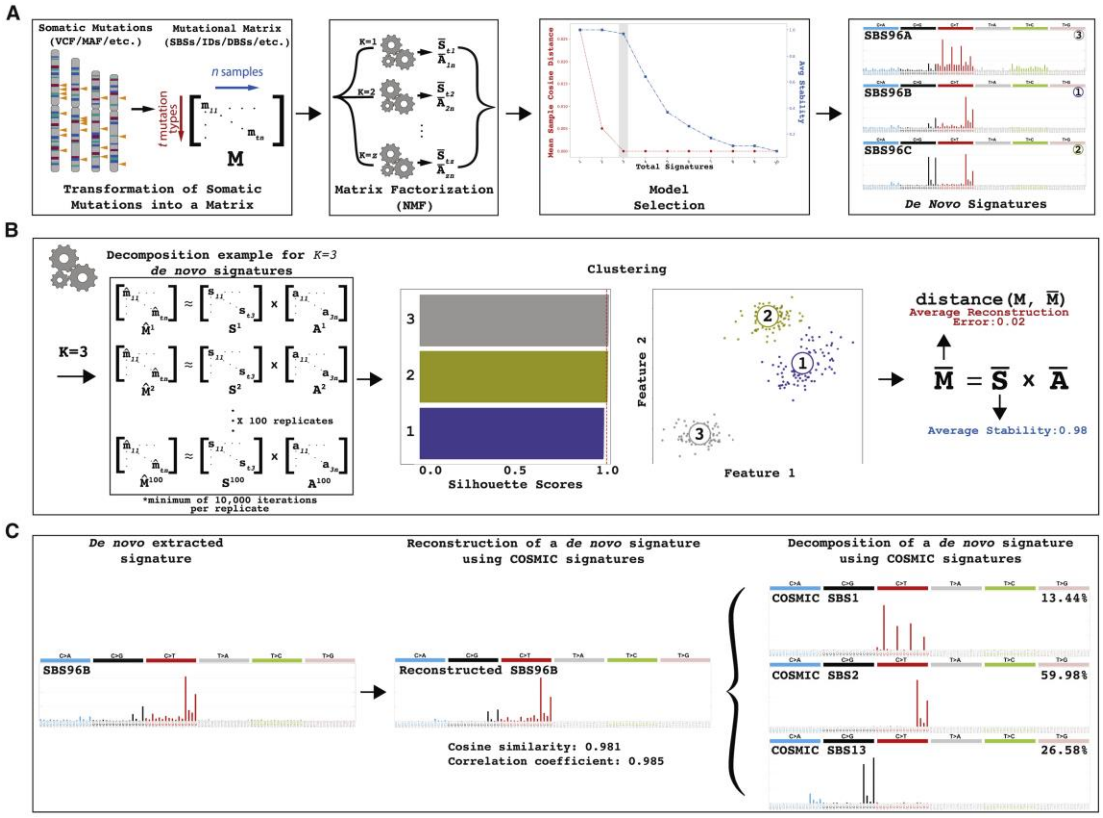
❊ رویکرد پیش‌فرض برای انتخاب تعداد کل امضاها وقتی ابزاری از انتخاب دستی و خودکار پشتیبانی می‌کند.

در **جدول 1** مروری بر ابزارهای بیوانفورماتیک توسعه داده شده جهت استخراج de novo امضاهای جهشی بررسی می‌گردد. ابزارها بر اساس حروف الفبا مرتب شده اند. 1 تا 1 به یک امضای de novo اشاره دارد که دقیقاً با یک امضای COSMIC مطابقت دارد. 1 تا 2 به یک امضای de novo اشاره دارد که با ترکیبی از حداکثر دو امضای COSMIC مطابقت دارد. 1to many به یک امضای de novo اشاره دارد که با ترکیبی از یک یا چند امضای COSMIC مطابقت دارد.

نمای کلی کار SigProfilerExtractor در **شکل 7 (A)** نشان داده شده است. کار SigProfilerExtractor از ورودی جهش‌های پیکری شروع می‌شود و منجر به خروجی امضاهای جهش de novo می‌گردد. **شکل 7** یک

[160] mutation annotation format
[161] variant call format
[162] expectation maximization algorithm
[163] principal component analysis
[164] nonnegative tensor factorization
[165] automatic relevance determination
[166] Bayesian information criterion

مثال برای راه حلی با سه امضای de novo را نشان داده است. جهش‌های پیکری ابتدا به یک ماتریس جهشی M تبدیل می‌شوند. پس از آن، ماتریس با رتبه‌های مختلف با استفاده از فاکتورسازی ماتریس غیرمنفی فاکتورسازی می‌شود. انتخاب مدل برای شناسایی رتبه فاکتورسازی بهینه بر اساس پایداری هر راه حل و بازسازی آن از داده های اصلی اعمال می‌شود.



شکل 7. نمای کلی SigProfilerExtractor

در قسمت (**B**) **شکل 7**، نمایش شماتیک برای یک مثال تجزیه[۱۶۷] با رتبه فاکتورسازی k=3 که منعکس کننده سه امضای جهش عملی است نمایش داده شده است. به‌طور پیش‌فرض، SigProfilerExtractor فاکتورسازی 100 ماتریس غیرمنفی مستقل را انجام می‌دهد که ماتریس M قبل از هر فاکتورسازی مجدداً نمونه‌برداری و نرمال‌سازی می‌شود (نشان داده شده با «^»). خوشه‌بندی پارتیشن از 100 فاکتورسازی برای ارزیابی رتبه ثبات

[167] decomposition

فاکتورسازی[۱۶۸]، اندازه‌گیری شده در مقادیر silhouette استفاده می‌کند. خوشه‌بندی همچنین می‌تواند به‌عنوان پیش‌بینی‌های دو بعدی ارائه شود که امضاهای جهشی مشابه بیشتری را نشان می‌دهد، همانطور که برای سه نمونه امضا نشان داده شده است. مرکز راه حل‌های خوشه‌ای (که با "-" نشان داده شده است) با ماتریس اصلی M مقایسه می‌شود. قسمت (**C**) **شکل 7** نشان میدهد که همه امضاهای de novo شناسایی شده با ترکیبی از امضاهای جهشی شناخته شده COSMIC مطابقت دارند. در این بخش یک مثال برای امضای de novo استخراج شده SBS96B، که با ترکیبی از امضاهای COSMIC SBS1، SBS2 و SBS13 مطابقت دارد، ارائه شده است.

یکی از چالش‌ها در مرحله کشف de novo امضاهای جهشی، این است که در آن امضاهای جهشی مستقیماً از یک گروه استخراج می‌شوند. ابزارهای متعددی توسعه داده شده است که بیشتر آن‌ها از نظر الگوریتمی، بر پایه فاکتورسازی ماتریس غیرمنفی (NMF) (Lee and Seung 1999) یا روی رویکردهای ریاضی مشابه NMF (Dempster, Laird et al. 1977, Févotte and Cemgil 2009, Suri and Roy 2017) متکی هستند. مزیت اصلی NMF نسبت به سایر رویکردهای فاکتورسازی، توانایی آن در به دست آوردن عوامل غیرمنفی است که بخشی از داده‌های اصلی هستند. بنابراین امکان تفسیر بیولوژیکی عوامل غیرمنفی شناسایی شده را فراهم می‌کند (Lee and Seung 1999). با این حال آن‌ها حتی برای یک مجموعه داده منجر به ایجاد نتایج بسیار متغیری می‌شوند (Omichessan, Severi et al. 2019, Alexandrov, Kim et al. 2020). بنابراین مقایسه امضاهای کشف شده در طول مطالعات دشوار است، به ویژه تعیین اینکه آیا یک امضا، جدید است یا صرفاً تغییری از امضای شناخته شده قبلی به دلیل بایاس الگوریتمی است. این موضوع ممکن است زیربنای تعداد زیادی از امضاهایی باشد که بیش از حد مشابه هستند یا می‌توانند به عنوان ترکیبات خطی یکدیگر در فهرست پرکاربرد امضاهای شناخته شده از کاتالوگ جهش‌های پیکری در سرطان (COSMIC) بیان شوند. همچنین ممکن است ویژگی بافت مشاهده شده برای برخی از امضاها را مخدوش کند (Degasperi, Amarante et al. 2020).

در این میان روش MuSiCal که توسط تیم Peter J. Park از دانشگاه هاروارد توسعه داده و در فوریه 2024 منتشر شد (Jin, Gulhan et al. 2024)، یک چارچوب جامع را ارائه می‌کند که تخصیص امضای دقیق و همچنین کشف امضای قوی و حساس را امکان‌پذیر می‌کند. MuSiCal برای حل چالش عنوان شده، از چندین روش جدید استفاده می‌کند، از جمله NMF حداقل حجم[۱۶۹] (mvNMF) (Craig 1994, Miao and Qi 2007, Ang

[168] factorization stability rank

[169] minimum-volume NMF (mvNMF)

and Gillis 2019, Leplat, Gillis et al. 2020)، حداقل مربعات غیرمنفی پراکنده مبتنی بر احتمال[۱۷۰] (NNLS) و یک رویکرد مبتنی بر داده برای بهینه‌سازی سیستماتیک پارامترها و اعتبارسنجی مناسب. این رویکرد اخیر در قیاس با روش بسیار معروف SigProfilerExtractor که توسط تیم L. Alexandrov از دانشگاه سن دیگو در 2022 معرفی شده بود نتایج بهتری را بدست آورد.

2.1.2 بازسازی و تخصیص امضاهای جهشی

تخصیص امضاهای جهشی به نمونه‌های سرطان فردی و جهش‌های پیکری فردی فرصتی را برای شناسایی فرآیندهای مسئول جهش‌های سوماتیکی به صورت نمونه به نمونه[۱۷۱] فراهم می‌کند و ما را قادر می‌سازد تا فرآیندهای جهشی را در ژنوم سرطان مشخص کنیم.

در دهه گذشته، ابزارهای متعددی برای بازسازی مجدد امضاهای شناخته شده، از جمله deconstructSigs (Rosenthal, McGranahan et al. 2016)، Mutational Patterns (Blokzijl, Janssen et al. 2018, Manders, Brandsma et al. 2022)، sigLASSO (Li, Crawford et al. 2020) و SignatureTool (Degasperi, Amarante et al. 2020, Degasperi, Zou et al. 2022) توسعه یافته‌اند (**جدول 2**). اکثر این ابزارها تقریباً به طور انحصاری از امضاهای SBS پشتیبانی می‌کنند و فاقد یک رابط آنلاین هستند. اگرچه چند ابزار وب، از جمله MuSiCa (Díaz-Gay, Vila-Casadesús et al. 2018) و Mutalisk (Lee, Lee et al. 2018) نیز وجود دارد.

