Article: EMu: probabilistic inference of mutational processes and their localization in the cancer genome

Journal: BMC genome biology

Authors: Fischer, A. et al.

Year: 2013

Citiations: 42

link: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r39

هدف:

- استخراج فرایندهای جهش‌زا ابتدایی و امضای آن‌ها از داده‌های جهش سرطانی کل ژنوم
- محدود کردن فعالیت فرایند به یک ناحیه در ژنوم‌های سرطان

انواع مختلفی از فرآیندهای فیزیکی ، شیمیایی و بیولوژیکی و همچنین غیرفعال شدن مکانیسم تعمیر mismatch منجر به جهش در سرطان می شود. ما در اینجا وظیفه شناسایی فرآیندهای جهش‌زا فعال را در طول توسعه سرطان در نظر می گیریم. ما فرض می‌کنیم که این فرایندها مجزا هستند ، به این ترتیب، هر فرآیند علامت مشخصه متفاوتی را بر روی ژنوم سرطان باقی می‌گذارد. با توجه به داده‌های توالی از تومورهای متعدد ، که در آن فرآیندهای جهش‌زا در ابعاد مختلف فعال هستند ، هدف ما استنباط تعداد فرآیندهای جهش‌زا اولیه ، امضاهای آنها و سهم هر فرآیند تا طیف جهش‌های مشاهده شده در هر تومور است. شناسایی فرایندهای جهش‌زا گامی در جهت شناخت مکانیسم های ایجادکننده سرطان است.

محدودیت روش‌های قبلی:

1. فرصت وقوع جهش‌ها در یک توالی ژنتیکی مشخص صریحاً مورد توجه قرار نگرفته است. این مهم است زیرا نتیجه مشاهده شده از یک فرایند جهش‌زا به محتویات توالی‌ای که در آن عمل می شود بستگی دارد. به عنوان مثال ، فرآیندی که انتقال‌های C> T را در محل‌های CpG تولید می‌کند، ممکن است با فعالیت یکنواخت در سراسر ژنوم انجام دهد ، اما بیشتر در مناطقی با چگالی بیشتر از این دونوکلئوتیدها مشاهده می‌شود.

۲. در حالی که با توجه به فرآیندهای جهش‌زا در کل ژنوم ، مطالعات قبلی عملکرد فرآیندهای مختلف را در مناطق کوچک‌تر ژنوم را بررسی نکرده اند. غیریکنواختی در میزان جهش در يك ژنوم سرطان مشاهده شده است ، که با تفاوت و bias در عملکرد مکانیسم‌هاي ترميم DNA و[5] با توزیع علائم اصلاح histone همراه است [14] ، نشان مي‌دهد که فرآیند هاي جهش‌زا ممکن است در مناطق مختلف ژنوم متفاوت عمل کنند. تغییر در فرصت جهش‌زایی ، که در بالا ذکر شد ، به ویژه در مناطق کوچک ژنوم از اهمیت ویژه‌ای برخوردار است. Copy number variation ، که می تواند در سرطان شایع باشد [15]، به طور قابل ملاحظه ای فرصت جهش‌زایی در مناطقی از تغییر پلوئیدی (ploidy) (تعداد دسته‌های کروموزوم‌ها) را تغییر می‌دهد ، در حالی که مناطق مختلف ژنوم انسان ممکن است از نظر بنیادی در محتوای داخلی توالی آنها متفاوت باشد [16].

۳. روشهای قبلی صراحتاً روش تصادفی که جهش‌ها در ژنوم وارد می‌شوند را در نظر نگرفته اند. تصادفی بودن ، همراه با ماهیت گسسته بودن جهش ، منجر به الگوهای خاص noise در طیف‌های جهش مشاهده شده می‌شود. هنگامی که زیرمجموعه های کوچکتر ژنوم مورد تجزیه و تحلیل قرار می گیرند ، این noise تقویت می‌شود ، و این یک راه‌حل احتمالی صریح از داده‌ها ضروری است.

روش جدید:

براساس expectation-maximization (EM)

برای استنتاج تعداد فرایندهای جهش‌زا ابتدایی و طیف‌های آنها از داده‌های توالی سرطان.

با امکان در نظر گرفتن bias ها در فرصت جهش‌زایی و داده‌های noise

با استفاده از نتایج شبیه‌سازی ، ما نشان می‌دهیم که اضافه کردن فرصت جهش‌زایی در شناسایی تعداد صحیح فرآیندهای جهش‌زا در این مدل بسیار مهم است.

