

**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 نقشه­برداری 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 = s0s1 … sm−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 … 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 Σ = {A, C, G, T}). ما به جای A، C، G و T به ترتیب از اعداد 0، 1، 2 و 3 استفاده می­کنیم. بنابراین، هر q-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 4q + 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:



 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 SA امرتب 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، w بین 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 Q طول 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 ] |Q| − 1 [ 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′ ≪ 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′ ، (m′ − q + 1) 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 می­ذهذ. روش ساخت به ترتیب زیر است: اول تبدیل نمادهای نوکلئئوتید 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 4q + 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 λ = εm 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 :



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 = 1.5λ تنظیم کنیم، مقدار ریدها با خطاهای 1.5λ به 0.9 میل می­کند و ما می­توانیم از q-gram با طول m/(1.5λ + 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 و RMUR، 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 α را 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 β – α 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 بازهای αدر تولی همپوشانی شوند، می­تواند بازهای α + β را حذف کرد (حذف 2). اگر بازهای β 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).HIA توالی مرجع FASTA را به عنوان ورودی می­گیرد ، شاخص هیبریدی را می سازد، و سپس آن را نمایش می دهد: فایل SA (.sa) ، فایل HT (.idx) ، 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، HIA همچنین 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 paired‑end 101 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. Time زمان 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 ایچاد شاخص را نشان می­دهد. این نتایج بیان کننده­ی زمان شاخص گذاری HIA با سایر ابزارهای همترازی قابل مقایسه است. به ویژه، HIA قادر 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 divsufsort را 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 تنطیمات 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 محموع 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 ، HIA به طور قابل توجهی سریع تر از سایر ابزارهای همترازی به جز BWA MEM و SOAP2 عمل می­کند. SOAP2 بسیار سریع است، اما حساسیت آن به اندازه­ی HIAنیست. 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 HIA حساس 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. SeqAlto کمی دقیق تر است، اما برای مجموع داده­های شبه – illumine حساسیت آن به اندازه­ی HIA نیست است. BWA MEM دقیق تر و حساس تر از HIA برای مجموع داده­های ILLUMINA- ماننداست ، اما دقت و حساسیت آن برای مجموع داده­های 454—مانند به اندازه­ی HIA نیست.

جدول 1. نتایج ایجاد شاخص



**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، HIA به 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. SOAP2 بسیار سریع است، ولی حساسیت آن به اندازه­ی HIAنیست، BWA و SeqAlto از نظر حساسیت شبیه HIA هستند، BEM MEM دقیق تر از سایر ابزارهای همترازی است. Bowtie2 نسبت به HIA، BWA و SeqAlto حساست پایین­تری دارد.

جدول 2. نتایج ریدهای single-end شبیه سازی شده



**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، HIA و 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 Bowtie2 بالاتر از سایر ابزارهای همترازی است، اما HIA سریع تر از تمام ترازهای دیگر است به جز SOAP2.

جدول 3. نتایج ریدهای paired-end شبیه سازس شده



**نتایج تست­های چند رشته­ای**

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

جدول 4. نتایج مجموع داده­های واقعی



جدول 5. نتایج تست­های چند رشته­ای



**نتیجه­گیری**

ما یک ابزار همترازی جدید برای همترازی ریدهای کوتاه و بلند با یک ژنوم مرجع ایجاد کردیم.، HIA دارای دو شاخص، HT و شاخص SA 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، HIA بسیار 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](http://www.seqan.de/projects/mason.html.%20Accessed%2015%20May%202012).

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.