Tool name	Input data (mutations)	Platform	Optimization Method	Algorithm	Computational engine	Penalties	Post hoc filter (TMB threshold)
DeconstructSigs (2016) (Rosenthal, McGranahan et al. 2016)	matrix, custom	R	Multiple linear regression with a nonnegative cutoff on activities	Golden-section search algorithm (Kiefer 1953)	Original implementation	Addition penalty (SSE; default: 0.001)	**Yes (6%)**
MutationalPatterns (2018) (Blokzijl, Janssen et al. 2018)	matrix, VCF	R	NNLS	Levenberg Marquardt algorithm (Levenberg 1944)	Pracma R package (Borchers 2022)	No penalties	**No**
MutationalPatterns (strict) (2022) (Manders, Brandsma et al. 2022)	matrix, VCF	R	NNLS	Levenberg Marquardt algorithm (Levenberg 1944)	Original implementation (penalty framework)	Removal penalty (cosine similarity; default: 0.004)	**No**

[170] likelihood-based sparse nonnegative least squares (NNLS)

[171] sample-by-sample

					and Pracma R package (Borchers 2022)		
sigLASSO (2020) (Li, Crawford et al. 2020)	matrix, VCF, MAF, custom	R	Non-negative linear LASSO regression	Alternative convex search algorithm (Gorski, Pfeuffer et al. 2007)	Original implementation (framework) and glmnet R package (Lasso regression) (Friedman, Hastie et al. 2010)	Optimized penalty (L1 norm). Priors. Lambda hyperparameter.	**No**
SignatureToolsLib (2020, 2022) (Degasperi, Amarante et al. 2020, Degasperi, Zou et al. 2022)	matrix, VCF, BEDPE, custom	R Web app	Non-negative linear regression (KL Divergence objective function)	Lee's multiplicative algorithm (Lee and Seung 1999)	NNLM R package (Lin and Boutros 2020)	No penalties	**Yes (5%)**
✦SigProfilerAssignment (2023) (Díaz-Gay, Vangara et al. 2023)	matrix, VCF, MAF, custom segmentation,	Python R Web app	NNLS	Lawson Hanson algorithm (Ling 1977)	Original implementation (penalty framework) and Scipy python package (NNLS) (Virtanen, Gommers et al. 2020)	Initial removal, addition, and removal penalties (L2 norm; default: 0.05, 0.05 and 0.01)	**No**

جدول 2. مروری بر ابزارهای بیوانفورماتیک توسعه داده شده جهت تخصیص امضاهای جهشی

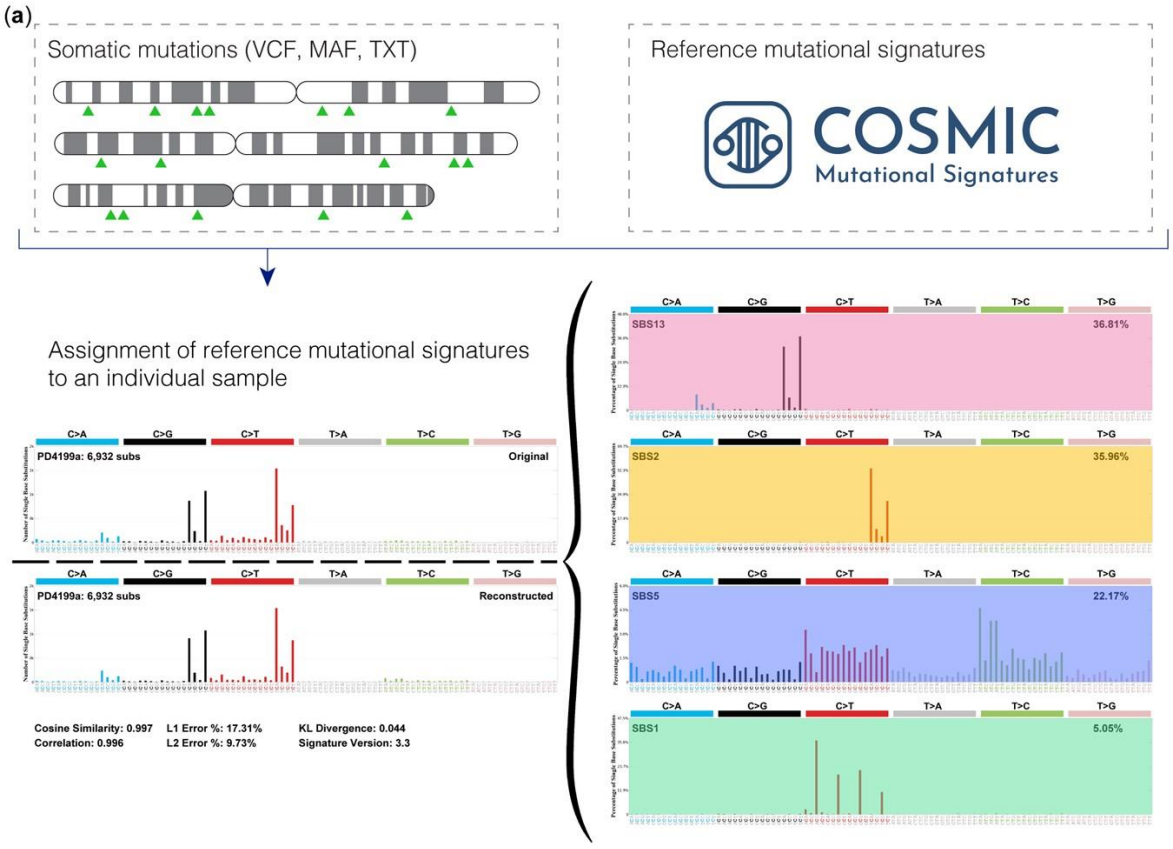
MAF، فرمت حاشیه‌نویسی جهش؛ matrix: ماتریس جهشی[۱۷۲]; NNLM، مدل‌های خطی غیر منفی[۱۷۳]؛ NNLS، کمترین مربعات غیر منفی[۱۷۴]؛ SSE، مجموع مربعات خطاها[۱۷۵]؛ TMB، بار جهش تومور[۱۷۶]؛ VCF، فرمت فراخوانی جهش.

جدول 2 مروری بر ابزارهای بیوانفورماتیک توسعه داده شده جهت تخصیص امضاهای جهشی را نشان می‌دهد. ابزارها بر اساس حروف الفبا مرتب شده اند. ستون‌های جدول موارد زیر را نشان می‌دهند: نام چهارچوب محک زدن ابزار، انواع داده‌های ورودی پشتیبانی‌شده، پلت‌فرم‌های عملیاتی سازگار، روش بهینه‌سازی استفاده‌شده، الگوریتم برازش اولیه[۱۷۷]، موتور محاسباتی استفاده‌شده، جریمه‌های اضافی اعمال‌شده، و آستانه درصد بار جهش تومور برای جلوگیری از بیش برازش امضاها.

در این میان SigProfilerAssignment، یک دسکتاپ و یک چارچوب محاسباتی آنلاین برای تخصیص تمام انواع امضاهای جهشی، از جمله مجموعه‌های COSMIC از امضاهای مرجع SBS، DBS، ID و CN به نمونه‌های

[172] mutational matrix
[173] non-negative linear models
[174] non-negative least squares
[175] sum of squared errors
[176] tumor mutational burden
[177] primary fitting algorithm

مجزا را ارائه می‌کند (**شکل a8 و b**) که توسط تیم L. Alexandrov در دسامبر 2023 معرفی شد (Díaz-Gay, Vangara et al. 2023).



شکل 8. اختصاص امضاهای جهش شناخته شده به یک نمونه فردی و جهش های فردی با SigProfilerAssignment، و محک زدن با چهار ابزار بیوانفورماتیک دیگر

مطابق **شکل 8** که اختصاص امضاهای جهش شناخته شده به یک نمونه فردی و جهش های فردی با SigProfilerAssignment، و محک زدن با چهار ابزار بیوانفورماتیک دیگر را نشان می‌دهد. SigProfilerAssignment از داده های ورودی در قالب استاندارد (VCF، MAF، یا text) پشتیبانی می‌کند و اجازه می‌دهد تا مجموعه‌ای از امضاهای شناخته شده (به عنوان مثال از پایگاه داده COSMIC) را به یک (a) نمونه فردی و (b) به احتمال زیاد به یک جهش سوماتیک فردی اختصاص دهیم. توجه داشته باشید که تخصیص احتمالی امضاهای جهش به یک جهش سوماتیک فردی تنها در صورتی امکانپذیر است که کاربر به جای بردار جهش[۱۷۸]، فهرستی از جهش‌های فردی (به عنوان مثال فایل VCF) را برای نمونه بررسی شده ارائه کند، زیرا یک بردار جهش فاقد اطلاعات برای جهش‌های فردی است.

با توجه به مجموعه‌ای از امضاهای جهشی شناخته شده و مجموعه‌ای از جهش‌ها در ژنوم سرطان، که هر دو تحت یک طرح جهش[۱۷۹] طبقه‌بندی شده‌اند (Alexandrov, Nik-Zainal et al. 2013, Bergstrom, Huang et al. 2019)، SigProfilerAssignment تعداد جهش‌های ایجاد شده توسط هر امضا را در آن ژنوم سرطان مشخص می‌کند (**شکل a8**).

اکنون یک رابط آنلاین کاربرپسند از ابزار SigProfilerAssignment جهت استفاده، به عنوان بخشی از وب سایت امضاهای جهشی COSMIC (Tate, Bamford et al. 2019) در https://cancer.sanger.ac.uk/signatures/assignment/ در دسترس است. SigProfilerAssignment اولین ابزاری است که امکان آنالیز امضاهای تعداد کپی و تخصیص احتمالی امضاها را به جهش‌های پیکری فردی فراهم می‌کند که پیش‌بینی کننده خوبی برای بقای بالینی هستند (Drews, Hernando et al. 2022, Steele, Abbasi et al. 2022). علاوه بر این، SigProfilerAssignment از تخصیص امضاهای جهشی استخراج شده de novo و مجموعه‌ای از امضاهای سفارشی ارائه شده توسط کاربر پشتیبانی می‌کند. این ابزار به عنوان موتور محاسباتی خود برای تعیین کمیت تعداد جهش‌های نقش شده توسط

[178] mutational vector
[179] mutational schema

هر امضا، از پیاده‌سازی سفارشی الگوریتم مرحله‌ای پیش‌رونده[۱۸۰] (Hastie, Tibshirani et al. 2009) برای رگرسیون پراکنده[۱۸۱] و کمترین مربعات غیرمنفی (NNLS) بر اساس روش لاوسون-هانسون[۱۸۲] (Ling 1977) برای بهینه‌سازی عددی استفاده می‌کند و نتایج مقایسه روش‌های مختلف در این زمینه نشان می‌دهد که SigProfilerAssignment از سایر ابزارهای رایج بهتر عمل می‌کند. الگوریتم ابزار در **الگوریتم 1** نشان داده شده است و در ادامه توضیح داده شده است. SigProfilerAssignment علاوه بر تعیین کمیت فعالیت هر امضای جهشی، امضاهای شناخته شده را به تک تک جهش‌ها (**شکل b8**) بر اساس زمینه جهش خاص آن‌ها تخصیص می‌دهد.

2.1.2.1 شرح الگوریتم SigProfilerAssignment

از نظر ریاضی، یک طرح جهش را می‌توان به عنوان یک الفبای محدود[۱۸۳] Ξ از انواع جهش نشان داد که در مجموع شامل حروف ξ است. در اینجا، یک امضای جهشی به عنوان یک تابع جرم احتمال[۱۸۴] با دامنه الفبای Ξ تعریف شده است. در نماد برداری[۱۸۵]، یک امضای جهشی را می‌توان به عنوان $\vec{s} = [s_1, s_2, \dots, s_\xi]^T$ نشان داد، که در آن s_k ، $1 \le \mathrm{k} \le \xi$، احتمالی است برای امضای جهش، $\vec{s}$، باعث ایجاد جهش‌هایی از نوع متناظر با حرف k ام الفبای Ξ شود. از آنجایی که یک امضای جهشی یک تابع جرم احتمالی است، $0 \le s_k \le 1$ و $\sum_{k=1}^{\xi} s_k = 1$. به این ترتیب، مجموعه‌ای از n امضای جهشی شناخته شده را می‌توان به عنوان یک ماتریس امضا، $S \in \mathbb{R}_+^{\xi \times \mathrm{n}}$، که در آن $S = [\vec{s}^1, \vec{s}^2, \dots, \vec{s}^n]$، بیان کرد. علاوه بر این، مجموعه‌ای از جهش‌ها در ژنوم سرطان را می‌توان به عنوان $v: \Xi \to \mathbb{N}_+^{\xi}$ تعریف کرد. در نماد برداری، مجموعه‌ای از جهش‌ها در ژنوم سرطان $\vec{v} = [v_1, v_2, \dots, v_\xi]^T$، که در آن v_k ، $1 \le \mathrm{k} \le \xi$، تعداد جهش‌های آن ژنوم سرطان از نوع جهش مربوط به حرف k ام الفبای Ξ را منعکس می‌کند.

