:HIA

یک نقشه بردار ژنومی با استفاده از همترازی توالی مبتنی بر شاخص هیبرید

خلاصه

زمینه: تعدادی از ابزارهای همترازسازی برای همتراز کردن ریدهای توالی یابی با ژنوم مرجع انسان ایجاد شده‌اند. مقیاس اطلاعات بدست امده از آزمایشات توالی یابی نسل جدید (NGS)، به سرعت در حال افزایش است. مطالعات اخیر انجام شده بر اساس فناوری NGS به طور مرتب اگزوم‌ها یا توالی‌های کامل ژنومی چندصد تا چندهزار نمونه را ایجاد کرده است. برای تامین نیاز روزافزون به آنالیز دیتاست‌های بسیار بزرگ NGS، لازم است که ابزارهای نقشه-برداری سریع‌تر، حساس‌تر و دقیق‌تری ایجاد شود.

نتایج: HIA از دو شاخص جدول درهم‌سازی و شاخص suffix array استفاده می‌کند. جدول درهم سازی جستجوی مستقیم q-gram را انجام می‌دهد و شاخص suffix array جستجوی بسیار سریع رشته‌های با طول متغییر را با استفاده از جستجوی دودویی اجرا می‌کند. ما پی بردیم که ترکیب جدول درهم‌سازی و suffix array بسیار سریعتر از روش suffix array می‌تواند یک زیر رشته را در توالی مرجع پیدا کند. در اینجا ما منطقه‌ی تطابق (MR) را تعریف کردیم که طولانی‌ترین زیر رشته‌ی مشترک بین یک مرجع و یک رید است. همچنین ما مناطق همترازی کاندید (CARs) را نیز به صورت لیستی از MRs تعریف کردیم که در کنار یکدیگر قرار دارند. شاخص هیبرید برای یافتن مناطق همترازی کاندید (CARs) بین مرجع و رید استفاده شد. ما پی بردیم که همترازی نواحی بدون تطابق در CAR بسیار سریعتر از همترازی کل CAR است. در بررسی‌های آزمایشی، HI در مقایسه با سایر ابزارهای همترازی نقشه‌برداری سریعتر و بدون کاهش چشمگیر در دقت نقشه را از خود نشان داد.

نتیجه‌گیری: آزمایشات ما نشان می‌دهد که هیبرید جدول درهم‌سازی و suffix array از نظر سرعت نقشه برداری ریدهای توالی‌یابی NGS به توالی مرجع ژنوم انسانی مفید است. در نتیجه، ابزار ما برای همتراز کردن دیتاست‌های عظیم به دست امده از توالی‌یابی NGS مناسب است.

کلمات کلیدی: شاخص هیبرید، NGS، مپر، همترازی توالی، شاخص درهم‌سازی، شاخص Suffix array

زمینه

مطالعات اخیر براساس فناوری توالی یابی نسل جدید (NGS) صدها یا هزاران اگزوم یا توالی کامل ژنومی را با کاهش هزینه‌های آزمایشات NGS ایجاد کرده است. با تکامل فناوری‌های NGS، این فناوری‌ها به تدریج طول ریدها را افزایش داده و از میزان خطاها کاسته‌اند. برای همگام شدن با فناوری‌های درحال توسعه‌ی NGS، ابزارهای همترازسازی زیادی برای ریدهای کوتاه و بلند ایجاد شده است. این ابزارها شامل SSAHA2، BWA ، AGILE، SOAP2، Bowtie2، SeqAlto و غیره هستند. از بین اینها، بسیاری از برنامه‌های همترازی از راهبرد نقشه‌برداری مبتنی بر شاخص استفاده می‌کنند. برای مثال، SSAHA2، AGILE و SeqAlto از یک شاخص جدول درهم‌سازی (HT) یک ژنوم مرجع استفاده می‌کنند در حالیکه BWA، SOAP2، و Bowtie2 از یک شاخص ژنومی مبتنی بر تبدیل باروز-ویلر استفاده می‌کند.

تمام ابزارهای همترازی مبتنی بر HT از استراتژی seed and extended، استفاده می‌کنند که با جستجوی مناطق همترازی کاندید (CARs) (همترازی هر جایگاه) و گزارش بهترین همترازی‌ها عمل می‌کند. شاخص HT از جستجوی سریع جایگاه‌های کاندید دارای q-gram ها پشتیبانی می‌کند (رشته‌های طول q). Q کوچکتر حساسیت و تعداد CARs را افزایش می‌دهد اما q بزرگتر حساسیت و تعداد CARs را کاهش می‌دهد. علاوه بر این از آنجایی که q ثابت می‌شود، زمانیکه q-gram ها برای یک طول جدید نیاز باشند، HT باید مجددا ساخته شود. بیشتر ابزارهای همترازی مبتنی بر BWT از شاخص full-text minute استفاده می‌کنند که از نظر حافظه کارامد است و شبیه

suffix tree است. از نظر زمان تطابق، suffix tree برای تطابق دقیق کارامد است اما برای تطابق غیر دقیق کند است. BWA و Bowtie2 از رویکردهای seed-and-extend مشابهی مانند استفاده از الگوریتم‌های مبتنی بر HT برای ریدهای طولانی پیروی می‌کنند.

پشتیبانی از تطابق ریدهای بلند، سرعت بالا، دقت و حساسیت، ویژگی های ضروری ابزارهای نقشه برداری NGS فعلی است. در اینجا، ما سعی کردیم مزایای استفاده از همترازی مبتنی بر HT و suffix tree در یک ابزار را بیان کنیم که این الزامات را برآورده می کند. برای این منظور، ما یک نقشه بردار ژنومی را با استفاده از همترازی توالی مبتنی بر شاخص هیبرید ایجاد کردیم.

در این مقاله، ما ابزار HIA را تشریح می‌کنیم، و نتایج مقایسه‌های انجام شده برروی کارایی این HIA و چهار ابزار همترازی محبوب دیگر شامل BWA، Bowtie2 ، SOAP2 و SeqAlto برروی داده‌های واقعی و شبیه‌سازی شده را نشان می‌دهیم. نتایج تجزیه و تحلیل ازمایشی نشان می دهد که HIA در مقایسه با دیگر ابزارهای همترازی بویژه از نظر سرعت، کارامدتر است.