تعداد فرایندها با بیشترین احتمال با Bayesian information criterion(BIC) می‌تواند بدست بیاید. تخمین خطای معنی دار از امضاهای جهش‌زا می‌تواند به روش تحلیلی یا عددی با روش های Markov chain Monte Carlo (MCMC) استنباط شود. با توسعه منطقی الگوریتم یادگیری-پارامتر به زیر مجموعه‌های کوچکتر یک ژنوم سرطان ، می‌توان فعالیت موضعی فرآیندهای مختلف را اندازه گرفت. ما از داده تفسیر (annotation) از

پروژه ENCODE [20] استفاده می کنیم تا تغییرات در فرآیندهای جهش‌زا مربوط به حالت chromatin را بررسی کنیم. پروژه ENCODE بخش هایی از ژنوم انسان را در حالتهای مختلف عملکردی کروماتین فراهم می‌کند [21] ، که عمدتا بر اساس توزیع علائم اصلاح histone است. در سرطان ، یک مطالعه جدید ارتباطات قوی جهش‌های جسمی را با برخی از این تغییرات histone ، به ویژه H3K9me3 مشخص کرده است [14]. ما در اینجا شواهد آماری را ارائه می‌دهیم که نشان می‌دهد فرآیندهای جهش‌زا در مناطق با حالت chromatin مختلف در سرطان سینه متفاوت عمل می‌کنند. به طور خاص ، یکی از فرآیندهای جهش ، تولید کننده جهش‌های C> T عمدتا در محل‌های CpG ، در مناطق promoter به شدت کمرنگ است و در مناطق heterochromatin تقویت می‌شود. این نشان می‌دهد که این فرایند با متیلاسیون DNA ارتباط دارد ، همانطور که در مورد حذف خود به خودی آمینواسید (deamination) وجود دارد [22].

نتایج

ارزیابی با داده‌های شبیه‌سازی شده

تعداد نمونه‌های سرطانی: $M \in \{10,20,50,100,500,1000\}$

تعداد فرایندهای جهش‌زا: $n \in \{5,10,15\}$

در هر حالت ، افزایش اندازه نمونه ، توانایی تمایز بین فرایندهای جهش را بهبود بخشید. در اکثر موارد ، پنج فرآیند جهش از بیست نمونه تومور قابل تشخیص است. در فرایندهای شناسایی شده، طیف جهش با میزان واقعی آن ارتباط قوی دارد. استفاده از BIC در انتخاب مدل، تعداد فرایندها هیچ‌وقت overestimated نشده است. هرچند با نادیده گرفتن فرصت جهش‌زایی باعث شد تعداد فرایند با BIC بیشتر از مقدار قابل انتظار شود.

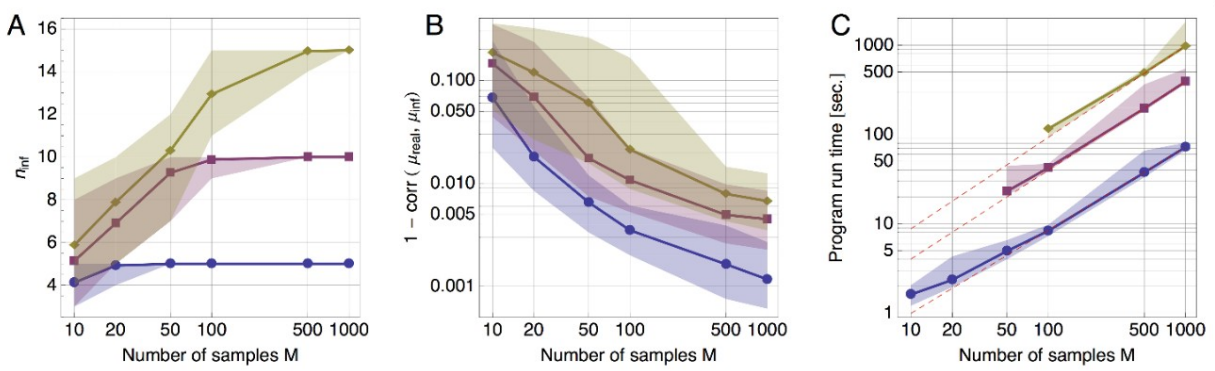


Figure 1 Performance. Results of the performance test of the EM method on simulated mutation data. Each simulated tumor belongs to one of up to three different cancer types, where in each type there are five independent processes active. The total number of processes per data set is thus 5 (blue), 10 (red) or 15 (yellow). Here, we show the scaling of different observables as a function of sample size (calculated over 50 replicates for each combination of *M* and *n*). **(A)** The number of processes present in the data is determined via the BIC. Shown are the median (line), the smallest and the largest (shaded area) number of inferred processes. **(B)** The correlation between the real and the inferred mutation spectra (the difference from 1 is plotted). **(C)** The time until completion of the inference program scales approximately linearly with *M* (for constant *n*; the fits above correspond to 0.93, 0.99 and 1.02). (B and C show the median with the 10% and 90% quantiles.). BIC: Bayesian information criterion; EM: expectation-maximization.

