مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره چهارم. ص ۴۳۵-۴۴۴، ۱۳۹۹

پیاده‌سازی و بهینه‌سازی مرحله حاشیه‌نویسی و تفسیر داده‌های نسل نوین توالی‌یابی برای بیماری ناشنوایی غیرسندرمیک اتوزومی مغلوب

مهدی شاه‌حسینی[۱]، نیوشا مولوی[۲]، محمدامین طباطبائی‌فر[۳]، محمدرضا صحتی[۴]*

• دریافت مقاله: ۹۸/۱۱/۱۵ • پذیرش مقاله: ۹۹/۴/۱

مقدمه: دقت و زمان لازم برای آنالیز داده‌های نسل نوین توالی‌یابی (NGS) بسته به ابزارهای استفاده شده برای هم‌ترازی، فراخوانی واریانت، حاشیه‌نویسی، اولویت‌بندی و فیلترینگ واریانت‌ها، تسلط افراد به تحلیل و تفسیر داده‌ها و ظرفیت محاسباتی آزمایشگاه متفاوت بوده و بهینه‌سازی آن یک مسئله چالش برانگیز است.

روش: یک نرم‌افزار کاربردی به منظور بهینه‌سازی مرحله سوم آنالیز داده‌های NGS طراحی و با زبان برنامه‌نویسی C# پیاده‌سازی شد. در این مطالعه روند حاشیه‌نویسی، فیلترینگ و تفسیر داده‌های NGS برای بیماری ناشنوایی غیرسندرمیک با وراثت اتوزومی مغلوب به طور اختصاصی بهینه شده است.

نتایج: داده مربوط به بیماری که دارای یک جهش بیماری‌زای تأیید شده توسط آنالیز ژنتیکی فامیلی بود و تعداد واریانت‌های اولیه در فایل حاصل از آنالیز مراحل اولیه وی شامل ۶۷۱۸۲۹ واریانت می‌شد توسط نرم‌افزار پیاده‌سازی شده مورد تحلیل قرار گرفت. بعد از انجام مرحله اولویت بندی خودکار واریانت‌ها با استفاده از فایل BED، تعداد واریانت‌ها ۵۰۸ شد. با توجه به شجره‌ی خانوادگی بیمار در مرحله بعدی آنالیز واریانت‌های هوموزیگوت انتخاب شدند و به این ترتیب تعداد واریانت‌ها به ۱۸۷ رسید. بعد از اعمال آستانه فراوانی جمعیتی ۰/۶٪ در پایگاه‌های داده genomAD و ExAC تعداد واریانت‌های باقی‌مانده به ترتیب ۱۱۰ و ۳ واریانت شد. پاتوژن شناسایی شده نهایی با نتیجه‌ی توالی‌یابی سنگر که به منظور بررسی هم‌تفکیکی واریانت مورد نظر در خانواده انجام شده بود، همخوانی داشت. مدت زمان آنالیز توسط نرم‌افزار طراحی شده بر روی یک کامپیوتر شخصی متوسط ۱۵ دقیقه بود.

نتیجه‌گیری: نرم‌افزار طراحی شده کاملاً گرافیکی و بدون نیاز به کدنویسی است که علاوه بر قابلیت مقایسه و یکپارچه کردن فایل‌های ورودی، امکان ایجاد یک دیتابیس داخلی از فایل‌های آنالیز شده، امکان اعمال محدودیت ناحیه آنالیز و آستانه‌گذاری بر فیلدهای مختلف پایگاه‌های داده انتخابی توسط کاربر را دارد.

کلید واژه‌ها: نسل نوین توالی‌یابی، حاشیه‌نویسی، تعیین اثر واریانت، فیلترینگ واریانت‌ها

• **ارجاع:** شاه حسینی مهدی، مولوی نیوشا، طباطبائی‌فر محمدامین، صحتی محمدرضا. پیاده‌سازی و بهینه‌سازی مرحله حاشیه‌نویسی و تفسیر داده‌های نسل نوین توالی‌یابی برای بیماری ناشنوایی غیر سندرمیک اتوزومی مغلوب. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۴): ۴۴-۴۳۵.

۱. کارشناسی ارشد مهندسی پزشکی- بیوالکتریک، گروه بیوانفورماتیک، دانشکده فناوری‌های نوین پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲. کارشناسی ارشد ژنتیک انسانی، گروه ژنتیک و بیولوژی مولکولی، دانشکده پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۳. دکتری تخصصی ژنتیک پزشکی، دانشیار، گروه ژنتیک و بیولوژی مولکولی، دانشکده پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۴. دکتری تخصصی مهندسی پزشکی- بیوالکتریک، استادیار، گروه بیوانفورماتیک، دانشکده فناوری‌های نوین پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران

* **نویسنده مسئول:** محمدرضا صحتی

آدرس: اصفهان، خیابان هزارجریب، دانشگاه علوم پزشکی اصفهان، دانشکده فناوری‌های نوین پزشکی، گروه بیوانفورماتیک