روش‌ها

شاخص هیبرید

Σ را یک الفبا و $S = s_0s_1 \dots s_{m-1}$ یک رشته بر روی Σ در نظر بگیرید. $|S| = m$ را طول S، $S[i] = si$ را نماد iم s، $S[i, j] = si \dots sj$ را یک زیر رشته و $Si = S[i,m-1]$ یک افزونه s در نظر بگیرید. ما q-gram را به عنوان یک زیر رشته s با طول q تعریف می‌کنیم. در متن توالی DNA، الفبای Σ شامل چهار نوکلئوتید A، T، C و G است (به عبارتی $\Sigma = \{A, C, G, T\}$). ما به جای A، C، G و T به ترتیب از اعداد 0، 1، 2 و 3 استفاده می‌کنیم. بنابراین، هرq-gram به عنوان یک عدد صحیح بدون علامت با دو بیت در هر نوکلئوتید کدگذاری می شود. با این وجود، بسیاری از توالی‌های ژنومی مرجع حاوی نوکلئوتیدی غیر از این چهار نوکلئوتید مانند N هستند. این

مسئله در توالی‌های NGS نیز اتفاق می‌افتد. ما N را با یک نوکلئوتید تصادفی یکنواخت مانند BWA و بسیاری از ابزارهای دیگر جایگزین می‌کنیم.

از نظر شاخص هیبرید، SOAP2 یک جدول درهم‌سازی را بر روی شاخص FM که یک SA فشرده است، اجرا می کند. از طرف دیگر، شاخص هیبرید دیگر حاوی یک توالی مرجع، suffix array (SA) و یک جدول درهم‌سازی (HT) است که در شکل 1 نشان داده شده است. به دلیل اینکه چهار نماد در این الفبا وجود دارد، توالی مرجع طول N می‌تواند به صورت N/4 بیت بسته بندی شود. SA یک ارایه‌ی موقعیت‌های آغازی (اعداد صحیح) افزونه‌های توالی مرجع در ترتیب لکسیکوگرافیکال است. تعداد افزونه‌های توالی اندازه‌ی N برابر N است. HT یک آرایه از اشاره گرها به SA است که نشان می دهد که کدام موقعیت در SA متعلق به q-gram است. از آنجا که ما یک q-gram را به عنوان یک رشته از طول q تعریف می کنیم، تعداد عناصر HT برابر $4^q + 1$ است. با توجه به q-gram ، HT[x] اولین موقعیت q-gram در SA است، که در آن x مقدار عددی q-gram است. ما محدوده‌ی (R) q-gram را در SA به صورت زیر تعریف می‌کنیم:

$$R(x) = [HT[x], HT[x + 1] - 1] \quad (1)$$

اگر q-gram در توالی وجود نداشته باشد، HT[x] اولین موقعیت q-gram بعدی در تولی است، بنابراین، HT[x] و HT[x + 1] برابر هستند و R(x) خالی است.

روش ساخت شاخص هیبرید شامل چهار فرایند است که عبارتند از : 1) بسته بندی توالی مرجع به صورت 2 بیت به ازای هر باز (توالی) ؛ 2) شمارش هر q-gram در توالی و ایجاد محدوده‌ی q-grams در SA (HT)؛ 3) وارد کردن موقعیت توالی هر q-gram در محدوده‌ی SA و 4) مرتب سازی لکسیکوگراقیکی هر q-gram (SA). بر اساس تعیین ذهنی موقعیت نهایی بسیاری از افزونه‌ها با استفاده از تنها چند نماد اول هر افزونه، الگوریتم ساخت SAمرتب سازی پیشوندهای طول w در افزونه‌های محدوده‌ی q-gram را انجام می‌دهد. اگر برخی از افزونه‌ها پیشوند طول W را به اشتراک بگذارند، دسته با دسته بندی طول W زیر رشته‌ی پیشوندهای طول W تکرار می‌شود. برای کاهش زمان دسترسی به توالی، پیشوندهای طول W به مقادیر صحیح تبدیل می‌شوند. در صورتی که اندازه کلمه حافظه 4

بایت و اندازه حروف الفبا 4 باشد، wبین 0 تا 16 تنظیم می شود. محدوده q-gram با هم تداخل پیدا نمی‌کنند به طوری که فرآیند چهارم نیز می تواند به صورت موازی باشد. شکل 1 شکل 1، ساختار داده‌های اساسی و روش ساخت یک شاخص هیبرید برای یک توالی مرجع را توصیف می کند.

بازیابی موقعیت توالی مورد نظر q در دو مرحله اجرا می شود: جستجوی HT و جستجوی دودویی SA. اگر پیشوند Qطول q (q-gram) زیر رشته ای از توالی است، ما محدوده ای را پیدا می‌کنیم که در آن q-gram به SA تعلق دارد (با استفاده از معادله‌ی 1) . به عبارت دیگر محدودی R یک محدوده‌ی خالی را بر می‌گرداند که نشان می‌دهد که Q در توالی قرار ندارد. اگر محدوده‌ی بازگردانی شده خالی نباشد، ما موقعیت‌هایی را پیدا می‌کنیم که Q در آنها با جستجوی دودویی زیر رشته‌ی Q [1 − |Q|] در توالی اتفاق می‌افتد. به صورت نظری، جستجوی زیر رشته‌ی طول m در یک رشته از طول N با استفاده از SA می‌تواند در زمان O (mlogN) در بد بینانه‌ترین حالت به کار گرفته شود. شاخص جدول درهم‌سازی می‌تواند طول رشته‌ی مورد جستجو مانند $(m' = m - q)$ و اندازه‌ی محدوده‌ی جستجو مانند $(n' \ll N)$ را کاهش دهد. زمانیکه توالی مرجع ساختمان GRCH37 ژنوم انسان است و q برابر 14 است، طول توالی بسته بندی شده 2، 861، 343، 766 و میانگین اندازه‌ی محدوده‌ی جستجو 14.12 است. ازمایش ما نشان داد که شاخص جدول درهم‌سازی زمان جستجو را به میزان چشمگیری کاهش می‌دهد (بنگرید به فایل پیوست 1).

نقشه‌برداری هیبرید: یافتن نواحی همترازی کاندید (CAR)

نقشه‌برداری هیبرید از روش seed-and-extend مشابهی با سایر ابزارهای مبتنی بر HT پیروی می‌کند. یک MR(ناحیه‌ی تطابق) زیر رشته‌ی مشترک بین توالی مرجع و رید است. ’sp‘ را به عنوان ناحیه‌ی شروع توالی مرجع در نظر بگیرید که MR در آن کار دارد.

ما هر MR را به صورت <dv, ro, L> مشخص می‌کنیم که در آن ’ro‘ (افست رید) ناحیه‌ی شروع در رید است که MR در آنجا قرار می‌گیرد، ’dv‘ (مقادیر دیاگونال) به صورت dv = sp − ro تعریف می‌شود و ’L‘ طول

MR است. با توجه به یک MR طول $m' - q + 1$ ، m' تعداد q-gram وجود دارد که مقادیر دیاگونال یکسان و افست‌های متوالی دارند. مقادیر دیاگونال برابر هستند که نشان می‌دهد که MRهای متناظر در توالی مرجع نزدیک یکدیگر قرار دارند. CAR (ناحیه‌ی همترازی کاندید) لیستی از MRهاست که نزدیک یکدیگر قرار دارند و بوسیله‌ی 'ro' مرتب شده‌اند. ما یک CAR را به صورت یک seed تعریف می‌کنیم و تنها مناطق ناجور در CAR را همتراز می‌کنیم.