فرایندهای جهش‌زا در سرطان سینه:

در سراسر ژنوم انسان ، فرصت جهش‌زایی بین کانالها به اندازه 17-fold تغییر می‌کند ، و این موضوع را به یک عامل مهم در استخراج طیف‌های جهش‌زا تبدیل می‌کند. در مجموعه داده‌های سرطان سینه ، این با وجود تغییرات copy number در مقیاس بزرگ همراه است. این منجر به طیفی از فرصت جهش‌زایی می‌شود که به نوع تومور بستگی دارد.

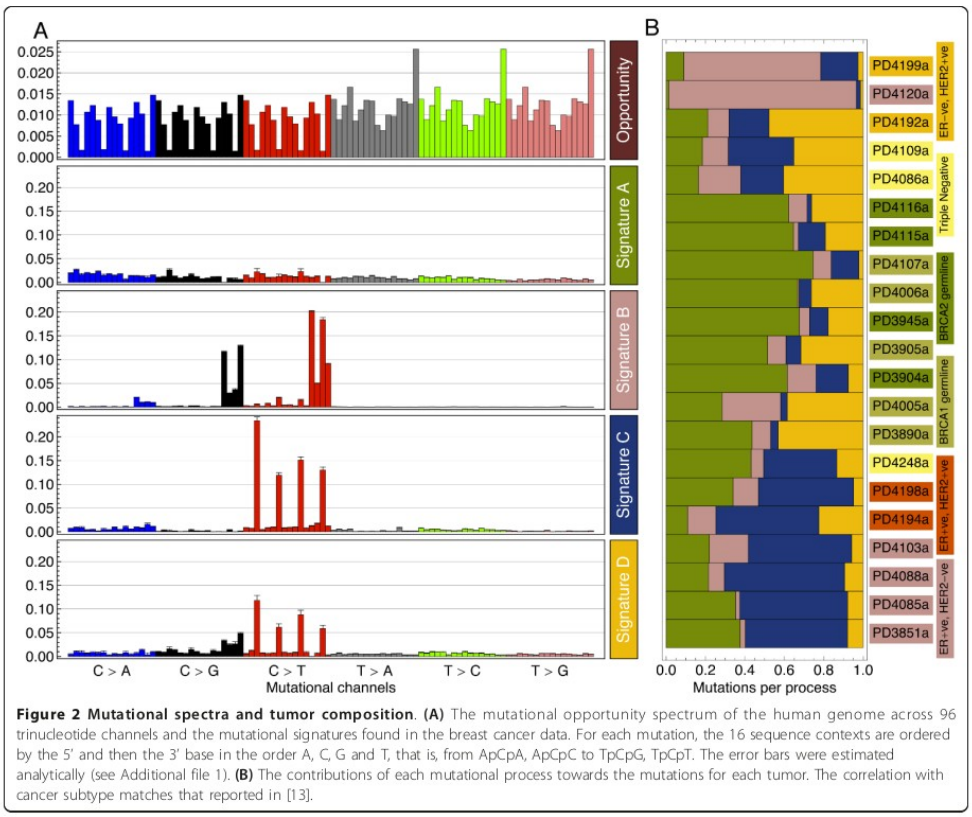
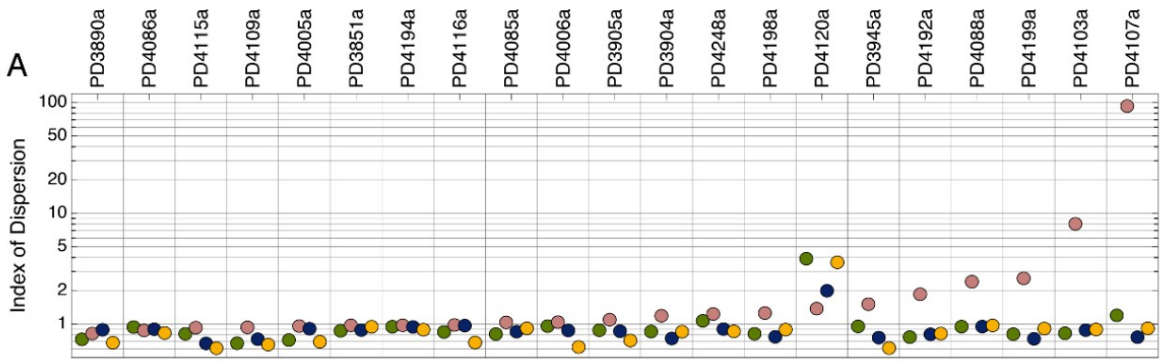


Figure 2 Mutational spectra and tumor composition. **(A)** The mutational opportunity spectrum of the human genome across 96 trinucleotide channels and the mutational signatures found in the breast cancer data. For each mutation, the 16 sequence contexts are ordered by the 5′ and then the 3′ base in the order A, C, G and T, that is, from ApCpA, ApCpC to TpCpG, TpCpT. The error bars were estimated analytically (see Additional file 1). **(B)** The contributions of each mutational process towards the mutations for each tumor. The correlation with cancer subtype matches that reported in [13].