SigProfilerAssignment یک ماتریس امضا، S و مجموعه‌ای از جهش‌ها، $\vec{v}$، را به عنوان ورودی می‌گیرد تا بردار ستونی از فعالیت‌های $\vec{a} = [a_1, a_2, \dots, a_\mathrm{n}]^T$ ، که در آن $a_t \in \mathbb{N}_0^{\mathrm{n}}$ ، $1 \le \mathrm{t} \le \mathrm{n}$، متناظر با تعداد

---

[180] forward stagewise algorithm
[181] sparse regression
[182] Lawson-Hanson method
[183] finite alphabet
[184] probability mass function
[185] vector notation

جهش‌های سوماتیک منتسب به t امین امضای جهش است را به عنوان خروجی تولید کند. فرض اساسی تخصیص امضاهای جهشی این است که جهش‌های درون یک نمونه را می‌توان به عنوان برهم‌نهی[۱۸۶] امضاهای جهشی شناخته شده و فعالیت‌های آن‌ها تقریب زد:

رابطه 1

$$\vec{v} \approx S\vec{a}$$

بنابراین، با توجه به $\vec{a} \geq 0$، باید بردار $\vec{a}$ را استخراج کرد که به بهترین وجه با داده‌های ورودی ارائه شده مطابقت دارد. برای حل این مشکل بهینه‌سازی، SigProfilerAssignment از پیاده‌سازی سفارشی الگوریتم مرحله‌ای پیش‌رونده (Hastie, Tibshirani et al. 2009) استفاده می‌کند و حداقل مربعات غیرمنفی (NNLS) (Lawson and Hanson 1995) را بر اساس روش لاوسون-هانسون (Lawson and Hanson 1995) اعمال می‌کند:

رابطه 2

$$\min_{\vec{a} \geq 0} \|\vec{v} - S\vec{a}\|_2^2$$

الگوریتم ابتدا با محاسبه حداقل خطای نسبی[۱۸۷]، $\epsilon_{min} \frac{\|\vec{v}-S\vec{a}\|_2^2}{\|\vec{v}\|_2^2}$، با استخراج بردار $\vec{a}$ غیرمنفی بهینه برای مجموعه کامل همه امضاهای مرجع، S، با استفاده از **رابطه 2** شروع می‌شود. این حداقل خطا بهترین توضیح ممکن را در مورد داده‌ها ارائه می‌دهد، اما همچنین منجر به بیش برازش [۱۸۸] می‌شود زیرا از همه امضاهای موجود استفاده می‌شود.

در مرحله بعد، این ابزار به ترتیب از مراحل حذف و اضافه کردن امضا بر اساس الگوریتم‌های مرحله‌ای پیش‌رونده و پس‌رونده[۱۸۹] استفاده می‌کند (Hastie, Tibshirani et al. 2009). ابتدا، امضاها با استفاده از یک الگوریتم مرحله‌ای پس‌رونده حذف می‌شوند (**الگوریتم 1**). به طور خاص، هر امضا از مجموعه امضای مرجع، S، به طور مکرر

[186] superposition
[187] minimum relative error
[188] overfitting
[189] backward and forward stepwise algorithms

حذف می‌شود و مجموعه امضای باقیمانده، $\hat{S}$، با اعمال **رابطه 2** به نمونه $\vec{v}$ نسبت داده می‌شود. افزایش خطای نسبی، $\epsilon_j \frac{\|\vec{v}-\hat{S}\vec{a}\|_2^2}{\|\vec{v}\|_2^2} - \epsilon_{min}$، به دلیل حذف یک امضا با اضافه کردن امضای j ام از S محاسبه می‌شود. امضایی با کمترین افزایش نسبی در میزان خطا از مجموعه امضا، S، حذف می‌شود، مشروط بر اینکه افزایش، کمتر از یک آستانه خاص (مقدار پیش فرض 0.01) باشد. پس از حذف نهایی امضا با حداقل افزایش نرخ خطای نسبی، حداقل خطای نسبی، ϵ_{min}، و مجموعه‌ی امضاها، S ، به روز می‌شوند تا این حذف را منعکس کنند. مرحله حذف تکرار می‌شود تا زمانی که همه امضاهایی که شرایط را برآورده می‌کنند از S حذف شوند. مراحل حذف با مراحل جمع بر اساس الگوریتم مرحله‌ای پیش‌رونده (**الگوریتم 1**) دنبال می‌شود. به طور خاص، هر یک از امضاهای مرجع حذف شده قبلی به طور مکرر به S مجدد اضافه می‌شود و مجموعه امضای جدید، $\hat{S}$، با اعمال **رابطه 2** برای نمونه $\vec{v}$ مناسب است. بنابراین، کاهش خطای نسبی، $\epsilon_l = \epsilon_{min} - \frac{\|\vec{v}-\hat{S}\vec{a}\|_2^2}{\|\vec{v}\|_2^2}$، به دلیل اضافه کردن یک امضا، با اضافه کردن امضای l ام به S محاسبه می‌شود. امضا با حداکثر کاهش نسبی میزان خطا، به مجموعه امضا، S ، برمی‌گردد، مشروط بر اینکه افزایش، بیش از یک آستانه خاص (مقدار پیش فرض 0.05) باشد. پس از افزودن نهایی امضا با بیشترین کاهش نرخ نسبی، حداقل خطای نسبی، ϵ_{min}، و مجموعه امضاها، S، به روز می‌شوند تا این اضافه را منعکس کنند. مرحله جمع تکرار می‌شود تا زمانی که همه امضاهایی که شرایط را برآورده می‌کنند به S مجدد اضافه شوند. در نهایت، مراحل جمع و حذف تا زمان همگرایی تکرار می‌شود، جایی که هیچ امضایی اضافه یا از لیست امضاها حذف نمی‌شود (**الگوریتم 1**).

SigProfilerAssignment علاوه بر تعیین کمیت فعالیت هر امضای جهشی، امضاهای شناخته شده را نیز بر اساس زمینه جهش خاصی به تک تک جهش‌ها اختصاص می‌دهد.

رابطه 3

$$p_k^t = \frac{s_k^t a_t}{[S\vec{a}]_k}$$

که در آن، p_k^t نشان دهنده احتمال جهش مربوط به حرف k ام الفبای Ξ است که توسط امضای t ام در نمونه ایجاد می‌شود؛ s_k^t احتمال امضای t ام است که منجر به ایجاد جهش مربوط به حرف k ام الفبای Ξ است؛ a_t تعداد جهش‌هایی است که به امضای جهش t ام نسبت داده می‌شود؛ و $[S\vec{a}]_k$ مقدار k امین عنصر برداری است که از ضرب ماتریسی ماتریس امضا، S، و فعالیت‌های امضای مشتق شده، $\vec{a}$ به دست می‌آید.

الگوریتم 1. تخصیص امضاهای جهشی به نمونه‌ها با SigProfilerAssignment

Input: $\vec{v} \in \mathbb{N}_{+}^{\xi \times 1}$ (a vector corresponding to a set of mutations in a sample) and
$S \in \mathbb{R}_{+}^{\xi \times n}$ (a matrix corresponding to a set of n known mutational signatures)
Output: $\vec{a} \in \mathbb{N}_{+}^{n \times 1}$ (the vector reflecting the activities of the n known signatures in sample $\vec{v}$)

$\epsilon_{\min}, \vec{a} =$ ***calcNNLS*** $(\vec{v}, S)$
$S^{\text{all}} = S$

While FLAG = True:
    $\epsilon_{\min}, S =$ ***removeSignatures*** $(\vec{v}, S, \epsilon_{\min})$
    $\epsilon_{\min}, S =$ ***addSignatures*** $(\vec{v}, S^{\text{all}}, S, \epsilon_{\min})$
    Set FLAG = False if S remains constant and there is no addition or removal of signatures
END While
$\epsilon_{\min}, \vec{a} =$ ***calcNNLS*** $(\vec{v}, S)$
Return $\vec{a}$

FUNCTION removeSignatures $(\vec{v}, S, \epsilon_{\min})$
    While FLAG = True:
        For j in **1** to size(S, 2) **do** *// loop from 1 to the total number of signatures in S*
            $\hat{S} = S[:, -j]$ *// remove the j^{th} signature from S*
            $\epsilon[j], \vec{a}_j =$ ***calcNNLS*** $(\vec{v}, \hat{S})$
        END For
        $\text{minIndex}, \text{minValue} = \min(\epsilon)$ *// find the signature set with least relative error*
        If ($\text{minValue} - \epsilon_{min} \leq 0.01$)
            $S = S[:, -\text{minIndex}]$
        else
            Return $\text{minValue}, S$
        END If
    END While
END removeSignatures

FUNCTION addSignatures $(\vec{v}, S^{\text{all}}, S, \epsilon_{\min})$
    While FLAG = True:
        For p in **1** to size(S^{all}, 2) **do** *// loop from 1 to the total number of signatures in S^{all}*
            $\hat{S} = \left[S; S^{\text{all}}[:, p]\right]$ *// add the p^{th} signature from S^{all}*
            $\epsilon[j], \vec{a}_j =$ ***calcNNLS*** $(\vec{v}, \hat{S})$
        END For
        $\text{minIndex}, \text{minValue} = \min(\epsilon)$ *// find the signature set with least relative error*
        If ($\epsilon_{min} - \text{minValue} \geq 0.05$)
            $S = \left[S; S^{\text{all}}[:, \text{minValue}]\right]$
        else
            Return $\text{minValue}, S$
        END If
    END While
END addSignatures

FUNCTION calcNNLS $(\vec{v}, S)$
    $\vec{a} = nnls(S, \vec{v})$ *// Calculating NNLS with the Lawson-Hanson method*
    $\epsilon = ||\vec{v} - S\vec{a}||_2^2 \,/\, ||\vec{v}||_2^2$ *// Computing relative error*
    Return $\epsilon, \vec{a}$
END calcNNLS

2.1.3 توزیع و کاربرد

داده‌های ورودی برای هر دو نسخه دسکتاپ و آنلاین را می‌توان با فراخوانی جهش و فایل‌های تقسیم‌بندی[۱۹۰]، بسته به نوع کلاس، ارائه کرد و به صورت داخلی توسط SigProfilerMatrixGenerator (Bergstrom, Huang et al. 2019, Khandekar, Vangara et al. 2023) پردازش می‌شود. این ابزار از فرمت‌های رایج برای جهش‌های سوماتیک SBS، DBS و ID، از جمله فرمت فراخوانی جهش[۱۹۱] (VCF)، فرمت حاشیه‌نویسی جهش[۱۹۲] (MAF) و فایل‌های متنی ساده پشتیبانی می‌کند. فایل‌های تقسیم‌بندی چند نمونه به‌دست آمده از ASCAT (Van Loo, Nordgard et al. 2010)، ABSOLUTE (Carter, Cibulskis et al. 2012)، Sequenza (Favero, Joshi et al. 2015)، FACETS (Shen and Seshan 2016)، Battenberg (Van Loo, Nordgard et al. 2010)، یا PURPLE (Shale, Cameron et al. 2022) برای تجزیه و تحلیل امضاهای CN پشتیبانی می‌شوند. علاوه بر این، SigProfilerAssignment می‌تواند از ماتریس‌های جهشی استاندارد استفاده کند، جایی که ردیف‌ها با کانال‌های جهشی و ستون‌ها با نمونه‌ها مطابقت دارند که از مجموعه ابزارهای SigProfiler استخراج شده‌اند (Bergstrom, Huang et al. 2019, Degasperi, Zou et al. 2022, Islam, Díaz-Gay et al. 2022). سنجش توالی‌یابی مختلف (توالی‌یابی کل ژنوم[۱۹۳]، توالی‌یابی کل اگزوم[۱۹۴]، و توالی‌یابی هدفمند[۱۹۵])، گونه ها (انسان، موش[۱۹۶] و موش صحرایی[۱۹۷])، ساختارهای ژنومی (GRCh37/38، mm9/10، و rn6)، و امضاها (پیش فرض COSMICv3.310 (Tate, Bamford et al. 2019)، نسخه‌های قبلی COSMIC و پایگاه‌های داده امضای سفارشی) پشتیبانی می‌شوند.