روش یافتن MRها و CARها به صورت زیر است: 1) بازیابی محدوده‌ی SA هر q-gram با استفاده از HT و SA؛ 2) محاسبه‌ی مقادیر دیاگونال 3) دسته بندی با مقدار دیاگونال و افست.؛ 4) گروه بندی MRها با مقادیر دیاگونال یکسان و افست‌های پشت سرهم 5) ترکیب کردن MRهای مجاور به درون CARها 6) دست‌بندی CARها با بازهای جور از بالا به پایین. برای مثال باتوجه به رید $(r = GCCATG)$ و طول q-gram $(q = 2)$ و شاخص هیبرید ساخته شده در شکل 1، ما می‌توانیم MRها و CARها را به صورت زیر پیدا کنیم:

I. بازیابی محدوده¬ی SA هر q-gram با استفاده از HT و SA

موقعیت صفرم q-gram : (GC: SA[10, 11])

موقعیت اول q-gram: (CC: SA[8])

موقعیت دوم q-gram: (CA: SA[6, 7])

موقعیت سوم q-gram:(AT: SA[3, 4])

موقعیت چهارم q-gram: (TG: SA[15])

II. محاسبه¬ی مقادیر دیاگونال

(GC, 5, 0)، (GC, 11, 0) ، (CC, 11, 1) ، (CA, 4, 2) ، (CA, 11,2) ، (AT, −2, 3)، (AT, 4, 3)، (TG, 4, 4)

III. دسته¬ بندی با مقدار دیاگونال و افست.؛

(AT, −2, 3)، (CA, 4, 2) ، (AT, 4, 3)، (TG, 4, 4)، (GC, 5, 0) ، (GC, 11, 0) ، (CC, 11, 1) ،

(CA, 11, 2)

IV. گروه بندی MRها با مقادیر دیاگونال یکسان و افست¬های پشت سرهم

MR0: (AT, −2, 3, 2) ← (AT, −2, 3)

MR1: (CATG, 4, 2, 4) ← (CA, 4, 2), (AT, 4, 3), (TG, 4, 4)

MR2: (GC, 5, 0, 2) ← (GC, 5, 0)

MR3: (GCCA, 11, 0, 4) ← (GC, 11, 0), (CC, 11, 1), (CA,

11, 2)

V. ترکیب کردن MRهای مجاور به درون CARها و تنظیم بازهای جور

CAR0: (MR0; 2)

CAR1: (MR1, MR2; 5)

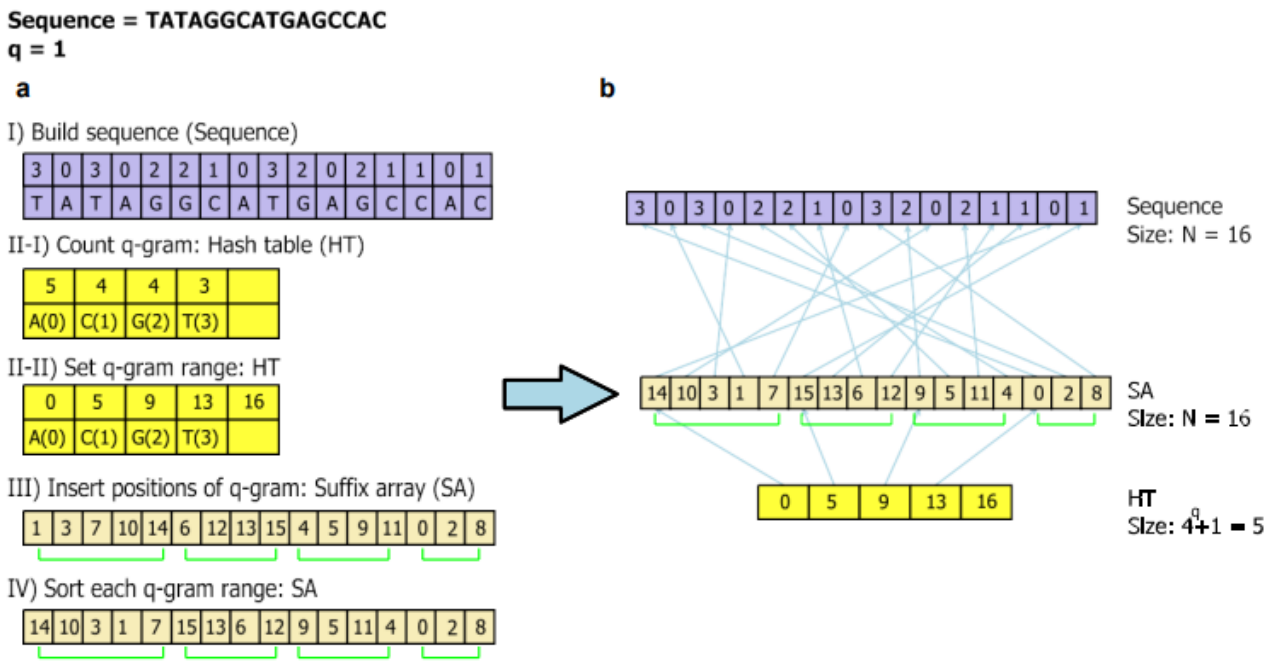
CAR2: (MR3; 4)

VI. دسته¬بندی CARها با بازهای جور

CAR1: (MR1, MR2; 5)

CAR0: (MR3; 2)

CAR2: (MR2; 2)



شکل 1. ساختن شاخص هیبرید. پنل a فرایند ساخت شاخص هیبرید را با دادن Sequence = TATAGGCATGAGCCAC و q = 1 نشان می‌دهد. روش ساخت به ترتیب زیر است: اول تبدیل نمادهای نوکلئوتید به معادل عددی آنها (I) . دوم شمارش هر q-gram و ذخیره سازی شمارش‌ها در HT (II-I) . سوم، تنظیم مناطق اغازی هر q-gram بر اساس تعداد q-gramها (II-II). چهارم، ذخیره سازی موقعیت‌های هر q-gram در SA مانند (II) . نهایتا، چینش محدوده‌ی هر q-gram در SA و ساختن شاخص هیبرید. اندازه‌های توالی، SA و HT به ترتیب 16، 16 و $4^q + 1 = 5$. پنل b شاخص هیبرید ساخته شده را نشان می‌دهد.

اگرچه مقادیر دیاگونال دو MR مجاور متفاوت است، اما این دو می‌توانند در یک CAR جای بگیرند به شرطی که بازهای اضافه شده یا حذف شده‌ای بین آنها وجود داشته باشد. در مورد CAR1، اختلاف مقدار بین مقدار دیاگونال MR1 و مقدار دیاگونال MR2 برابر 1 است و یک باز وارد شده (C) بین MR1 و MR2 وجود دارد. ما این مقدار را به مجاورت نسبت می‌دهیم و از این مقدار به منظور تنظیم مقدار کجاز اندازه‌ی حذف و اضافه بین MRها استفاده می‌کنیم.