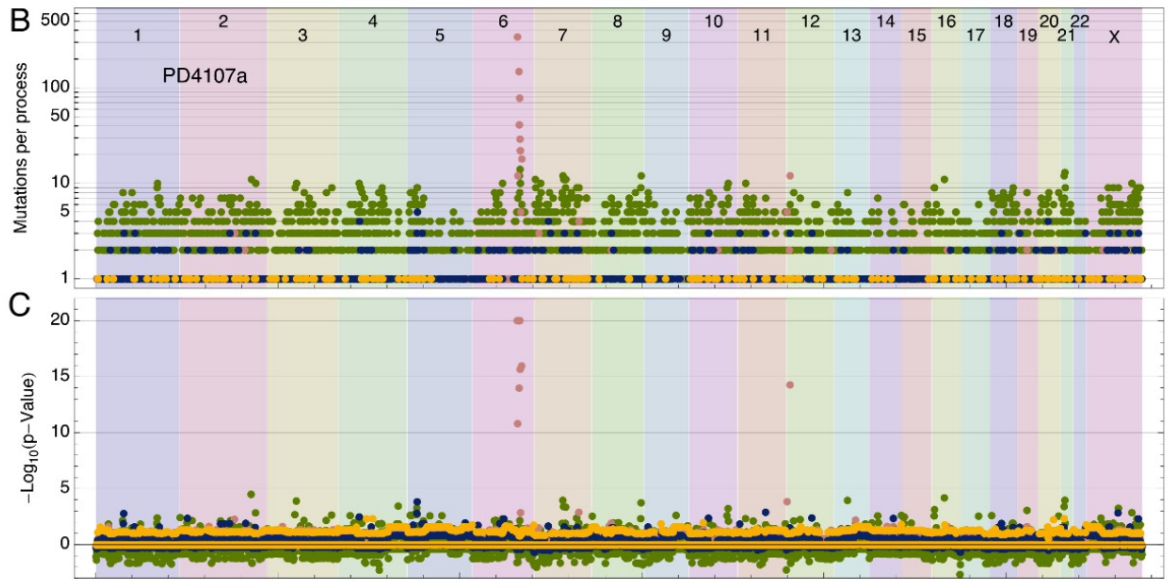
تجزیه و تحلیل موضعی، مناطقی را نشان می‌دهد که به شدت توسط فرآیندهای تکی مورد هدف قرار می‌گیرند:

برای موضعی کردن مناطق ژنوم که در آن فرآیندهای مختلف جهش‌زایی فعال هستند ، از امضاهای جهش‌زابه دست آمده برای تخمین فعالیت هر فرآیند برای هر پنجره توالی ۱ Mb استفاده کردیم. تنوع در فعالیت موضعی فرآیندهای جهش‌زا تکی در هر ژنوم سرطان به ویژه در فرآیند B مشاهده شد. برای به دست آوردن میزان ناهمگنی جهش‌زایی ، برای هر فرآیند در هر نمونه سرطانی یک شاخص پراکندگی (نسبت واریانس به میانگین جهش‌ها در هر پنجره) ، معادل فرصت جهش‌زایی ، جایی که مقدار 1 مطابق توزیع unbiased و همگن جهش‌ها در سراسر ژنوم است.



خوشه بندي تعداد زيادي جهش در ناحيه كوچكي از ژنوم كه kataegis ناميده مي‌شـود ، قبلاً در ژنوم‌هاي سرطان سـينه در نظـر گرفتـه شـده اسـت كـه در 13 مـورد از 21 تومور وجود دارد. در كار قبلی ، رویدادهای kataegis تا حدود زیادی جهش‌هـای C> T و C> G را نشان میدادند. اکنون می بینیم کـه kataegis بـه شـدت بـا امضـای جهش‌زا B مرتبط است.

یک دید بهتر از تغییرات جهش‌یافته مناطق خوشه بندی جهش را شناسایی کرد. بـرای هر منطقه 1 مگابایت ژنوم ، تعداد جهش‌های شناسایی شده از هـر فرآینـد در برابـر این فرض صفر ارزیابی شد که هر فرآیند جهش‌های خـود را بطـور تصـادفی در ژنـوم سرطان توزیع می‌کند ، با توجه به یک فعـالیت یکنـواخت در سرتاسـر ژنـوم ، تنهـا بـا فرصت موضعی تنظیم شده. بررسـی P-value بـه عنـوان تـابعی از مـوقعیت ژنـوم در نمونـه تومـور PD4107a بسـیار واضـح اسـت کـه رویـداد kateagis قبلاً در کرومـوزوم 6 گزارش شده است ، اما علاوه بر این ، یک واقعه کوچکـتر امـا بسـیار مهم در ابتـدای کروموزوم 12 پیدا کرده است.