خروجی اصلی SigProfilerAssignment شامل فعالیت هر امضای جهش شناخته شده برای هر یک از نمونه‌های ارائه شده، بازسازی مجموعه داده اصلی، و احتمال ایجاد هر جهش فردی توسط یک امضای خاص است. زمانی که فایل ورودی یک بردار جهش یا ماتریس جهشی است، مورد دوم ارائه نمی‌شود، زیرا این قالب ورودی فاقد اطلاعات مربوط به جهش‌های سوماتیکی فردی است. فعالیت‌های امضا با تعداد خاصی از جهش‌ها از کاتالوگ اصلی ناشی از یک فرآیند جهشی خاص مطابقت دارد. با در نظر گرفتن این فعالیت‌ها، و همچنین مجموعه ارائه شده از

[190] segmentation
[191] Variant Call Format
[192] Mutation Annotation Format
[193] whole genome sequencing
[194] whole exome sequencing
[195] targeted sequencing
[196] mouse
[197] rat

امضاهای جهش شناخته شده، بازسازی کاتالوگ جهشی اصلی برای هر نمونه مشتق شده است. معیارهای دقت[۱۹۸] مختلف برای این بازسازی توسط SigProfilerAssignment، از جمله شباهت کسینوس[۱۹۹]، واگرایی Kullback-Leibler [۲۰۰]، همبستگی پیرسون[۲۰۱]، خطای نسبی L1 [۲۰۲] و خطای نسبی L2 به دست می‌آید.

نتایج تخصیص امضا با استفاده از سه بصری‌سازی مستقل خلاصه می‌شود: (1) نمودار نواری که فعالیت‌های همه امضاهای جهش‌یافته را در یک نمونه نشان می‌دهد. (2) نمودار امضای بار جهشی تومور[۲۰۳] (TMB) که فعالیت‌های هر امضای جهشی را نشان می‌دهد. و (3) یک نمودار بازسازی فردی در هر نمونه، که شامل پروفایل‌های جهش برای هر دو نمونه ورودی اصلی و بازسازی شده، معیارهای دقت مختلف، و پروفایل‌های جهش برای هر یک از امضاهای جهش شناخته شده اختصاص داده شده به آن نمونه است. برای نسخه آنلاین ابزار، نمودار نقشه حرارتی تعاملی، شامل فعالیت‌های امضاها و دقت بازسازی نمونه‌ها نیز ارائه شده است. فایل‌های داده خام حاوی فعالیت‌ها، معیارهای بازسازی، و احتمالات امضا برای جهش‌های فردی توسط ابزار دسکتاپ تولید می‌شوند و می‌توانند از نسخه آنلاین دانلود گردند.

2.1.4 محک زدن ابزارهای بیوانفورماتیک برای بازسازی مجدد امضاهای جهشی شناخته شده

برای ارزیابی عملکرد ابزارها برای بازسازی امضاهای جهش شناخته شده، از مجموعه استانداردی از معیارهای ارزیابی استفاده و SigProfilerAssignment با چهار رویکرد رایج دیگر مقایسه شد: deconstructSigs (Rosenthal, McGranahan et al. 2016)، MutationalPatterns (Blokzijl, Janssen et al. 2018, Manders, Brandsma et al. 2022)، sigLASSO (Li, Crawford et al. 2020)، و SignatureToolsLib (Degasperi, Amarante et al. 2020, Degasperi, Zou et al. 2022). به طور خاص، هر ابزار روی 2700 ژنوم سرطان قبلا شبیه‌سازی شده (Islam, Díaz-Gay et al. 2022)، مربوط به 300 تومور شبیه‌سازی شده از 9 نوع سرطان مختلف استفاده شد. ژنوم سرطان این نمونه‌ها با استفاده از 21 امضای مرجع COSMIC SBS مختلف شبیه‌سازی شده‌اند.

---

[198] accuracy metrics
[199] cosine similarity
[200] Kullback–Leibler divergence
[201] Pearson correlation
[202] L1 relative error
[203] tumor mutational burden (TMB)

برای تقلید از یک بازسازی معمولی از امضاهای جهشی، هر ابزار با استفاده از مجموعه کامل 79 امضای COSMICv3.3 SBS استفاده شد. پس از تخصیص امضاها، انتساب هر امضا به هر نمونه به عنوان نتیجه مثبت واقعی[۲۰۴] (TP)، مثبت کاذب[۲۰۵] (FP)، یا منفی کاذب[۲۰۶] (FN) طبقه‌بندی شد. اگر حداقل یک جهش توسط یک ابزار خاص به امضا اختصاص داده شود و فعالیت یافته‌های عینی امضا بزرگتر از صفر باشد، یک امضای شناخته شده TP در نظر گرفته می‌شود. در مقابل، زمانی که امضا توسط یک ابزار تخصیص داده شد، اما فعالیت یافته‌های عینی صفر بود، به عنوان FP طبقه‌بندی شد. در نهایت، نتایج FN امضاهایی با فعالیت‌های یافته‌های عینی بالای صفر بود که هیچ جهش سوماتیکی به آن‌ها اختصاص داده نشد. این معیارهای استاندارد امکان محاسبه دقت[۲۰۷]، حساسیت[۲۰۸] و امتیاز F1 [۲۰۹] هر ابزار در هر نمونه را فراهم می‌کند که به صورت زیر تعریف می‌شود:

رابطه 4

$$Precision = \frac{TP}{TP + FP}$$

رابطه 5

$$Sensitivity = \frac{TP}{TP + FN}$$

رابطه 6

$$F_1\text{score} = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

این معیارها برای هر نمونه تولید شده به صورت مصنوعی محاسبه شده و متعاقباً برای به دست آوردن یک مقدار دقت نهایی برای هر سطح نویز تصادفی (0٪، 5٪ و 10٪) میانگین گرفته می‌شود.

---

[204] *true positive* (TP)
[205] *false positive* (FP)
[206] *false negative* (FN)
[207] precision
[208] sensitivity
[209] F1 score

برای محک زدن ID و DBS، پروفایل‌های جهش مصنوعی با پیروی از همان روش مورد استفاده برای ساخت مجموعه داده SBS منتشر شده قبلی (Islam, Díaz-Gay et al. 2022)، با استفاده از تابع GenerateSyntheticTumors پکیج SynSigGen R تولید شدند:

https://github.com/steverozen/SynSigGen

این پکیج از فعالیت‌های اصلی از تجزیه و تحلیل PCAWG امضاهای جهشی (Alexandrov, Kim et al. 2020) برای استخراج پروفایل‌های جهش مصنوعی در هر نوع سرطان استفاده می‌کند. این فرآیند شبیه‌سازی مستلزم آن است که حداقل دو امضای مختلف به هر نمونه از هر نوع سرطان خاص اختصاص داده شود. با در نظر گرفتن این موضوع، ما مجموعه داده‌های مصنوعی را برای کلاس‌های نوع DBS و ID با استفاده از همان 9 نوع سرطان که قبلاً در معیار SBS استفاده شده بود (300 نمونه شبیه‌سازی شده از هر نوع سرطان)، از جمله کارسینوم سلول انتقالی مثانه، آدنوکارسینوم مری، آدنوکارسینوم سینه، سلول سنگفرشی ریه، کارسینوم سلول کلیه، آدنوکارسینوم تخمدان، استئوسارکوم، آدنوکارسینوم دهانه رحم و آدنوکارسینوم معده تولید کردیم. با این حال، به دلیل محدودیت ذکر شده در بالا، آدنوکارسینوم دهانه رحم برای تولید پروفایل ID مصنوعی حذف شد زیرا تنها ID1 در فعالیت‌های اصلی نمونه‌های PCAWG وجود داشت. برای تولید مجموعه داده مصنوعی DBS، نیز آدنوکارسینوم دهانه رحم حذف شد (فقط DBS4 به یکی از نمونه‌های PCAWG اختصاص داده شد)، همراه با کارسینوم سلول سنگفرشی ریه، زیرا تنها امضای DBS2 مرتبط با تنباکو به چندین مورد از PCAWG اختصاص داده شد. به طور خلاصه، 2100 نمونه DBS مصنوعی و 2400 نمونه ID مصنوعی تولید شد (به ترتیب 300 نمونه برای هر یک از هفت و هشت نوع سرطان). در مورد تغییرات CN، از آنجایی که این نوع جهش توسط SynSigGen پشتیبانی نمی‌شود، فعالیت‌های pan-cancer از مطالعه اصلی که امضاهای CN COSMICv3.3 را توصیف می‌کند (Steele, Abbasi et al. 2022) برای به دست آوردن یک مجموعه داده مصنوعی شامل 9699 نمونه مصنوعی از 33 نوع سرطان مختلف استفاده و در مرجع ضرب شد. با توجه به مجموعه ورودی امضاهای جهشی شناخته شده، در هر سه مورد از جدیدترین نسخه COSMICv3.3 از امضاهای مرجع، شامل 18 ID، 11 DBS، و 24 امضای CN استفاده شد.

برای محک زدن عملکرد محاسباتی ابزارهای مختلف بیوانفورماتیک، زمان سپری شده CPU و حداکثر استفاده از حافظه آن‌ها نظارت و میانگین برای سه سطح نویز محاسبه شد.

SigProfilerAssignment نسخه 0.0.28 با استفاده از پارامترهای پیش فرض اجرا شد. deconstructSigs v1.8.0 (Rosenthal, McGranahan et al. 2016) با پارامترهای پیش فرض همانطور که در https://github.com/raerose01/deconstructSigs/ نشان داده شده است استفاده شد. MutationalPatterns v3.0.1 (Manders, Brandsma et al. 2022) با پارامترهای پیش‌فرض به‌طور مستقل با استفاده از حالت‌های استاندارد و سخت اجرا شد که به ترتیب مربوط به توابع ***fit_to_signatures*** و ***fit_to_signatures_strict*** هستند. طبق دستورالعمل‌های نویسندگان در:

https://bioconductor.org/packages/release/bioc/vignettes/MutationalPatterns/inst/doc/Introduction_to_MutationalPatterns.html

پارامتر ***max_delta*** به مقدار پیش‌فرض 0.004 برای حالت سخت‌گیرانه ثابت شد. sigLASSO v1.1 (Li, Crawford et al. 2020) با پارامترهای پیش‌فرض (بدون اولویت) به دنبال دستورالعمل‌های https://github.com/gersteinlab/siglasso استفاده شد. اگرچه از تولید نمودارها برای مقایسه عملکرد محاسباتی اجتناب می شود. SignatureToolsLib (Degasperi, Zou et al. 2022) نسخه 2.1.2 با امضاهای سراسری با استفاده از تابع ***Fit*** و پارامترهای پیش‌فرض همانطور که در https://github.com/Nik-Zainal-Group/signature.tools.lib نشان داده شده است اجرا شد.

2.1.5 بررسی کد و گیت هاب SigProfilerAssignment

2.1.5.1 در دسترس بودن داده ها

تمام داده‌های محک زده شده مصنوعی مورد استفاده در این مقاله در FigShare در: https://doi.org/10.6084/m9.figshare.24457114 موجود است و داده‌های چهارچوب معیار SBS در اصل به عنوان بخشی از (Islam, Díaz-Gay et al. 2022) تحت مجوز Creative Commons Attribution 4.0 International در دسترس عموم هستند:
https://doi.org/10.6084/m9.figshare.20409430.