به منظور یافتن کارامدتر MRها و CARها، ما از سه فراین کاوشی استفاده می‌کنیم. طول رید و میزان خطا را به ترتیب m و ε در نظر بگیرید. اولین روش کاوشی این است که یک زیر رشته با طول حداقل $m/(k + 1)$ بین دو

رید طول m با اختلاف k وجود دارد. فرض کنید $\lambda = \varepsilon m$ تعداد خطاهای مورد انتظار در یک رید باشد و x را یک متغییر تصادفی در نظر بگیرید. ما می‌توانیم شانس مشاهده‌ی یک رید با خطاهای k را به صورت زیر محاسبه کنیم :

$$P\{X \leq k\} = \sum_{i=0}^{k} e^{-\lambda} \lambda^{i}/i! \qquad (2)$$

فرمول 2 معادله توزیع تجمعی x است. بر اساس معادله‌ی 2 ما قادر هستیم مقدار ریدها با خطای k را محاسبه کنیم. اگر ما $k = 1.5\lambda$ تنظیم کنیم، مقدار ریدها با خطاهای 1.5λ به 0.9 میل می‌کند و ما می‌توانیم از q-gram با طول $m/(1.5\lambda + 1)$ استفاده کنیم. استفاده از یک q-gram با طول یکسان به عنوان زیر رشته‌ی مشترک تعداد MRها و CARها را کاهش می‌دهد.

ثانیا، از آنجایی که q-gram که در مناطق زیادی از توالی وجود دارد تفکیک کننده‌ی خوبی نیست، چنین q-gram نسبت به q-gram که در مناطق کمتر وجود دارد وزن کمتری می‌گیرد. این روش مبتنی بر عکس فراوانی سند (IDF) است که معمولا در زمینه بازیابی اطلاعات استفاده می شود. IDF یک شیوه‌ی وزن دهی است که مشخص می‌کند که آیا یک کلمه در کل سند مشترک است یا نه. با استفاده از این روش، می‌توانیم q-gram های با وزن پایین را فیلتر کنیم و در نتیجه از MRها و و CARهای نامطلوب خلاص شویم.

نهایتا، با توجه به رید (r) و رشته‌های S1 و S2 ، هر دو طول m ، اگر تعداد بازهای جور بین r و S1 بیشتر از تعداد بازهای جور بین r و S2 باشد مقدار اختلاف بین r و S1 کوچکتر از تعداد اختلاف بین r و S2 خواهد بود. این روش را می‌توان برای دسته‌بندی CARها با استفاده از تعداد بازهای جور و فیلتر کردن CARهای با رتبه‌ی پایین استفاده کرد.

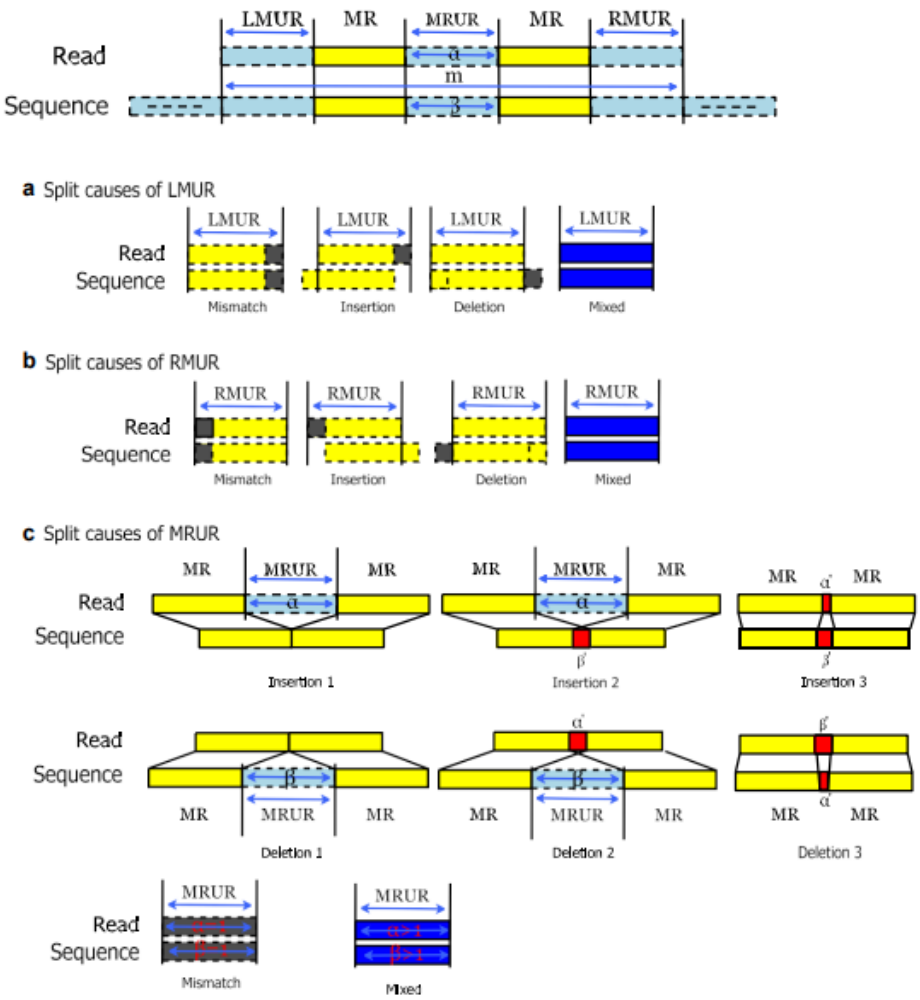
نقشه‌برداری هیبرید : همترازی ناحیه‌ی همترازی کاندید (CAR).

برای همترازی CARهای با رتبه‌ی بالا، ما نواحی ناجور در هر CAR را همتراز می‌کنیم، زیرا MRها کاملا همتراز شده‌اند (جور). ما نواحی ناجور را به سه دسته تقسیم می‌کنیم: سمت چپ‌ترین ناحیه‌ی ناجور (LMUR) که ناحیه‌ی ناجور سمت چپ اولین MR در CAR است، سمت راست‌ترین ناحیه‌ی ناجور (RMUR) که ناحیه‌ی ناجور راست

اخرین MR در CAR است و نواحی ناجور بین دو MR (MRURs) . نواحی جور و ناجور به دلیل وجود جفت شدگی‌های اشتباهی، دخول و حذف می‌توانند از هم جدا شوند. ما این علل جدایی را در شکل 2 بررسی می‌کنیم. در مورد LMUR وRMUR ، اگر تنها بازهای مجاور بین این ناحیه‌ی ناجور و MR جفت شدگی اشتباهی باشند و بازهای دیگر جور باشند، علت جدایی آن عبارت است از جفت شدگی اشتباهی؛ اگر این بازها نتیجه‌ی دخول باشند پس عامل جدایی دخول است، و اگر این بازها حذف شده باشند و بقیه جور باشند عامل جدایی حذف است در غیر اینصورت عامل جدایی مرکب است. اطلاعات جفت شدگی اشتباه، دخول و حذف به وضوح نشان می‌دهد که این نواحی چطور همتراز می‌شوند. بنابراین، ما می‌توانیم از الگوریتم نیدلمن- وونش تنها برای نواحی ناجوری که نتیجه‌ی چند عامل هستند استفاده کنیم بدین ترتیب می‌توانیم زمان محاسبات روش برنامه‌نوبسی پویای الگوریتم نیدلمن- وونش را کاهش دهیم.