فرآیندهای جهشی و حالت chromatin:

تجزیه و تحلیل موضعی را می‌توان برای هر زیر مجموعه تعریف شده ژنوم انجام داد. تقسیم بندی ژنوم به حالت‌های chromatin از لحاظ عملکردی مرتبط با پروژه ENCODE مورد توجه ویژه ای است[20،21]. در حقیقت ، تراکم منطقه‌ای جهش‌ها در حالت chromatin در human mammary epithelial cells (HMECs) متفاوت است. مقایسه فرصت جهش‌زایی در هر حالت chromatin با جهش‌های واقعی مشاهده شده نشان داد که 7٪ جهش بیشتر از آنچه انتظار می‌رفت در بخش های heterochromatin وجود دارد ، اما تقریباً 50٪ جهش کمتری در مناطق promoter داشت. با استفاده از روش استنباط موضعی ما حساسیت فرآیندهای جهش‌زا به حالت chromatin را بررسی کردیم. برای این منظور ، میزانی که یک فعالیت موضعی فرآیند جهش‌زا به حالت chromatin یک بخش ژنومی بستگی دارد، اندازه‌گیری می‌شود. فرایند جهش‌زا C (C> T در محل‌های XpCpG) یک تغییر بسیار مهم در فعالیت هر دو منطقه heterochromatin و promoter را نشان می دهد. این سیگنال به طور مداوم در تمام 21 سرطان مشاهده می‌شود. برای تعیین میزان این اثر ، تعداد جهش‌هایی را که به یک فرآیند خاص در هر حالت کروماتین اختصاص داده شده را با تعداد جهش‌هایی که انتظار می‌رود در صورت عملکرد یکنواخت فرایند در سراسر ژنوم دیده شود، مقایسه کردیم. نسبت این دو عدد را برای هر فرایند در هر حالت کروماتین و برای هر ۲۱ تومور نشان دادیم. در مناطق heterochromatin فعالیت فرایند C در همه تومورها افزایش می‌یابد، به طور متوسط تا ۴۰٪، نسبت به به اندازه آن بخش ژنومی اثر بزرگی دارد. در مناطق promoter ، فرایند C به طور مداوم کاهش می‌یابد ، به طور متوسط 80٪ـ این مشاهده این احتمال را ایجاد می‌کند که فرایند C به حالت DNA methylation وابسته باشد. فرآیند C عمدتا از جهش‌های C> T در محل‌های CpG تشکیل شده است. چنین جهش‌هایی مشاهده شده است که بیشتر در سایت های CpG هنگام متیله شدن cytosine اتفاق می‌افتد ، در نتیجه یک deamination خود به خودی [22،23]ـ در مناطق promoter ژن‌های فعال ، تعداد محل‌های CpG بیشتر از حد معمول وجود دارد (افزایش فرصت برای فرآیند C)ـ ، اما به طور عمده آنها متیله نشدند (کاهش فعالیت مشاهده شده فرآیند C).

پیشنهادها:

تفسیر حالت کروماتین که در اینجا مورد استفاده قـرار گـرفت ، از آزمایشـات انجـام شده بر روی بافت سالم بدست آمد. یک توسعه جالب برای این کار ، بررسی ارتبـاط بین فرآیندهای جهش‌زا و تفسـیر عملکـردی و اپی‌ژنـتیکی ژنوم‌هـای سـرطانی واقعی است که در آنها یافت شده است. این نوع داده‌ها هنوز در دسترس نیست اما مطمئناً در دست فناوری‌های فعلی است.

روش استنباط ارائه شده در این مطالعه محـدود بـه ۹۶ کانـال جهش سـه‌نوکلئوتیدی نمی باشد. علاوه بر این، می‌توان اطلاعاتی از قبیل رشته رونویسی بودن جهش‌ها یـا اینکه آیا جهش دیرتر یا زودتر در زمان سرطان رخ داده است[19].

به دلیل بازآرایی های کرومـوزومی (chromosomal rearrangements) و تغیـیرات در تعـداد نسخه‌ها (copy number changes) در مقیاس بزرگ در طی تکامل جسمی ، این فرصـت جهش‌زایی در واقع یک مقـدار پویـا اسـت [19]ـ بـه عنـوان یـک تقـریب ، مـا بـرای محاسبه فرصت جهش‌زایی از copy number state ژنوم‌های سرطان در زمان توالی‌یابی استفاده می‌کنیم. به عنوان مثال ، ممکن است در اواخر مرحله تکامـل سـرطان یـک واقعه کپی بر روی یک نسخه والـدین از یـک کرومـوزوم اتفـاق افتـاده باشد. بنـابراین فرصت جهش‌زایی برای این قطعه کپی 50٪ـ (از دو تا سه نسخه) بـرای زمـان بـاقی مانده تا توالی‌یـابی افـزایش می‌یابد. در این کـار ، مـا فـرض کـردیم کـه این فرصـت افـزایش جهش‌زایی در همـه فرآینـدهای جهش‌زا در سراسـر ژنـوم وجـود دارد. بـا اطلاعات دقیق‌تر در مورد زمان‌بندی نسبی تغییرات در تعداد نسـخه‌ها ، می‌تـوان یـک فرصت جهش‌زایی مؤثر را ایجاد کرد.