2.1.5.2 در دسترس بودن کد

SigProfilerAssignment به عنوان یک پکیج Python توسعه داده شده و تحت یک مجوز مجاز BSD 2-clause در:

https://github.com/AlexandrovLab/SigProfilerAssignment
https://pypi.org/project/SigProfilerAssignment/

در دسترس است. یک بسته R نیز با استفاده از همان مجوز در:
https://github.com/AlexandrovLab/SigProfilerAssignmentR
ارائه می شود. SigProfilerAssignment از اکثر سیستم عامل‌ها، از جمله ویندوز، macOS، و سیستم‌های مبتنی بر لینوکس پشتیبانی می‌کند و دارای اسناد گسترده‌ای در https://osf.io/mz79v/wiki/home/ است.
علاوه بر این، یک رابط آنلاین کاربرپسند از ابزار به عنوان بخشی از وب سایت امضاهای جهشی COSMIC (Tate, Bamford et al. 2019) در https://cancer.sanger.ac.uk/signatures/assignment/ ارائه شده است. برای انطباق با قوانین حریم خصوصی اتحادیه اروپا و بریتانیا، وب سایت COSMIC قبل از استفاده از SigProfilerAssignment به ثبت نام رایگان نیاز دارد. این تضمین می کند که تمام داده‌های آپلود شده کاربر به صورت خصوصی نگهداری می‌شوند و به درستی پاک می‌شوند.

SigProfilerAssignment تخصیص امضاهای جهش شناخته شده قبلی را به نمونه‌های فردی و جهش‌های سوماتیکی فردی امکان پذیر می‌کند. این ابزار انواع مختلفی از امضاهای جهش مرجع، از جمله امضاهای COSMIC و همچنین پایگاه‌های داده امضای سفارشی را بازسازی می‌کند. بارسازی امضاهای جهش شناخته شده یک رویکرد بهینه‌سازی عددی است که نه تنها مجموعه‌ای از امضاهای جهشی عملیاتی را در یک نمونه خاص شناسایی می‌کند، بلکه تعداد جهش‌های اختصاص داده شده به هر امضای یافت شده در آن نمونه را نیز کمیت می‌دهد. SigProfilerAssignment از SigProfilerMatrixGenerator و SigProfilerPlotting استفاده می‌کند که به طور یکپارچه با سایر ابزارهای SigProfiler یکپارچه می‌شود.

برای کاربرانی که ترجیح می‌دهند در محیط R کار کنند، یک بسته wrapper ارائه شده است که می‌توان آن را پیدا و نصب کرد: https://github.com/AlexandrovLab/SigProfilerAssignmentR. مستندات دقیق را می‌توان در: https://osf.io/mz79v/wiki/home یافت.

2.1.5.2.1نصب

نسخه پایدار PyPi فعلی SigProfilerAssignment را نصب کنید:

```
$ pip install SigProfilerAssignment
```

اگر از فایل‌های فراخوان جهش (MAF، VCF یا فایل‌های متنی ساده) به عنوان ورودی استفاده می‌شود، لطفا ژنوم مرجع مورد نظر خود را به شرح زیر نصب کنید (ژنوم‌های مرجع موجود عبارتند از: GRCh37، GRCh38، mm9، mm10 و rn6):

```
$ python
from SigProfilerMatrixGenerator import install as genInstall
genInstall.install('GRCh37')
```

2.1.5.2.2اجرا

تخصیص امضاهای جهش شناخته شده به نمونه‌های فردی با استفاده از تابع ***cosmic_fit*** انجام می‌شود. نمونه‌های ورودی با استفاده از پارامتر نمونه‌ها در قالب فایل‌های فراخوانی جهش (VCF، MAF یا فایل‌های متنی ساده)، فایل‌های تقسیم‌بندی[۲۱۰] یا ماتریس‌های جهش ارائه می‌شوند. امضاهای جهشی COSMIC نسخه 3.4 به عنوان امضاهای مرجع پیش فرض استفاده می‌شوند، اگرچه نسخه‌های قبلی COSMIC و پایگاه‌های داده امضای سفارشی نیز با استفاده از پارامترهای ***cosmic_version*** و ***signature_database*** پشتیبانی می‌شوند. نتایج در پوشه مشخص شده در پارامتر خروجی یافت می‌شود.

```
from SigProfilerAssignment import Analyzer as Analyze
Analyze.cosmic_fit(samples, output, input_type="matrix", context_type="96",
collapse_to_SBS96=True, cosmic_version=3.4, exome=False,
genome_build="GRCh37",
signature_database=None,
exclude_signature_subgroups=None,
export_probabilities=False,
export_probabilities_per_mutation=False,make_plots=False,
sample_reconstruction_plots=False, verbose=False)
```

[210] segmentation files

2.1.5.2.3پارامترهای اصلی

Parameter	Variable Type	Parameter Description
samples	String	Path to the input somatic mutations file (if using segmentation file/mutational matrix) or input folder (mutation calling file/s).
output	String	Path to the output folder.
input_type	String	Three accepted input types: • "vcf": if using mutation calling file/s (VCF, MAF, simple text file) as input • "seg:TYPE": if using a segmentation file as input. Please check the required format at https://github.com/AlexandrovLab/SigProfilerMatrixGenerator#copy-number-matrix-generation. • "matrix": if using a mutational matrix as input The default value is "matrix".
context_type	String	Required context type if `input_type` is "vcf". `context_type` takes which context type of the input data is considered for assignment. Valid options include "96", "288", "1536", "DINUC", and "ID". The default value is "96".
cosmic_version	Float	Defines the version of the COSMIC reference signatures. Takes a positive float among 1, 2, 3, 3.1, 3.2, 3.3, and 3.4. The default value is 3.4.
exome	Boolean	Defines if the exome renormalized COSMIC signatures will be used. The default value is False.
genome_build	String	The reference genome build, used for select the appropriate version of the COSMIC reference signatures, as well as processing the mutation calling file/s. Supported genomes include "GRCh37", "GRCh38", "mm9", "mm10" and "rn6". The default value is "GRCh37". If the selected genome is not in the supported list, the default genome will be used.
signature_database	String	Path to the input set of known mutational signatures (only in case that COSMIC reference signatures are not used), a tab delimited file that contains the signature matrix where the rows are mutation types and columns are signature IDs.
exclude_signature_subgroups	List	Removes the signatures corresponding to specific subtypes to improve refitting (only available when using default COSMIC reference signatures). The usage is explained below. The default value is None, which corresponds to use all COSMIC signatures.
export_probabilities	Boolean	Defines if the probability matrix per mutational context for all samples is created. The default value is True.
export_probabilities_per_mutation	Boolean	Defines if the probability matrices per mutation for all samples are created. Only available when `input_type` is "vcf". The default value is False.
make_plots	Boolean	Toggle on and off for making and saving plots. The default value is True.

جدول 3. پارامترهای اصلی کد اجرای SigProfilerAssignment

2.1.5.2.4زیر گروه‌های امضا

هنگام استفاده از امضاهای مرجع COSMIC، برخی از زیر گروه‌های امضا را می‌توان حذف کرد تا بازسازی آنالیزها بهبود یابد. برای استفاده از این ویژگی، پارامتر **exclude_signature_subgroups** باید به دنبال دستور زیر اضافه شود:

```
exclude_signature_subgroups = ['MMR_deficiency_signatures',
                               'POL_deficiency_signatures',
                               'HR_deficiency_signatures',
                               'BER_deficiency_signatures',
                               'Chemotherapy_signatures',
                               'Immunosuppressants_signatures'
                               'Treatment_signatures'
                               'APOBEC_signatures',
                               'Tobacco_signatures',
                               'UV_signatures',
                               'AA_signatures',
                               'Colibactin_signatures',
                               'Artifact_signatures',
                               'Lymphoid_signatures']
```

لیست کامل زیرگروه‌های امضا در جدول زیر آمده است:

Signature subgroup	SBS signatures excluded	DBS signatures excluded	ID signatures excluded
MMR_deficiency_signatures	6, 14, 15, 20, 21, 26, 44	7, 10	7
POL_deficiency_signatures	10a, 10b, 10c, 10d, 28	3	-
HR_deficiency_signatures	3	13	6
BER_deficiency_signatures	30, 36	-	-
Chemotherapy_signatures	11, 25, 31, 35, 86, 87, 90, 99	5	-
Immunosuppressants_signatures	32	-	-
Treatment_signatures	11, 25, 31, 32, 35, 86, 87, 90, 99	5	-
APOBEC_signatures	2, 13	-	-
Tobacco_signatures	4, 29, 92	2	3
UV_signatures	7a, 7b, 7c, 7d, 38	1	13
AA_signatures	22a, 22b	20	23
Colibactin_signatures	88	-	18

Artifact_signatures	27, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 95	14	-
Lymphoid_signatures	9, 84, 85	-	-

جدول 4. لیست زیرگروه‌های امضاهای جهشی

2.1.5.2.5 مثال‌ها

Using mutation calling files (VCFs) as input:

```
import SigProfilerAssignment as spa
from SigProfilerAssignment import Analyzer as Analyze

Analyze.cosmic_fit(samples=spa.__path__[0]+"/data/tests/vcf_input",
                   output="example_vcf",
                   input_type="vcf",
                   context_type="96",
                   genome_build="GRCh37",
                   cosmic_version=3.4)
```

Using a multi-sample segmentation file as input:

```
import SigProfilerAssignment as spa
from SigProfilerAssignment import Analyzer as Analyze

__path__[0]+"/data/tests/cnv_input/all.breast.ascat.summary.sample.tsv".Analyze.cosmic_fit(samples=spa
                                                                                           ,
                   output="example_sf",
                   input_type="seg:ASCAT_NGS",
                   cosmic_version=3.4,
                   collapse_to_SBS96=False)
```

Using a mutational matrix as input:

```
import SigProfilerAssignment as spa
from SigProfilerAssignment import Analyzer as Analyze

Analyze.cosmic_fit(samples=spa.__path__[0]+"/data/tests/txt_input/sample_matrix_SBS.txt",

                   output="example_mm",
                   input_type="matrix",
                   genome_build="GRCh37",
                   cosmic_version=3.4)
```

2.1.5.2.6 استخراج نوین امضاهای جهشی آنالیز پایین دست

کارکردهای اضافی برای آنالیز پایین‌دستی استخراج de novo امضاهای جهشی نیز به عنوان بخشی از SigProfilerAssignment موجود است، از جمله اختصاص امضاهای جهشی استخراج شده de novo و تجزیه امضاهای de novo با استفاده از مجموعه‌ای از امضاهای شناخته شده. اطلاعات بیشتر را می‌توانید در صفحه ویکی در https://osf.io/mz79v/wiki/5.%20Advanced%20mode بیابید.