عوامل زیادی برای جداییMRUR وجود دارد که در شکل 2 اشاره شده است. این عوامل را می‌توان به صورت جفت شدگی ناجور، دخول، حذف و ترکیبی از اینها دسته‌بندی کرد. در رابطه با دخول، اگر چندین باز (α) بین دو MR مجاور در یک رید وجود داشته باشد، ولی بازی در توالی وجود نداشته باشد، پس می توان باز αرا وارد کرد (دخول 1). اگر بازهای α بین دو MR مجاور در یک رید وجود داشته باشند و بازهای β در تولی همپوشانی شوند، می‌تواند بازهای α + β را وارد کرد (دخول 2). اگر بازهای α در یک رید و بازهای β در توالی همپوشانی داشته باشند و α کوچکتر از β باشد می‌توان بازهای α − β را وارد کرد (دخول 3). در رابطه با حذف، اگر چندین باز β بین دو MR مجاور در یک توالی وجود داشته باشد، ولی بازی در رید وجود نداشته باشد، پس می توان باز β راحذف کرد (حذف 1). اگر بازهای β بین دو MR مجاور در یک توالی وجود داشته باشند و بازهایα در تولی همپوشانی شوند، می‌تواند بازهای α + β را حذف کرد (حذف 2). اگر بازهایβ در یک توالی و بازهای α در یک رید همپوشانی داشته باشند و α بزرگتر از β باشد می‌توان بازهای α − β را حذف کرد (حذف 3). در رابطه با جفت شدگی اشتباهی، اگر یک باز بین دو MR مجاور در یک رید وجود داشته باشد و یک باز در توالی موجود باشد این بازهای می‌توانند اشتباهی جفت شوند (جفت شدگی اشتباهی). به جز تمام عواملی که در بالا به آنها اشاره شد یکی دیگر از این عوامل ترکیب

چندعامل است. ما همچنین الگوریتم نیدلمن- وونش را تنها برای MRUR چندعاملی به کار ببریم تا بار برنامه نویسی پویای این الگوریتم را کاهش دهیم. برای مثال در مورد CAR1 که در شکل 3 نشان داده شده است، در اینجا تنها یک MRUR بین MR1 و MR2 وجود دارد. از طریق دخول 2 ما یک دخول بین MR1 و MR2 اعمال می‌کنیم و 2M1I3M ((CIGAR format) را بدست می‌آوریم. در مورد CAR2، یک RMUR وجود دارد و عامل ترکیبی است. بنابراین ما می‌توانیم نیدلمن-وونش را بر RMUR اعمال کنیم و 4M1I1M یا 5M1I (CIGAR format) را بدست اوریم.



شکل 2. دیاگرام دسته‌بندی عوامل جدایی بین نواحی جور و ناجور. پنل a عوامل جدایی LMUR را نشان می‌دهد. پنل b عوامل جدایی .RMUR را نشان می‌دهد. چندین عامل برای جدایی در LMUR و RMUR وجود دارد (جفت شدن اشتباهی، دخول، حذف و مرکب). پنل c عوامل جدایی MRUR را نشان می‌دهد که می‌توان آنها را با

جزئیات دقیق دسته بندی کرد. بلوک زرد نشان می دهد که هر دو بلوک ریدها و توالی کاملا جور هستند. بلوک خاکستری تیره به معنای منطقه عدم جور است. بلوک سفید نشان دهنده‌ی منطقه‌ی gap بین بلوک است. بلوک قرمز نشان دهنده‌ی محل همپوشانی دو بلوک است. نهایتا بلوک ابی نشان می‌دهد که بیش از دو عامل جدایی روی بلوک رید و توالی وجود دارد.

نقشه‌برداری هیبرید ممتد

نقشه برداری هیبرید ممکن است برخی از CAR ها را از دست بدهد و در نتیجه به شکست نقشه (unmapped) منجر شود. دلایل اصلی از دست دادن CARها احتمالا یکی از موارد زیر است: 1) وجود خطاهای بسیار زیاد در رید 2) وجود تداد زیادی از q-grams های با فراوانی بالا در رید و 3) وجود تعداد زیادی از CARهای با رتبه‌ی بالا. در مورد اول، یک q-gram بلند بر اساس روش کاوش اول به محدوده‌های q-gram خالی منجر می‌شود. در مورد دوم، روش کاوشی دوم به میزان زیادی باعث از دست رفتن q-grams های حاوی اطلاعات می‌شود. در مورد سوم، CAR های با رتبه پایین اما حاوی اطلاعات به دلیل روش کاوشی سوم از بین می‌روند. ما از نقشه‌برداری هیبرید ممتد برای نواحی بدون نقشه استفاده می‌کنیم، که از q-gramهای کوتاه‌تر شامل برخی از q-gramهای با فراوانی بالا استفاده می‌کند و CARهای رده بندی شده را بیشتر امتداد می‌دهد. روی هم رفته، این تغییرات حساسیت تکنیک را افزایش می‌دهد.

اجرا

ما HIA را برروی جاوا اجرا کردیم تا پلتفرم‌های چندگانه را پشتیبانی کند (بنگرید به فایل 2). HIAتوالی مرجع FASTAرا به عنوان ورودی می‌گیرد ، شاخص هیبریدی را می سازد، و سپس آن را نمایش می دهد: فایل (.sa) SA ، فایل (.idx) HT، فایل فشرده‌ی توالی (.seq) و فایل اطلاعات توالی مرجع (.seqInfo) . برای همترازی HIA شاخص هیبرید و یک فایل FASTQ مورد نظر را به عنوان ورودی می‌گیرد همترازی‌های نقشه‌برداری

شده و بدون نقشه را به فرمت SAM نمایش می‌دهد. برای کاهش مشکل در خواندن ریدها و نوشتن نتایج نقشه برداری، ما روش همترازی را به روش خواندن، نقشه برداری و نوشتن تقسیم کردیم. هر روش در یک رشته مستقل اجرا می شود و با استفاده از سه صف پردازش داده برنامه ریزی شده است. علاوه بر این، HIAهمچنین گزارشی را شامل می شود که شامل یک فایل خلاصه نتایج نقشه برداری و دو نمودار پیمایش نمودار نرخ نقشه برداری است تا کاربر را در مورد نتایج نقشه برداری اگاه سازد. علاوه بر این، در طول همترازی، HIA ورودی FASTQ را خلاصه کرده و امارهای اساسی و اصلاعات کیفیت باز را گزارش می‌کند: امار FASTQ؛ کیفیت باز در هر موقعیت رید، به صورت یک نمودار خطی؛ کیفیت باز به صورت یک نقشه‌ی حرارتی؛ و امتیاز کیفیت، به صورت یک نمودار جعبه‌ای. این خلاصه برای نعیین کیفیت داده‌های توالی NGS مفید است. ما برای رسم تمام نمودارها از JFreeChart استفاده کردیم.