کاربردهای آینده این روش شامل فهرست‌نویسی امضاهای جهش‌زا در انـواع مختلـف سرطان ، شناسـایی فرآینـدهای جهش‌زا مشـترک و خـاص از نـوع سـرطان (و از این رو ، طبقه‌بندی بهتر سـرطان‌ها براسـاس تـرکیب فرآینـد آنهـا) و سـرانجام یک مـدل ابتدایی وابسته‌ـ به نوع ـ سرطان (cancer-type-dependent) برای شناسایی انواع گذرگر (driver) مسبب سرطان، از طریق سیگنال های انتخاب ، می‌شود.

Materials and methods

داده جهش:

183,916 جهش جسمی نقطه‌ای

$M = ۲۱$ نمونه سرطان سینه

$N_c = ۹۶$ کانال جهش سه‌نوکلئوتیدی

۶ جفت تغییر نوکلئوتیدی ممکن، C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, T:A > G:C

به همراه ۱۶ حالت مختلف از دو نوکلئوتید مجاور هر جهش

← مجموعه مشاهدات:

X_j^m (j = 1, . . . , N_c, m = 1, . . . , M) of channel- and tumor-specific mutation counts

فرآیندهای ابتدایی جهش‌زا:

n فرایند جهش‌زا ابتدایی مجزا

امضاهای جهش‌زا:

$\{\mu_{aj}\}$ (a = 1,...,n, j = 1,...,N_c), where $\sum_{j=1}^{N_c} \mu_{aj} = 1$

x^m_a فعالیت فرایند a در تومور m ، میزان عملکرد یک فرآیند در تومور مشخص (سهم فرایند در تومور)

توجه به این نکته ضروری است که استخراج میزان جهش واقعی یک فرآیند امکان پذیر نیست. مجموع فعالیت یک فرآیند در یک سرطان معین ، حالضرب میزان جهش آن ، زمانی که فرایند فعال بوده است و نسبت ژنومی که در واقع مورد حمله قرار گرفته است. در حالی که این مورد آخر را می توان با یک تجزیه و تحلیل مکانی حل کرد ، دو مورد اول بدون توالی‌یابی time-resolved تومور غیرقابل حل است.

با توجه به مطالب فوق ، با استفاده از توزیع پواسون نتیجه ایجاد تصادفی جهش‌ها در ژنوم را توصیف می‌کنیم. با فرض اینکه فرآیندهای جهش‌زا مستقل از یکدیگر باشند ، احتمال مشاهده بردار x^m از تعداد جهش در تومور m در کانال j :

$$P\left(X^m | x^m, \omega^m, \mu\right) \equiv \prod_{j=1}^{N_c} \text{Pois}\left(X_j^m | \sum_{a=1}^{n} x_a^m \mu_{aj} \omega_j^m\right), \quad (1)$$

ω_j^m فرصت جهش‌ها در کانال j که در تومور m رخ داده است

در داده‌های مورد بررسی ، تنوع تعداد کپی (copy number variation) در فرصت جهش‌زایی بین تومورها متفاوت است. ما در اینجا در اطلاعات ploidy موجود [19] به

شرح زیـر اسـتدلال کـرده ایم: هـر جفت نوکلئوتیـد بسـته بـه محتـوای تـوالی مجـاور نوکلئوتید جهش یافته می‌تواند به سه روش مختلف متناظر با سه کانـال 96 جهش‌زا، جهش یابد. بردار مجموع فرصت جهش‌زایی ω^m ، مجموع تجمعی در کل نوکلئویتـدها است ، که در آن هر نوکلئویتید سه کانال نه با یک بلکه با تعـداد کـپی آن نوکلئوتیـد در ژنوم m افزایش می یابد. در روشی که فرصـت جهش‌زایی را نادیـده گرفتـه اسـت ، مقادیر ω_j^m به 1 تنظیم شده است. توجه داشته باشیم که ، از پارامترهـای موجـود در معادله بالا ، فعـالیت هـای x و طیـف μ ناشـناخته هسـتند. تخمین حـداکثر احتمـال این مقادیر با استفاده از الگوریتم EM انجام شد.