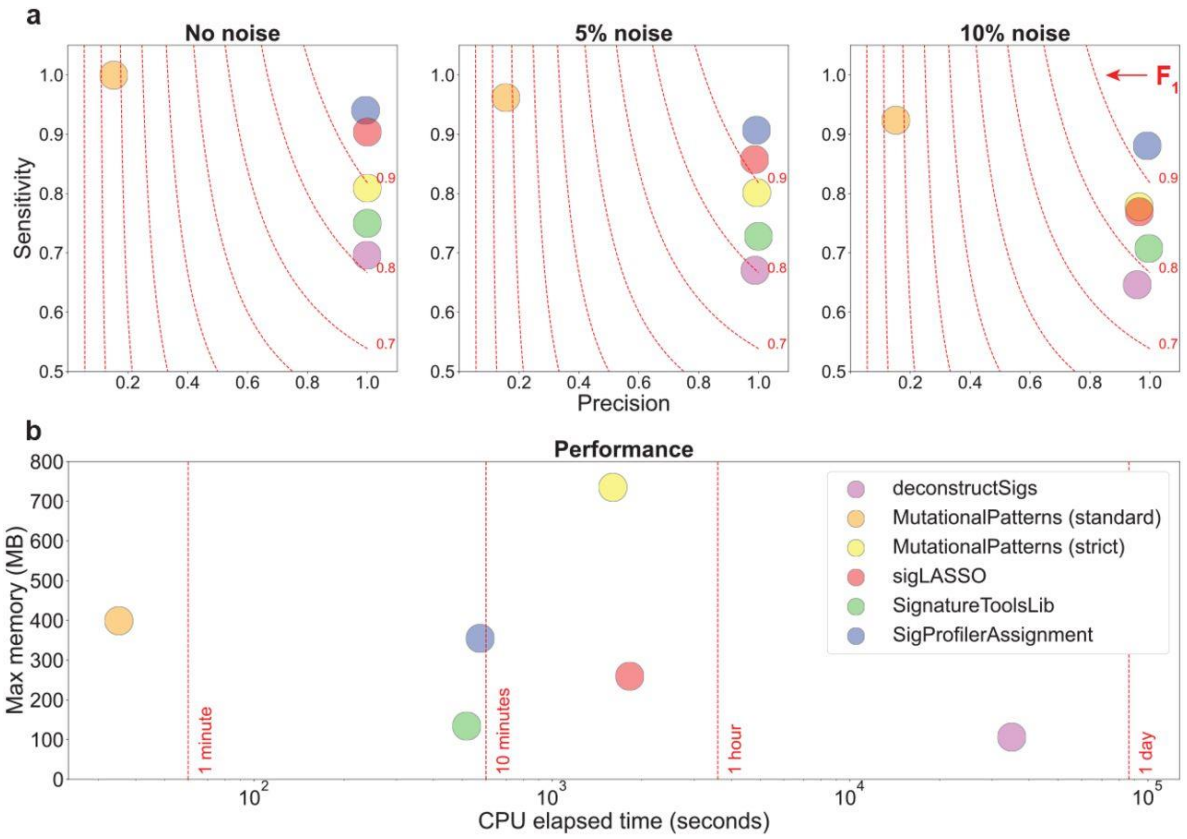
فصل سوم

نتایج

3.1 نتایج

برای ارزیابی عملکرد SigProfilerAssignment و چهار ابزار رایج دیگر در تنظیم مجدد امضاهای جهشی (Rosenthal, McGranahan et al. 2016, Blokzijl, Janssen et al. 2018, Degasperi, Amarante et al. 2020, Li, Crawford et al. 2020, Degasperi, Zou et al. 2022, Manders, Brandsma et al. 2022)، یک معیار مقایسه‌ای با استفاده از یک مجموعه داده مصنوعی مستقل که قبلاً تولید شده بود (Islam, Díaz-Gay et al. 2022) انجام گرفت (**شکل a9 و b**). مجموعه داده شامل الگوهای SBS از 2700 ژنوم سرطان شبیه‌سازی شده، مربوط به 300 تومور از 9 نوع سرطان مختلف است که با استفاده از 21 امضای مرجع COSMIC مختلف تولید شده‌اند. برای تقلید از تنظیم مجدد معمول امضاهای جهشی، مجموعه کامل 79 امضای COSMICv3.3 SBS به عنوان ورودی استفاده شد. فعالیت‌های امضاهای جهشی به‌دست‌آمده توسط هر ابزار با فعالیت‌های یافته‌های عینی مورد استفاده برای تولید مصنوعی این نمونه‌ها مقایسه شد. سه سطح مختلف نویز تصادفی (0٪، 5٪ و 10٪) برای ارزیابی قدرت الگوریتم های مختلف در یک زمینه بیولوژیکی واقعی آزمایش شد. برای ارزیابی دقت تنظیم مجدد امضا، حساسیت[۲۱۱]، ویژگی[۲۱۲] و امتیاز F1 [۲۱۳] محاسبه گردید. علاوه بر این، زمان اجرا و استفاده از حافظه هر ابزار نیز بررسی شد.

[211] sensitivity
[212] specificity
[213] F1 score

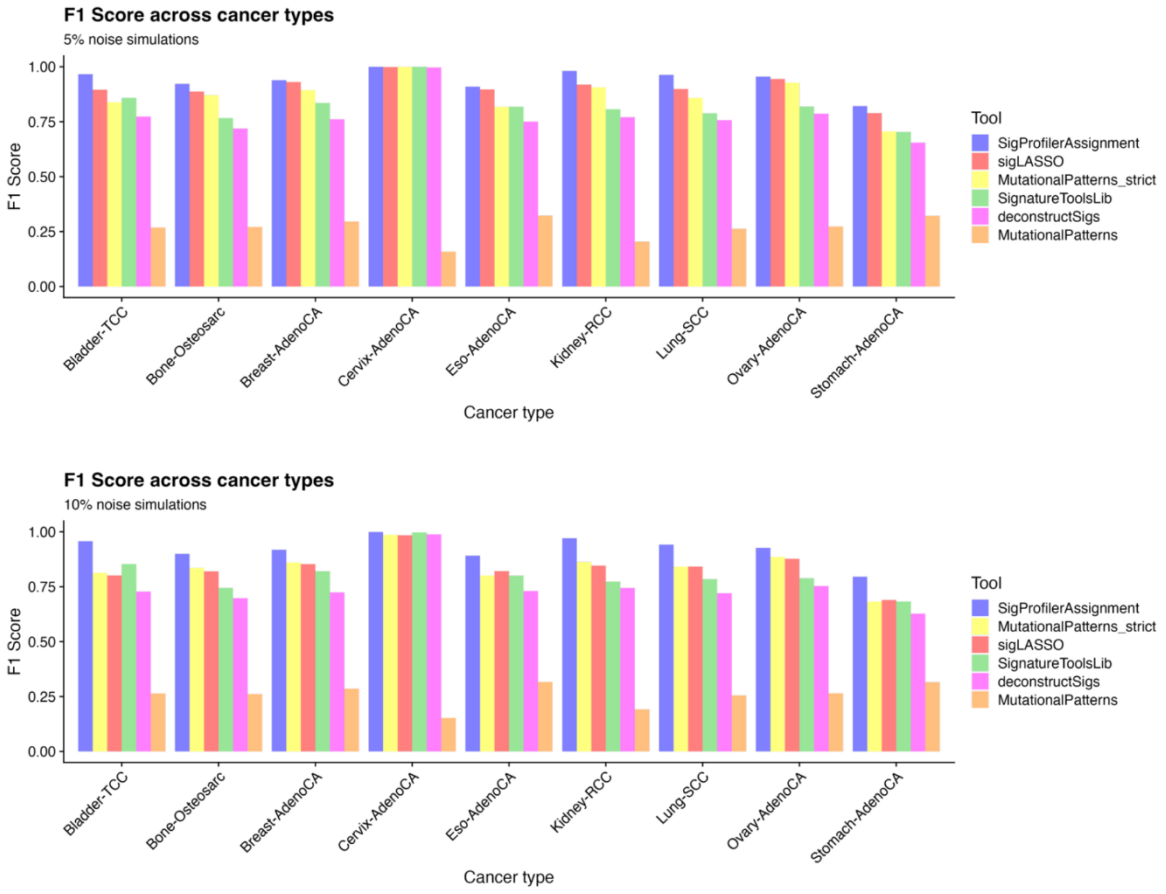


شکل 9. محک زدن دقت[۲۱۴] SigProfilerAssignment و چهار ابزار دیگر برای تخصیص امضاهای جهشی

شکل 9 محک زدن دقت SigProfilerAssignment و چهار ابزار دیگر برای تخصیص امضاهای جهشی را نشان می‌دهد. در قسمت (**a**) هر ابزار با استفاده از 2700 ژنوم سرطان مصنوعی تولید شده با استفاده از 21 امضای جهشی مرجع COSMIC مورد ارزیابی قرار گرفته است. همه امضاهای COSMICv3.3 به عنوان مجموعه ورودی امضاهای جهش شناخته شده استفاده شد. برای ارزیابی دقت (محورهای x)، حساسیت (محورهای y) و امتیازات F1 (میانگین هارمونیک[۲۱۵] دقت و حساسیت؛ خطوط قرمز نقطه‌چین) هر ابزار، از سه سطح مختلف نویز تصادفی غیرسیستماتیک (0٪، 5٪ و 10٪) استفاده شد. قسمت (**b**) محک زدن محاسباتی براساس زمان سپری شده CPU (محور x؛ log-scaled) و حداکثر استفاده از حافظه (محور y) برای هر ابزار را نمایش می‌دهد.

[214] accuracy benchmarking
[215] harmonic mean

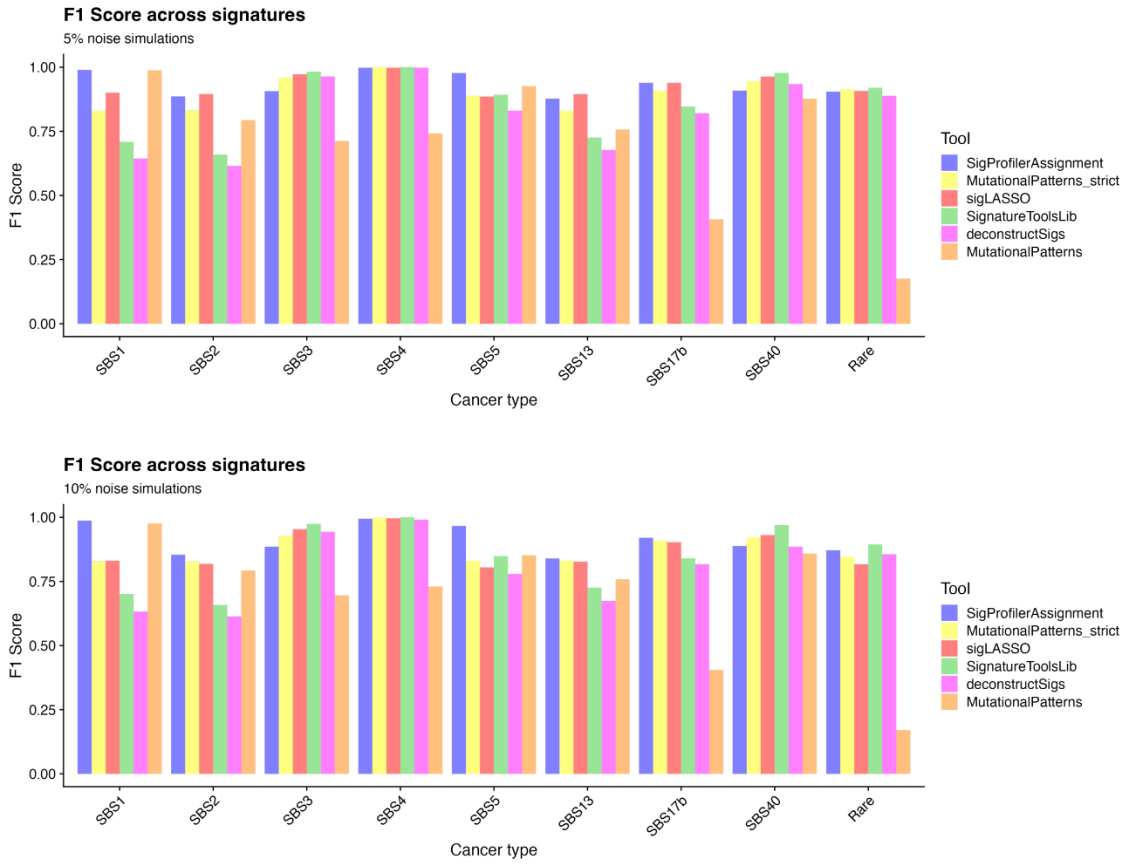
چهارچوب محک زدن [۲۱۶] مصنوعی مقاله نشان داد که SigProfilerAssignment از تمام روش‌های دیگر برای سطوح نویز بررسی شده بهتر عمل می‌کند (**شکل 9 a**). برای 10٪ نویز تصادفی، فقط SigProfilerAssignment امتیاز F1 score >0.90 را به دست آورد. در همه موارد، SigProfilerAssignment دقت بالایی را نشان داد در حالی که حساسیت بهبود یافته را نیز در مقایسه با سایر رویکردها به نمایش گذاشت (**شکل 9 a**)، در کنار عملکرد بالای ثابت در انواع سرطان (**شکل 10**) و اکثر امضاهای جهشی (**شکل 11**).