نتایج و بحث

ارزیابی مجموع داده‌ها و معیارهای ارزیابی

ما با استفاده از Mason **از ساختمان** GRCH37 ژنوم انسان 6 پایگاه داده ساختیم. دو مورد از آنها مجموع داده‌های illuminaمانند جفت نشده هستند که به ترتیب دارای یک میلیون رید 100 جفت بازی و یک میلیون رید 150 جفت‌بازی هستند، که MASON با پارامترهای "illumina hn 2 sq n 100 N 1000000" و "illumina hn 2 sq n 150 N 1000000" شبیه سازی کرده است. دو مجموعه داده‌ی بعدی مجموعه داده‌های جفت شده‌ی illumia – مانند هستند که MASON با پارامترهای - 'illumina -hn 2 -sq -rn 2 -mp -ll 375 -le 100 n 100 –N 1000000' و 'illumina -hn 2 -sq -rn 2 -mp -ll 375 -le 100 -n 150 -N 1000000' شبیه‌سازی کرده است. MASON نتایج همترازی صحیح 6 مجموعه داده‌های رید را در فرمت SAM تولید کرد . پارامترهای خط فرمان دقیق و توضیحات برای هر مجموعه داده در فایل پیوست شماره 1 یافت می شود.

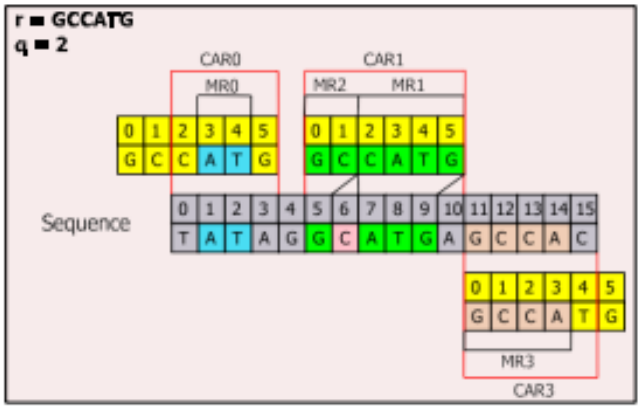
برای ارزیابی عملکرد در داده‌های واقعی، مجموعه داده‌های Illumina را از یک مطالعه توالی یابی مجدد انسان و مجموعه داده 454 از پروژه 1000 Genomes Project Pilot (1000 Genomes Project Consortium2010) بدست اوردیم. مجموعه‌داده‌ی Illumina حاوی 1,296,188,286 101 bp × 99 bp رید جفت شده است. شماره دسترسی NCBI مجموعه داده 454 عبارت است از SRR003161 که شماره‌ی دسترسی کوتاهی است و حاوی 1,375,489 رید با ظول میانگین 355 bp است. ما سه مجموعه داده‌ی آزمایشی از مجموعه داده‌های Illumina و یک مجموعه داده‌ی آزمایشی از مجموعه داده‌های 454 مانند (1) یک میلیون رید جفت شده‌ی HiSeq، (2) یک میلیون رید HiSeq با 101 paired-end جفت بازی ، 3) کل ریدهای HiSeq paired-endو 4) تمام ریدهای 454 .

چهار معیار ارزیابی زیر را در مطالعه‌ی ازمایشی مورد استفاده قرار دادیم: همترازی (٪)، منحصر به فرد(٪)، Q10 (%) و زمان (S). Aligned (٪) نشان دهنده درصد ریدهای همتراز شده نسبت به کل ریدهاست و نشان دهنده نرخ کلی نقشه برداری است. Unique (٪) درصد دفعات ریدهای همتراز شده‌ی منحصربفرد را نسبت به کل ریدها نشان می‌دهد و به MAPQ ≥ 1 اشاره می کند. Q10 (%) کسری از ریدهای نقشه‌برداری شده MAPQ ≥ 10 را اندازه‌گیری می‌کند. Timeزمان سپری شده (ثانیه) شامل زمان بارگذاری فهرست و زمان همترازی است. در مورد مجموعه داده‌های شبیه سازی شده، Err% درصد ریدهایی را نشان می‌دهد که اشتباهی همتراز شده‌اند را نسبت به ریدهای با Unique (%) و Q10 (%) رضایت بخش اندازه‌گیری می‌کند. ما مفهوم همترازی صحیح را از Langmead و Salzberg اخذ می‌کنیم،آنها همترازی صحیح را تعیین کردند به شرطیکه همترازی در همان رشته قرار داشته باشد و چپ‌ترین موقعیت در حدود 50 جفت‌بازی از موقعیت تعیین شده باشد.

نتایج ارزیابی

برای ارزیابی کارایی HI، آن را با BWA، Bowtie2، SOAP2 و SeqAlto روی 6 مجموعه داده‌ی شبیه سازی شده و دو مجموعه داده‌ی واقعی مقایسه کردیم. در تمام نست‌ها، از ساختمان GRCH37 ژنوم انسان به عنوان توالی

مرجع برای همترازی استفاده کردیم. ما همترازی را با استفاده از کامپیوتری دارای دو پردازشگر Intel Xeon 6-Core X5670 2.93-GHz و 48 GB RAM انجام دادیم. تمام ابزارهای همترازی را با یک رشته برای همترازی به جز تست‌های چند رشته‌ای اجرا کردیم.



شکل 3. یافتن MR ها و CAR ها . MRها و CAR ها را با قرار دادن رید (r = GCCATG) و شاخص هیبرید ساخته شده پیدا کنید.

عملکرد ایجاد شاخص

جدول 1 نتایج ایجاد شاخص را نشان می‌دهد. این نتایج بیان کننده‌ی زمان شاخص گذاری HIA با سایر ابزارهای همترازی قابل مقایسه است. به ویژه، HIAقادر است زمان ایجاد شاخص را با استفاده از چندین رشته از کامپیوترهای چند هسته‌ای مدرن کاهش دهد (فایل پیوست شماره 1).