پیدا کردن طیف‌های جهش‌زایی (امضاهای جهش‌زا):

الگوریتم EM تخمین حـداکثر درسـت‌نمایی (ML) پارامترهـای نـامعلوم مـدل (در اینجـا طیف μ) و داده‌های پنهان (در اینجـا فعـالیت x) را می‌یابـد. پس از یـک حـدس اولیـه ، $\mu(0)$ ساخته شد ، دو مرحلـه بـه صـورت iterative تکـرار می‌شـوند. در مرحلـه اول ، بـا توجه به بهـترین حـدس فعلی ، $\mu(k)$ و داده‌هـای مشـاهده شـده ، X ، تخمیـنی بـرای x یافت می‌شود. در مرحله دوم ، این برآورد ، $\hat{x}$ از داده‌های پنهان برای به دسـت آوردن برآورد پارامتر بـه روز شـده ، $\mu\ (k+1)$ اسـتفاده می‌شـود. این دو مرحلـه تـا زمـانی رسیدن همگرایی به حداکثر (محلی) درست‌نمایی داده، $P(X \mid \mu)$ ، تکرار می شود [17]. به صورت دقیق‌تر روال به شرح زیر است:

0. (Initialize) باتوجه به نرمال سازی $\sum_{j=1}^{N_c} \mu_{aj} = 1$ مقداردهی اولیه μ
1. (E-step) باتوجه به تخمین حاضر، μ^k ، برآورد حداکثر درست‌نمایی را برای همـه فعالیت‌های پنهان پیدا کنید:

$$\hat{x}^m \equiv \underset{x\in\mathbb{R}_+^n}{\operatorname{argmax}} \log P\left(X^m|x,\omega^m,\mu^{(k)}\right) = \underset{x\in\mathbb{R}_+^n}{\operatorname{argmin}} \sum_{j=1}^{N_c}\left[\sum_{a=1}^{n} x_a\mu_{aj}^{(k)}\omega_j^m - X_j^m \log\left(\sum_{a=1}^{n} x_a\mu_{aj}^{(k)}\omega_j^m\right)\right]$$

2. (M-step) با استفاده از تخمین‌های داده‌های پنهان، طیف جهش را به روزرسانی

$$\mu^{(k+1)} \equiv \underset{\mu\in\mathbb{R}^{n\times N_c}}{\operatorname{argmax}} \sum_{m=1}^{M} \log P\left(X^m|\hat{x}^m,\omega^m,\mu\right), \text{with constraint } \sum_{j=1}^{N_c}\mu_{aj}^{(k+1)} = 1.$$

کنید:

3. (Finish) تست برای همگرایی. در صورت نیاز به مرحله E-step و تکرار برگردید.

محاسبه مجموع داده‌های درست‌نمایی:

پس از همگرایی الگوریتم EM به حداکثر تخمین درست‌نمایی ، $\hat{\mu}$ ، از n طیف ، مقدار عددی درست‌نمایی با ادغام متغیرهای نهفته ، x محاسبه شد. برای جلوگیری از هزینه ادغام عددی ، این کار با استفاده از تقریب نقطه زینی انجام شد:

$$\log P\left(X|\hat{\mu}\right) = \sum_{m=1}^{M} \log \int d^m x\, P\left(X^m|x, \omega^m, \hat{\mu}\right) \approx \sum_{m=1}^{M} \left[\frac{n}{2}\log 2\pi - L\left(\hat{x}^m\right) - \frac{1}{2}\log\det H\left(L\right)\left(\hat{x}^m\right)\right]$$

که در آن $L(x) = -\log P(X|x, \omega, \mu)$ لگاریتم درست‌نمایی داده شرطی است و $H(L)(x)$ ماتریس Hessian مشتق دوم آن است.

پیدا کردن تعداد فرایندهای جهش‌زا:

مقایسه درست‌نمایی داده‌هاي به دست آمده با مقادیر مختلف n ، محتمل‌ترین تعداد فرایندهای جهش‌زا را نشان می‌دهد . با افزایش مقدار n تعداد پارامترهای مدل موجود برای fitting افزایش می‌یابد ، و به طور کلی توضیحی بهتر برای داده‌ها، X ایجاد می‌کند. این به معنای افزایش ارزش درست‌نمایی ، $P(X \mid \hat{\mu})$ است. برای جلوگیری از overfitting داده‌ها ، از Bayesian Information Criterion (BIC) برای اصلاح پیچیدگی مدل استفاده شد. در نهایت مدل با بیشترین مقدار BIC انتخاب شد :

$$\mathrm{BIC} = 2\log P\left(X|\hat{\mu}\right) - n\left(N_c - 1\right)\log M.$$

برآورد فعالیت‌های فرآیند و اختصاص جهش‌ها به فرآیندها:

تجزیه و تحلیل سراسری:

تکمیل الگوریتم EM تخمین‌های حداکثر درست‌نمایی ، $\hat{\mu}$ از امضاهای جهش‌زا n فرآیند را فراهم می‌کند. به همین ترتیب ، E-step تخمین‌های ، $\{\hat{x}^m_a\}$ ، از فعالیت‌های فرآیند برای هر نمونه را بدست می‌آورد. در مجموع ، می‌توان از این برای تخمین تعداد جهش‌هایی استفاده $X^m = \sum_{a=1}^{n} \hat{X}^m_a$ تمال زیاد توسط هر فرآیند در مجموع همه

جهش‌های مشاهده شده شرکت می‌کند: . برای بهبود برآورد فعالیت ، یک تخمین ساده از حداکثر احتمال لگاریتم درست‌نمایی از

$$\hat{x}^m \equiv \underset{x\in\mathbb{R}_+^n}{\operatorname{argmax}} \log P\left(X^m|x,\omega^m,\mu^{(k)}\right) = \underset{x\in\mathbb{R}_+^n}{\operatorname{argmin}} \sum_{j=1}^{N_c}\left[\sum_{a=1}^{n} x_a\mu_{aj}^{(k)}\omega_j^m - X_j^m \log\left(\sum_{a=1}^{n} x_a\mu_{aj}^{(k)}\omega_j^m\right)\right]$$

را تعمیم می‌دهیم تا اطلاعات پیشین را به صورت pseudocounts درج کنیم. همانطور که در زیر نشان می‌دهیم ، این pseudocounts به طور کلی در کانال‌ها یکنواخت نیستند. فعالیت تخمین زده می‌شود ، $\{\hat{x}^m_a\}$ ، و دارای وضعیت بیشینه احتمال پسین است. سرانجام ، تعداد جهش ها ، $\hat{x}^m_a$ ، در سرطان m که از طریق فرایند a در کانال j منتشر شده‌اند به شرح زیر است:

$$\hat{x}_{a,j}^{m,g} = \frac{|X^m|\,\hat{x}_a^{m,g}\hat{\mu}_{a,j}\omega_j^m}{\sum_{b=1}^{n}\hat{x}_b^{m,g}\left(\hat{\mu}\omega^m\right)_b}, \text{where } \left|X^m\right| \equiv \sum_{j=1}^{N_c} X_j^m, \text{ and} \left(\hat{\mu}\omega^m\right)_a \equiv \sum_{j=1}^{N_c}\mu_{aj}\omega_j^m.$$

در اینجا ، اندیس بالا g مقدار سراسری بدست آمده را نشان می‌دهد. قاعده تخصیص را می‌توان به شرح زیر درک کرد: کسری از جهش‌های مشاهده شده که توسط یک فرآیند خاص تولید شده‌اند باید متناسب با فعالیت آن ، طیف جهش آن و فرصت جهش در آن کانال باشد. به هر مشاهده ، X^m ، در مجموع n تعداد pseudocounts اضافه شد. این شمارش دقیقاً به روش فوق انجام شد ، اما بدون هیچ گونه bias ناشی از فعالیتهای (هنوز ناشناخته):

$$\tilde{X}_{aj}^m \equiv \frac{n\hat{\mu}_{aj}\omega_j^m}{\sum_{b=1}^{n}\left(\hat{\mu}\omega^m\right)_b}$$

بنابراین ، آنها نمایانگر مشاهده‌ای هستند که فقط براساس اطلاعات شناخته شده پیشین از برآورد فعالیت انجام می‌شود. pseudocounts با یک عبارت اضافی وارد تابع لگاریتم درست‌نمایی می‌شود:

$$L\left(x; X^m, \omega^m, \mu\right) \rightarrow L\left(x; X^m, \omega^m, \mu\right) - \sum_{j=1}^{N_c} \sum_{a=1}^{n} \tilde{X}_{aj}^m \log\left(x_a \mu_{aj} \omega_j^m\right)$$

تجزیه و تحلیل موضعی:

با همان منطق مشابه در بالا ، ما تخمین‌هایی از فعالیت موضعی هر فرآیند در هر سرطان گرفتیم و هر ژنوم را به پنجره های غیر همپوشانی با طول 1Mb تقسیم می‌کنیم. با این حال ، ما اکنون می‌توانیم از فعالیت‌های سراسری قبلاً محاسبه شده ، $\{\hat{X}^{m,g}_a\}$، به عنوان اطلاعات پیشین آگاه‌تر برای استنباط فعالیت‌های فرآیند در هر پنجره megabase ، به طول l استفاده کنیم:

$$\tilde{X}_{aj}^{m,l} = \frac{n \hat{x}_a^{m,g} \hat{\mu}_{aj} \omega_j^{m,l}}{\sum_{b=1}^{n} \hat{x}_a^{m,g} \left(\hat{\mu} \omega^{m,l}\right)_b} \xrightarrow[\text{using } X^{m,l}, \omega^{m,l}]{\text{E - step with eq. 8}} \hat{x}_a^{m,l}.$$