شکل 10. محک زدن نوع خاص بافت SigProfilerAssignment و چهار ابزار دیگر برای تخصیص امضاهای جهشی

[216] benchmarking

در **شکل 10** امتیاز F1 (میانگین هماهنگی دقت و حساسیت[۲۱۷]) برای 9 نوع سرطان موجود در مجموعه داده مصنوعی (300 ژنوم شبیه‌سازی شده[۲۱۸] برای هر نوع سرطان) برای ارزیابی دقت انتساب امضا در سراسر شبیه‌سازی‌ها با نویز تصادفی غیر سیستماتیک[۲۱۹] استفاده شده است.



شکل 11. محک زدن مخصوص امضای SigProfilerAssignment و چهار ابزار دیگر برای تخصیص امضاهای جهشی

[217] harmonic mean of precision and sensitivity

[218] simulated genomes

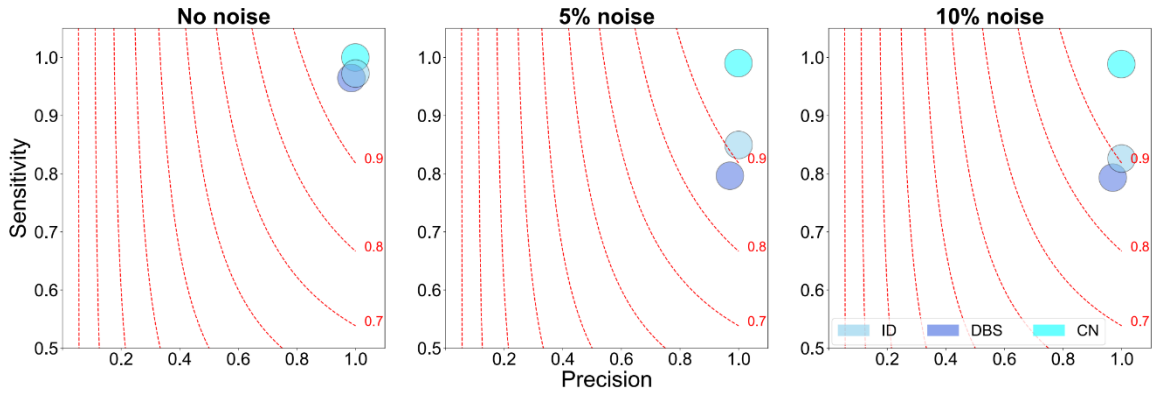
[219] non-systematic random noise

در **شکل 11** امتیاز F1 (میانگین هماهنگی دقت و حساسیت) برای هشت امضای جهش رایج در سراسر فعالیت‌های یافته‌های عینی و میانگین 13 امضای نادر باقی‌مانده برای ارزیابی دقت تخصیص امضا در سراسر شبیه‌سازی‌ها با با نویز تصادفی غیر سیستماتیک استفاده شده است.

از نظر عملکرد محاسباتی، SigProfilerAssignment 2700 نمونه را در 9.6 دقیقه (0.21 اینچ در هر نمونه) پردازش کرد. فقط حالت استاندارد MutationalPatterns نتایج قابل ملاحظه‌ای سریع‌تر تولید می‌کند (**شکل 9 b**). با این حال، حالت استاندارد MutationalPatterns عملکردی کمتر از حد بهینه را نشان داد، البته با کاهش قابل توجهی در دقت برای تمام سطوح نویز، احتمالاً به دلیل بیش برازش داده‌های ورودی (**شکل 9 a**) (Blokzijl, Janssen et al. 2018). این موضوع در جدیدترین نسخه MutationalPatterns با افزودن یک حالت سخت[220] (Manders, Brandsma et al. 2022)، البته با هزینه عملکرد محاسباتی قابل توجه مورد بررسی قرار گرفته است (**شکل 9 b**). سایر رویکردها با اجرای جریمه‌های مختلف بر اساس خطای L1 [221] (به عنوان مثال در sigLASSO) (Li, Crawford et al. 2020) یا خطای مجموع مربع[222] (به عنوان مثال در deconstructSigs) (Rosenthal, McGranahan et al. 2016) و فیلترهای post-hoc بر اساس درصد تعداد کل جهش‌های نسبت داده شده به یک امضای معین (به عنوان مثال در deconstructSigs و SignatureToolsLib) (Rosenthal, McGranahan et al. 2016, Degasperi, Zou et al. 2022) (**جدول 2**) بیش برازش را محدود می‌کنند. هیچ حافظه قابل توجه مورد نیازی برای هیچ یک از ابزارها مشاهده نشد (**شکل 9 b**).

تجزیه و تحلیل مجموعه داده‌های چهارچوب‌های مشابه برای امضاهای جهشی DBS، ID و CN (داده‌های Supplementary) نشان داد که SigProfilerAssignment دقت و حساسیت بالایی را با امتیاز F1 score > 0.85 برای تمام سطوح نویز ارزیابی‌شده، نشان می‌دهد (**شکل 12**).

[220] strict mode
[221] L1 error
[222] sum-squared error



شکل 12. محک زدن SigProfilerAssignment در انواع مختلف جهش

شکل 12 محک زدن SigProfilerAssignment در انواع مختلف جهش را نمایش می‌دهد. نمونه‌های DBS، ID و CN مصنوعی برای آزمایش دقت تخصیص امضای SigProfilerAssignment با استفاده از امضاهای COSMICv3.3 به عنوان امضاهای جهش شناخته شده ورودی استفاده شد. سه سطح مختلف نویز تصادفی غیر سیستماتیک (0، 5، و 10٪) برای ارزیابی دقت (محورهای x)، حساسیت (محورهای y)، و امتیازات F1 (میانگین هماهنگی دقت و حساسیت؛ خطوط خط چین قرمز رنگ) هر ابزار استفاده شد.

فصل چهارم
بحث و نتیجه گیری

4.1 بحث

تخصیص امضاهای جهشی به نمونه‌های فردی فرصتی را برای شناسایی فرآیندهای مسئول جهش‌های سوماتیکی به صورت نمونه به نمونه[۲۲۳] فراهم می‌کند. با در نظر گرفتن چهارچوب محک زدن مصنوعی ما، SigProfilerAssignment به عنوان دقیق‌ترین و حساس‌ترین ابزار در عین حفظ عملکرد محاسباتی بالا و ارائه قابلیت‌های جدید برجسته می‌شود. تا آنجا که ما می‌دانیم، SigProfilerAssignment اولین ابزار محاسباتی برای تخصیص احتمالات امضا به جهش‌های فردی است که می‌تواند به کشف فرآیندهای جهشی مسئول تغییرات ژنومی محرک خاص منجر به تکامل تومور کمک کند. SigProfilerAssignment همچنین اولین ابزاری است که از تخصیص امضاهای copy number اخیراً توسعه یافته پشتیبانی می‌کند (Steele, Abbasi et al. 2022)، که پیش‌بینی کننده خوبی برای بقای بالینی[۲۲۴] هستند (Drews, Hernando et al. 2022, Steele, Abbasi et al. 2022).

[223] sample-by-sample
[224] copy number

به طور خلاصه، SigProfilerAssignment یک بسته محاسباتی جدید و یک رابط آنلاین قابل دسترس برای تخصیص دقیق امضاهای جهشی شناخته شده به یک سرطان فردی و جهش‌های جسمی فردی ارائه می‌کند، بنابراین، کاربران را قادر می‌سازد تا فرآیندهای جهشی را در ژنوم سرطان مشخص کنند.

مراجع

Jumper, J., et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.

" COSMIC. Signatures of mutational processes in human cancer. http://cancer.sanger.ac.uk/cosmic/signatures (27 April 2017, date last accessed).".

(2020). "Pan-cancer analysis of whole genomes." Nature **578**(7793): 82-93.

Abadi, M., et al. (2016). "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467.

Alexandrov, L. B., et al. (2015). "Clock-like mutational processes in human somatic cells." Nature genetics **47**(12): 1402-1407.

Alexandrov, L. B., et al. (2020). "The repertoire of mutational signatures in human cancer." Nature **578**(7793): 94-101.

Alexandrov, L. B., et al. (2015). "A mutational signature in gastric cancer suggests therapeutic strategies." Nature communications **6**(1): 8683.

Alexandrov, L. B., et al. (2013). "Signatures of mutational processes in human cancer." Nature **500**(7463): 415-421.

Alexandrov, L. B., et al. (2013). "Deciphering signatures of mutational processes operative in human cancer." Cell reports **3**(1): 246-259.

Alexandrov, L. B. and M. R. Stratton (2014). "Mutational signatures: the patterns of somatic mutations hidden in cancer genomes." Current opinion in genetics & development **24**: 52-60.

Ang, A. M. S. and N. Gillis (2019). "Algorithms and comparisons of nonnegative matrix factorizations with volume regularization for hyperspectral unmixing." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(12): 4843-4853.

Ardin, M., et al. (2016). "MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes." BMC bioinformatics **17**: 1-10.

Attolini, C. S. O. and F. Michor (2009). "Evolutionary theory of cancer." Annals of the New York Academy of Sciences **1168**(1): 23-51.

Bauer, H., et al. (1938). "X-ray induced chromosomal alterations in Drosophila melanogaster." Genetics **23**(6): 610.

Bayard, Q., et al. (2018). "Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress." Nature communications **9**(1): 5235.

Beerenwinkel, N., et al. (2007). "Genetic progression and the waiting time to cancer." PLoS computational biology **3**(11): e225.

Bergstrom, E. N., et al. (2019). "SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events." BMC genomics **20**(1): 1-12.

Blokzijl, F., et al. (2018). "MutationalPatterns: comprehensive genome-wide analysis of mutational processes." Genome medicine **10**: 1-11.

Blokzijl, F., et al. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med, BioMed Central.

Borchers, H. W. (2022). "Pracma: Practical numerical math functions (2.4. 2)." from https://CRAN.R-project.org/package=pracma last accessed.

Brunet, J.-P., et al. (2004). "Metagenes and molecular pattern discovery using matrix factorization." Proceedings of the National Academy of Sciences **101**(12): 4164-4169.

Carpenter, B., et al. (2017). "Stan: A probabilistic programming language." Journal of statistical software **76**.

Carter, S. L., et al. (2012). "Absolute quantification of somatic DNA alterations in human cancer." Nature biotechnology **30**(5): 413-421.

Cartolano, M., et al. (2020). "CaMuS: simultaneous fitting and de novo imputation of cancer mutational signature." Scientific reports **10**(1): 19316.

Ceccaldi, R., et al. (2015). "Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair." Nature **518**(7538): 258-262.

Craig, M. D. (1994). "Minimum-volume transforms for remotely sensed data." IEEE Transactions on Geoscience and Remote Sensing **32**(3): 542-552.

Davies, H., et al. (2017). "HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures." Nature medicine **23**(4): 517-525.

Davies, H., et al. (2017). "Whole-genome sequencing reveals breast cancers with mismatch repair deficiency." Cancer research **77**(18): 4755-4762.

Degasperi, A., et al. (2020). "A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies." Nature cancer **1**(2): 249-263.

Degasperi, A., et al. (2022). "Substitution mutational signatures in whole-genome–sequenced cancers in the UK population." Science **376**(6591): abl9283.

Dempster, A. P., et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society: series B (methodological) **39**(1): 1-22.

Díaz-Gay, M., et al. (2023). "Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment." Bioinformatics **39**(12): btad756.