از نظر ساخت SA، ما چندین تست ایجاد شاخص را انجام دادیم و نتایج حاصل از الگوریتم ایجاد شاخص خودمان و divsufsortرا که یکی از بهترین الگوریتم های ساخت SA است مقایسه کردیم. واضح است که الگوریتم divsufsort از الگوریتم ما در ساخت SA از ژنوم انسان برتری است. ما از طریق اجرای رشته‌های چندگانه در ساخت SA ژنوم انسان، این مشکل را رفع کردیم (بنگرید به فایل پیوست مقایسه‌ی عملکرد رشته‌های چندگانه و الگوریتم divsufort در فایل پیوست1). علاوه بر این، از آنجا که ساخت SA نیاز اولیه در همترازی داده‌های NGS است، ما معتقدیم که این امر نمی‌تواند مشکل جدی در کاربرد عملی باشد.

نتایج ریدهای شبیه‌سازی شده‌ی single-end

ما 6 ابزار همترازی را با تنظیمات پارامترهای متفاوت برای ریدهای single-end 100، 150، 250 و 400 جفت بازی انجام دادیم. تمام نتایج در پیوست 1 قابل مشاهده است. SeqAlto تنها می‌تواند ریدهای Illumina- مانند را همتراز کند و بنابراین از تست برای مجموع داده‌های 454- مانند کنار گذاشته شد.

جدول 2 بهترین نتایج را از دیدگاه حساسیت و دقت نشان می دهد. برای هر دو مجموع داده‌های Illumina–مانند و مجموعه داده‌های 454 –مانند ، HIAبه طور قابل توجهی سریع تر از سایر ابزارهای همترازی به جز BWA MEMو SOAP2 عمل می‌کند. SOAP2بسیار سریع است، اما حساسیت آن به اندازه‌ی HIA نیست. BWA کمی دقیق تر است، اما حساسیت آن برای مجموعه داده های شبه ILLUMINA به اندازه‌ی HI نیست، با این حال، برای مجموع داده‌های 454- مانند HIAحساس تر و دقیق تر از BWA است. Bowtie2 از نظر حساسیت شبیه به HIA است، اما دقت آن برای هر دو مجموعه به اندازه‌ی HIA نیست. SeqAltoکمی دقیق تر است، اما برای مجموع داده‌های شبه – illumine حساسیت آن به اندازه‌ی HIA نیست است. BWA MEMدقیق تر و حساس تر از HIA برای مجموع داده‌های ILLUMINA - مانندست ، اما دقت و حساسیت آن برای مجموع داده‌های 454—مانند به اندازه‌ی HIA نیست.

جدول 1. نتایج ایجاد شاخص

Aligner	Options	Time	Memory (GB)	Size (GB)
HIA	-t 1 -q 14	165	20.32	12.63
HIA	-t 12 -q 14	28	20.47	12.63
BWA		65	4.53	5.40
bowtie2		99	5.35	4.10
soap2		55	3.39	5.90
seqAlto	-l 0 genome.fa 28	33	37.99	22.40
seqAlto	-l 1 genome.fa 22	12	13.19	5.52

Time measurement is elapsed time (minute). Memory is the peak memory for the index construction. Size is the sum of all generated files

نتایج ریدهای شبیه‌سازی شده‌ی paired-end

ما همچنین شش ابزرار همترازی را با همان تنظیمات پارامتر در ریدهای single-end، برای ریدهای paired-end 100 و 150 جفت بازی اجرا کردیم. تمام نتایج را می توان در فایل پیوست شماره‌ی 1 مشاهده کرد.

جدول 3 بهترین نتایج را از نظرحساسیت و دقت نشان می دهد. برای هر دو مجموعه داده، HIAبه طور قابل توجهی سریع تر از سایر همترازها به جز BWA MEM و SOAP2 عمل می‌کند و در عین حال از حساسیت و دقت خوبی‌هم برخوردار است. SOAP2بسیار سریع است، ولی حساسیت آن به اندازه‌ی HIAنیست، BWAو SeqAlto از نظر حساسیت شبیه HIA هستند، BEM MEMدقیق تر از سایر ابزارهای همترازی است. Bowtie2نسبت بهHIA ، BWAو SeqAlto حساست پایین‌تری دارد.

جدول 2. نتایج ریدهای single-end شبیه سازی شده

Aligner	Time	% Aligned	% Unique [% Err]	% Q10 [% Err]
(a) Illumina-like 100 bp reads (unpaired)				
HIA	464	100.00	96.57 [0.4314]	95.80 [0.2151]
BWA	1242	98.11	94.73 [0.1711]	94.60 [0.1562]
BWA MEM	265	100.00	96.30 [0.0497]	95.27 [0.0153]
Bowtie2	1291	99.95	99.63 [2.4252]	94.22 [0.0208]
SOAP2	264	79.37	76.27 [0.4679]	
SeqAlto	1459	99.69	96.33 [0.2861]	96.04 [0.2156]
(b) Illumina-like 150 bp reads (unpaired)				
HIA	530	100.00	97.56 [0.2552]	97.26 [0.1643]
BWA	2464	98.00	95.55 [0.0953]	95.48 [0.0866]
BWA MEM	355	100.00	97.36 [0.0210]	96.41 [0.0053]
Bowtie2	2069	99.97	99.87 [1.6663]	95.99 [0.0094]
SOAP2	525	68.72	66.78 [0.2806]	
SeqAlto	3608	99.68	97.25 [0.1947]	97.10 [0.1490]
(c) 454-like 250 bp reads (unpaired)				
HIA	1009	99.96	98.28 [0.4189]	94.38 [0.1772]
BWA-SW	3157	99.86	97.61 [0.6735]	94.38 [0.0357]
BWA MEM	1497	100.00	97.92 [0.0767]	97.26 [0.0346]
Bowtie2	2947	99.59	83.40 [0.5980]	36.44 [0.0011]
(d) 454-like 400 bp reads (unpaired)				
HIA	1378	99.76	98.48 [0.1557]	96.17 [0.0397]
BWA-SW	5144	100.00	95.89 [0.2084]	94.00 [0.0284]
BWA MEM	2426	99.99	98.46 [0.0471]	97.98 [0.0238]
Bowtie2	6597	99.96	88.35 [0.3048]	32.93 [0.0000]

Time measurement is elapsed time (second). Unique refers to MAPQ ≥ 1 if MAPQ available. Q10 refers to MAPQ ≥ 10

نتایج مجموع داده‌های واقعی

ما شش ابزار همترازی با تنظیمات پارامترهای مختلف را برای یک رید paired-end و دو رید single-end اجرا کردیم. تمام نتایج را می توان در فایل پیوست شماره‌ی 1 مشاهده کرد.

جدول 4 بهترین نتایج را از تعداد کل ریدهای همتراز شده به ترتیب برای ریدهای single-end و ریدهای paired-end نشان می‌دهد. برای دو single-end، HIAو BWA MEM رتبه بالاتری از سایر ابزارهای همترازی از نظر سرعت و تعداد کل ریدهای همتراز شده داشتند . برای ریدهای paired-end ، می توان گفت که درصدهمترازی Bowtie2بالاتر از سایر ابزارهای همترازی است، اما HIA سریع تر از تمام ترازهای دیگر است به جزSOAP2 .

جدول 3. نتایج ریدهای paired-end شبیه سازی شده

Aligner	Time	% Aligned	% Unique [% Err]	% Q10 [% Err]
(a) Illumina-like 100 bp reads (paired)				
HIA	1009	99.96	97.17 [0.0859]	96.75 [0.0510]
BWA	2554	99.79	97.70 [0.0954]	97.49 [0.0692]
BWA MEM	646	99.99	98.03 [0.0282]	97.95 [0.0172]
Bowtie2	1691	97.13	97.09 [1.2711]	93.73 [0.0128]
SOAP2	586	84.54	82.75 [0.3356]	
SeqAlto	2945	99.61	97.15 [0.0833]	97.02 [0.0788]
(b) Illumina-like 150 bp reads (paired)				
HIA	1153	99.99	98.10 [0.0780]	97.89 [0.0537]
BWA	5380	99.78	98.17 [0.0983]	98.06 [0.0876]
BWA MEM	649	99.99	98.43 [0.0137]	98.39 [0.0083]
Bowtie2	2348	97.13	97.12 [0.9904]	94.14 [0.0083]
SOAP2	856	75.25	74.04 [0.3487]	
SeqAlto	7191	99.58	97.80 [0.0723]	97.74 [0.0698]

Time measurement is elapsed time (second). Unique refers to MAPQ ≥ 1 if MAPQ available. Q10 refers to MAPQ ≥ 10

نتایج تست‌های چند رشته‌ای

ما 6 ابزار همترازی با 6 رشته و 12 حالت را برای کل ریدهای paired-end HiSeq اجرا کردیم. همه نتایج در جد.ل 5 امده است. HIAسریع تر از دیگر ابزارهای همترازی برای هر 6 رشته و 12 حالت است.

جدول 4. نتایج مجموع داده‌های واقعی

Aligner	Time	% Aligned	% Unique	% Q10
(a) Illumina 100 bp reads (unpaired)				
HIA	369	97.71	91.23	86.41
BWA	2877	85.87	81.82	81.68
BWA MEM	272	96.86	89.32	86.85
Bowtie2	1291	94.96	92.11	83.69
SOAP2	283	87.29	82.29	
SeqAlto	1567	89.16	85.22	84.60
(b) 454 400 bp reads (unpaired)				
HIA	964	99.05	96.90	95.92
BWA-SW	6369	99.53	96.48	92.63
BWA MEM	830	99.73	96.36	94.86
Bowtie2	6597	98.37	96.96	91.02
(c) Illumina 100 bp reads (paired)				
HIA	1111	91.53	87.48	85.25
BWA	2871	88.90	86.59	86.26
BWA MEM	690	93.49	90.55	89.80
Bowtie2	1646	93.13	91.60	84.93
SOAP2	725	82.18	79.56	
SeqAlto	3370	92.04	87.82	87.55

Time measurement is elapsed time (second). Unique refers to MAPQ ≥ 1 if MAPQ available. Q10 refers to MAPQ ≥ 10

جدول 5. نتایج تست‌های چند رشته‌ای

Aligner	Time (6 threads)	Time (12 threads)
HIA	932	505
BWA	4006	2586
BWA MEM	1162	645
bowtie2	1180	789
soap2	2217	1616
seqAlto	3945	2077

Time measurement is elapsed time (minute)

نتیجه‌گیری

ما یک ابزار همترازی جدید برای همترازی ریدهای کوتاه و بلند با یک ژنوم مرجع ایجاد کردیم.، HIAدارای دو شاخص، HT و شاخصSA است. HT قادر به جستجو مستقیم q-gram ااست ، و SA می تواند به سرعت q-gram طول متغییر را جستجور کند. بررسی ما نشان می‌دهد که ترکیب HT و SA از نظر سرعت نقشه‌برداری ریدهای توالی یابی NGS ژنوم مرجع مفید است. HIA همچنین از چند رشته‌ای بودن نقشه برداری پشتیبانی می‌کند. به

طور خاص، HIAبسیار سریعتر از سایر ابزارهای همترازی دیگر است؛ بنابراین ابزار ما برای همترازی مجموع داده‌های غول پیکر ایجاد شده با توالی یابی NGS مفید است.

دقت هم ترازی در توالی‌یابی مجدد بسیار مهم است، زیرا هدف اصلی همترازکردن پیدا کردن واریانس نسبت یه یک ژنوم مرجع است. اگر چه این واریانس‌ها (و یا خطاهای توالی‌یابی) باعث جفت شدن غیر دقیق ریدها با توالی مرجع می‌شوند با این حال، ابزارهای همترازی باید این ریدها به درستی نسبت به توالی‌های مرجع نقشه برداری کنند. با توجه به نتایج حاصل از تجزیه و تحلیل ازمایشی ، می توان نتیجه گرفت که HIA با چهار ابزار همترازی محبوب مقایسه قابل مقایسه است.

References

1. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
2. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour.2011;11:759–69.
3. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001;11:1725–9.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows– Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows– Wheeler transform. Bioinformatics. 2010;26:589–95.
6. Misra S, Agrawal A, Liao WK, Choudhary A. Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. Bioinformatics. 2011;27:189–95
7. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24:713–4.
8. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
9. Mu JC, Jiang H, Kiani A, Mohiyuddin M, Bani Asadi N, Wong WH. Fast and accurate read alignment for resequencing. Bioinformatics. 2012;28(18):2366–73.
10. Burrows M,Wheeler DJ. A block-sorting lossless data compression algorithm. Technical report 124, Palo Alto, CA, Digital Equipment Corporation; 1994.
11. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One. 2009;4:e7767.
12. Li H, Homer N. A survey of sequence alignment algorithms for nextgeneration sequencing. Brief Bioinform. 2010;11:473–83.
13. Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings of the 41st annual symposium on foundations of computer science IEEE Computer Society, Los Alamitos, CA, USA; 2000. p. 390–8.
14. Larsson NJ, Sadakane K. Faster suffix sorting. Theoret Comput Sci. 2007;387:258–72.
15. Pevzner PA, Waterman MS. Multiple filtration and approximate pattern matching.Algorithmica.1995;13(1/2):135–54.
16. Robertson SE. Understanding inverse document frequency: on theoretical arguments for IDF. J Document. 2004;60(5):503–20.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
18. JFreeChart. http://www.jfree.org/jfreechart/. Accessed 20 June 2012.
19. Mason. http://www.seqan.de/projects/mason.html. Accessed 15 May 2012.
20. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 2012;30(1):78–82.
21. divsufsort. https://code.google.com/p/libdivsufsort/. Accessed 10 Jan 2015.