Díaz-Gay, M., et al. (2018). "Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples." BMC bioinformatics **19**(1): 1-6.

Drews, R. M., et al. (2022). "A pan-cancer compendium of chromosomal instability." Nature **606**(7916): 976-983.

Fantini, D., et al. (2020). "MutSignatures: an R package for extraction and analysis of cancer mutational signatures." Scientific Reports **10**(1): 18217.

Favero, F., et al. (2015). "Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data." Annals of Oncology **26**(1): 64-70.

Févotte, C. and A. T. Cemgil (2009). Nonnegative matrix factorizations as probabilistic inference in composite models. 2009 17th European Signal Processing Conference, IEEE.

Fischer, A., et al. (2013). "EMu: probabilistic inference of mutational processes and their localization in the cancer genome." Genome biology **14**(4): 1-10.

Fox, E. J., et al. (2016). "Exploring the implications of distinct mutational signatures and mutation rates in aging and cancer." Genome medicine **8**(1): 1-3.

Friedman, J., et al. (2010). "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software **33**(1): 1.

Gaujoux, R. and C. Seoighe (2010). "A flexible R package for nonnegative matrix factorization." BMC bioinformatics **11**(1): 1-9.

Gehring, J. S., et al. (2015). "SomaticSignatures: inferring mutational signatures from single-nucleotide variants." Bioinformatics **31**(22): 3673-3675.

Georgeson, P., et al. (2022). "Identifying colorectal cancer caused by biallelic MUTYH pathogenic variants using tumor mutational signatures." Nature communications **13**(1): 3254.

Georgeson, P., et al. (2021). "Evaluating the utility of tumour mutational signatures for identifying hereditary colorectal cancer and polyposis syndrome carriers." Gut **70**(11): 2138-2149.

Gori, K. and A. Baez-Ortega (2018). "sigfit: flexible Bayesian inference of mutational signatures." bioRxiv: 372896.

Gorski, J., et al. (2007). "Biconvex sets and optimization with biconvex functions: a survey and extensions." Mathematical methods of operations research **66**: 373-407.

Govindan, R., et al. (2012). "Genomic landscape of non-small cell lung cancer in smokers and never-smokers." Cell **150**(6): 1121-1134.

Grolleman, J. E., et al. (2019). "Mutational signature analysis reveals NTHL1 deficiency to cause a multi-tumor phenotype." Cancer cell **35**(2): 256-266. e255.

Haradhvala, N. J., et al. (2016). "Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair." Cell **164**(3): 538-549.

Harris, R. S. (2013). "Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications." Genome medicine **5**: 1-3.

Hastie, T., et al. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer.

Helleday, T., et al. (2014). "Mechanisms underlying mutational signatures in human cancers." Nature reviews genetics **15**(9): 585-598.

Hillman, R. T., et al. (2018). "Genomic rearrangement signatures and clinical outcomes in high-grade serous ovarian cancer." JNCI: Journal of the National Cancer Institute **110**(3): 265-272.

Hoang, P. H., et al. (2019). "Mutational processes contributing to the development of multiple myeloma." Blood cancer journal **9**(8): 60.

Howard, B. D. and I. Tessman (1964). "Identification of the altered bases in mutated single-stranded DNA: III. Mutagenesis by ultraviolet light." Journal of molecular biology **9**(2): 372-375.

Huang, X., et al. (2018). "Detecting presence of mutational signatures in cancer with confidence." Bioinformatics **34**(2): 330-337.

Hunter, C., et al. (2006). "A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy." Cancer research **66**(8): 3987-3991.

Islam, S. A., et al. (2022). "Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor." Cell Genomics **2**(11).

Jin, H., et al. (2024). "Accurate and sensitive mutational signature analysis with MuSiCal." Nature genetics: 1-12.

Kamp, J., et al. (2020). "BRCA1-associated structural variations are a consequence of polymerase theta-mediated end-joining." Nature communications **11**(1): 3615.

Kasar, S., et al. (2015). "Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution." Nature communications **6**(1): 8866.

Khandekar, A., et al. (2023). "Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator." bioRxiv: 2023.2002. 2003.527015.

Kiefer, J. (1953). "Sequential minimax search for a maximum." Proceedings of the American mathematical society **4**(3): 502-506.

Koh, G., et al. (2021). "Mutational signatures: emerging concepts, caveats and clinical applications." Nature reviews cancer **21**(10): 619-637.

Koh, G., et al. (2020). "Mutational signatures: experimental design and analytical framework." Genome biology **21**: 1-13.

Kucab, J. E., et al. (2019). "A compendium of mutational signatures of environmental agents." Cell **177**(4): 821-836. e816.

Lawson, C. L. and R. J. Hanson (1995). Solving least squares problems, SIAM.

Ledford, H. (2022). "Trove of tumour genomes offers clues to cancer origins." Nature **604**(7907): 609-609.

Lee-Six, H., et al. (2019). "The landscape of somatic mutation in normal colorectal epithelial cells." Nature **574**(7779): 532-537.

Lee, D. D. and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." Nature **401**(6755): 788-791.

Lee, J., et al. (2018). "Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures." Nucleic acids research **46**(W1): W102-W108.

Leplat, V., et al. (2020). "Blind audio source separation with minimum-volume beta-divergence NMF." IEEE Transactions on Signal Processing **68**: 3400-3410.

Letouzé, E., et al. (2017). "Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis." Nature communications **8**(1): 1315.

Levatić, J., et al. (2022). "Mutational signatures are markers of drug sensitivity of cancer cells." Nature communications **13**(1): 2926.

Levenberg, K. (1944). "A method for the solution of certain non-linear problems in least squares." Quarterly of applied mathematics **2**(2): 164-168.

Li, S., et al. (2020). "Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood." Nature communications **11**(1): 3575.

Li, X., et al. (2016). "Distinct subtypes of gastric cancer defined by molecular characterization include novel mutational signatures with prognostic capability." Cancer research **76**(7): 1724-1732.

Li, Y., et al. (2020). "Patterns of somatic structural variation in human cancer genomes." Nature **578**(7793): 112-121.

Lin, X. and P. C. Boutros (2020). "Optimization and expansion of non-negative matrix factorization." BMC bioinformatics **21**(1): 1-10.

Ling, R. F. (1977). Solving least squares problems, JSTOR.

Macintyre, G., et al. (2018). "Copy number signatures and mutational processes in ovarian carcinoma." Nature genetics **50**(9): 1262-1270.

Manders, F., et al. (2022). "MutationalPatterns: the one stop shop for the analysis of mutational processes." BMC genomics **23**(1): 134.

Mateos-Gomez, P. A., et al. (2015). "Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination." Nature **518**(7538): 254-257.

Mayakonda, A., et al. (2018). "Maftools: efficient and comprehensive analysis of somatic variants in cancer." Genome research **28**(11): 1747-1756.

Miao, L. and H. Qi (2007). "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization." IEEE Transactions on Geoscience and Remote Sensing **45**(3): 765-777.

Morganella, S., et al. (2016). "The topography of mutational processes in breast cancer genomes." Nature communications **7**(1): 11383.

MULLER, H. (1932). Further studies on the nature and causes of gene mutations. Proceedings of the 6th International Congress of Genetics, 1932.

Nebgen, B. T., et al. (2021). "A neural network for determination of latent dimensionality in non-negative matrix factorization." Machine Learning: Science and Technology **2**(2): 025012.

Nik-Zainal, S., et al. (2012). "Mutational processes molding the genomes of 21 breast cancers." Cell **149**(5): 979-993.

Nik-Zainal, S., et al. (2016). "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." Nature **534**(7605): 47-54.

Omichessan, H., et al. (2019). "Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance." PloS one **14**(9): e0221235.

Papaemmanuil, E., et al. (2014). "RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia." Nature genetics **46**(2): 116-125.

Petljak, M., et al. (2019). "Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis." Cell **176**(6): 1282-1294. e1220.

Pfeifer, G. P. (2010). "Environmental exposures and mutational patterns of cancer genomes." Genome medicine **2**: 1-4.

Pfeifer, G. P., et al. (2002). "Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers." Oncogene **21**(48): 7435-7451.

Pfeifer, G. P., et al. (2005). "Mutations induced by ultraviolet light." Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **571**(1-2): 19-31.

Pich, O., et al. (2019). "The mutational footprints of cancer therapies." Nature genetics **51**(12): 1732-1740.

Poon, S., et al. (2015). Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med 7: 38.

Poon, S. L., et al. (2014). "Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention." Genome medicine **6**: 1-14.

Roberts, S. A. and D. A. Gordenin (2014). "Hypermutation in human cancer genomes: footprints and mechanisms." Nature reviews cancer **14**(12): 786-800.

Rosales, R. A., et al. (2017). "signeR: an empirical Bayesian approach to mutational signature discovery." Bioinformatics **33**(1): 8-16.

Rosenthal, R., et al. (2016). "DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution." Genome biology **17**(1): 1-11.

Rubin, A. F. and P. Green (2009). "Mutation patterns in cancer genomes." Proceedings of the National Academy of Sciences **106**(51): 21766-21770.

Sax, K. (1938). "Chromosome aberrations induced by X-rays." Genetics **23**(5): 494.

Schulze, K., et al. (2015). "Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets." Nature genetics **47**(5): 505-511.

Schwarz, G. (1978). "Estimating the dimension of a model." The annals of statistics: 461-464.

Secrier, M., et al. (2016). "Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance." Nature genetics **48**(10): 1131-1141.

Setlow, R. and W. Carrier (1966). "Pyrimidine dimers in ultraviolet-irradiated DNA's." Journal of molecular biology **17**(1): 237-254.

Shale, C., et al. (2022). "Unscrambling cancer genomes via integrated analysis of structural variation and copy number." Cell Genomics **2**(4).

Shen, R. and V. E. Seshan (2016). "FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing." Nucleic acids research **44**(16): e131-e131.

Stacklies, W., et al. (2007). "pcaMethods—a bioconductor package providing PCA methods for incomplete data." Bioinformatics **23**(9): 1164-1167.

Steele, C. D., et al. (2022). "Signatures of copy number alterations in human cancer." Nature **606**(7916): 984-991.

Steele, C. D., et al. (2019). "Undifferentiated sarcomas develop through distinct evolutionary pathways." Cancer Cell **35**(3): 441-456. e448.

Stratton, M. R., et al. (2009). "The cancer genome." Nature **458**(7239): 719-724.

Suri, P. and N. R. Roy (2017). Comparison between LDA & NMF for event-detection from large text stream data. 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), IEEE.

Tan, V. Y. and C. Févotte (2012). "Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence." IEEE transactions on pattern analysis and machine intelligence **35**(7): 1592-1605.

Tate, J. G., et al. (2019). "COSMIC: the catalogue of somatic mutations in cancer." Nucleic acids research **47**(D1): D941-D947.

Taylor-Weiner, A., et al. (2019). "Scaling computational genomics to millions of individuals with GPUs." Genome biology **20**(1): 1-5.

Thorndike, R. L. (1953). "Who belongs in the family?" Psychometrika **18**(4): 267-276.

Van Loo, P., et al. (2010). "Allele-specific copy number analysis of tumors." Proceedings of the National Academy of Sciences **107**(39): 16910-16915.

Virtanen, P., et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python." Nature methods **17**(3): 261-272.

Vogelstein, B., et al. (2013). "Cancer genome landscapes." science **339**(6127): 1546-1558.

Vöhringer, H., et al. (2021). "Learning mutational signatures and their multidimensional genomic properties with TensorSignatures." Nature communications **12**(1): 3628.

Wang, S., et al. (2021). "Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes." Plos Genetics **17**(5): e1009557.

Yates, L. R. and P. J. Campbell (2012). "Evolution of the cancer genome." Nature reviews genetics **13**(11): 795-806.

Zou, X., et al. (2018). "Validating the concept of mutational signatures with isogenic cell models." Nature communications **9**(1): 1744.