• **شماره تماس:** ۰۳۱۳۷۹۲۳۸۵۴ • **Email:** mr.sehhati@amt.mui.ac.ir

مقدمه

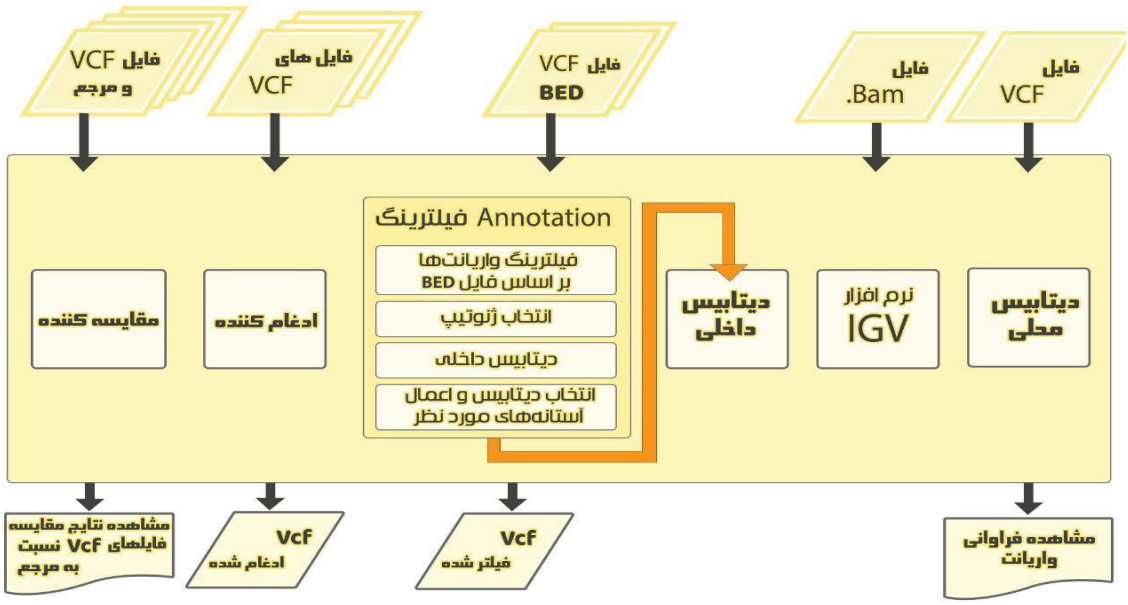
ناشنوایی ارثی پس از عقب‌ماندگی ذهنی دومین اختلال مادرزادی شایع در ایران محسوب می‌شود [۱]. ناشنوایی اختلالی چندعاملی است که عوامل محیطی یا ژنتیکی یا ترکیبی از هر دو در بروز آن نقش دارند [۲]. حداقل ۶۰ درصد از ناشنوایی‌ها علت ژنتیکی دارند و شامل انواع سندرمی و غیرسندرمی می‌شوند. بیش از ۷۰ درصد از موارد ارثی، غیر سندرمی است و الگوی اصلی وراثت آن، اتوزومی مغلوب است [۳]. از آنجایی‌که زمان طلایی برای درمان ناشنوایی تعریف شده است، شناسایی نوزادان ناشنوا در ماه‌های اول زندگی برای رشد زبان و گفتار، موفقیت‌های اجتماعی، عاطفی و تحصیلی آن‌ها بسیار ضروری است. تشخیص ژنتیکی می‌تواند برای تصمیم‌گیری و ارزیابی موفقیت درمان‌هایی مانند پیوند حلزون و درمان‌های پیشرفته آتی اهمیت داشته باشد [۱]. تاکنون بیش از ۷۰ ژن در رابطه با ناشنوایی غیرسندرمی اتوزومی مغلوب شناخته شده است؛ بنابراین می‌توان گفت ناشنوایی از جمله ناهمگن‌ترین صفات ژنتیکی انسان است [۴–۷]. به دلیل تعداد زیاد ژن‌های درگیر در ناشنوایی، طویل بودن اکثر این ژن‌ها و همچنین تعداد کم نقاط جهش خیز، بهترین رویکرد تشخیصی بعد از منفی شدن نتیجه توالی‌یابی سنگر ژن *GJB2*، استفاده از توالی‌یابی کل اگزون‌ها است. پیشرفت‌های اخیر در زمینه تکنولوژی‌های توالی‌یابی ژنوم این فرصت را به وجود آورده است که بتوان ژنوم هر انسانی را در کمتر از یک هفته به طور کامل توالی‌یابی کرد و به این ترتیب بازدهی تشخیص افزایش یافته است. با وجود تکنولوژی نسل نوین توالی‌یابی (Next Generation Sequencing) NGS امروزه محدودیتی در رابطه با توالی‌یابی یک، چندین یا همه ژن‌های انسانی وجود ندارد؛ اما چالش اساسی در این زمینه شناسایی سریع و دقیق یک یا دو واریانت عامل بیماری ژنتیکی از بین خطاهای توالی‌یابی و میلیون‌ها پلی‌مورفیسم تک نوکلئوتیدی است. آنالیز داده‌های NGS در سه مرحله کلی انجام می‌شود. مرحله اول آن شامل فرآیند تولید داده اولیه و گزارش کیفیت آن است. در مرحله دوم فرآیند نگاشت داده توالی‌یابی بر روی ژنوم مرجع و فراخوانی واریانت‌ها یا همان تعیین تغییرات توالی نسبت به ژنوم مرجع انجام می‌گیرد. در مرحله سوم واریانت‌های فراخوانی شده در مرحله قبل به منظور تعیین تأثیر این واریانت‌ها بر بیماری‌ها یا سایر صفات مورد بررسی قرار خواهند گرفت. بدین‌منظور اطلاعاتی از قبیل موقعیت ژنومیکی واریانت مورد نظر، نقش ژن و پروتئینی که واریانت در آن واقع شده، اثر واریانت بر عملکرد پروتئین و فراوانی جمعیتی واریانت مورد نیاز است [۸]. اطلاعات فوق از طریق دسترسی آزاد به پایگاه داده‌های جمعیتی نظیر ExAC، GenomeAD و سایر پایگاه‌ها و ابزارهای تخمین و پیش‌بینی اثر واریانت‌ها قابل دستیابی است. با استفاده از پایگاه‌های داده ذکر شده، تفسیر و آنالیز اثر بالینی یک واریانت کار سخت و پیچیده‌ای به نظر نمی‌رسد؛ اما استخراج این اطلاعات برای صدها هزار واریانت گزارش شده در فایل VCF (Variant Calling Format) مربوط به هر بیمار کاری بسیار پرزحمت و زمان‌بر است؛ بنابراین جمع‌آوری کلیه اطلاعات مربوط به یک واریانت با سرعت و دقت بالا در قالب یک فایل به منظور اجرای سریع‌تر مراحل بعدی آنالیز که اولویت‌بندی و فیلترینگ واریانت‌ها است معقول به نظر می‌رسد [۹]. Krunic و همکاران، یک چارچوب تحت وب به نام VARIFI را به منظور آنالیز مراحل دوم و سوم داده‌های NGS ارائه کردند. این ابزار دارای محدودیت در آپلود حجم فایل ورودی، عدم قابلیت ارزیابی فایل VCF و مخصوص تکنولوژی IonTORRENT است. همچنین کاربر قادر به تغییر آستانه‌ها و انتخاب پایگاه‌های داده مورد نظر در گام‌های مختلف آنالیز نیست. از مزیت‌های این ابزار این است که قادر به فیلتر کردن خودکار واریانت‌ها با استفاده از فایل (Browser Extensible Data) BED است [۱۰]. Wang و همکاران، پلت‌فرم تحت وبی را به نام wANNOVAR به منظور آنالیز مرحله سوم داده‌های NGS ارائه کردند. این ابزار نیز دارای محدودیت در آپلود حجم فایل VCF ورودی و عدم فیلتر کردن خودکار واریانت‌ها با استفاده از فایل BED است. در این ابزار تغییر آستانه‌ها توسط کاربر امکان‌پذیر است؛ اما امکان انتخاب پایگاه‌های داده مورد نظر کاربر وجود ندارد [۱۱]. چندین نرم‌افزار نیز در این حوزه توسعه یافته است، Chang و Wang، نرم‌افزاری را به نام ANNOVAR طراحی کردند. کاربر برای استفاده از این ابزار نیاز به دانلود مجموعه‌ای از پایگاه‌های داده با حجم بالا و مهارت کدنویسی در محیط لینوکس دارد. از مزیت‌های ابزار فوق فراهم کردن انواع پایگاه داده برای دانلود و وجود انجمن پشتیبانی بسیار قوی است [۱۲]. Cingolani و همکاران نرم‌افزاری را به نام Snpeff به منظور آنالیز مرحله سوم NGS طراحی کردند. از معایب این ابزار استفاده از پایگاه داده محدود به واریانت‌های تک نوکلئوتیدی می‌باشد [۱۳]. هدف از این مطالعه طراحی یک نرم‌افزار کاربردی با گرافیکی کاربرپسند و بدون نیاز به کدنویسی است که قابلیت مقایسه و یکپارچه کردن فایل‌های

VCF را دارا است. علاوه بر این امکان ایجاد یک دیتابیس داخلی، فیلتر کردن خودکار واریانت‌ها با استفاده از فایل BED، امکان آستانه‌گذاری برای فیلدهای مختلف پایگاه‌های داده مورد نظر و انتخاب آن‌ها بدون نیاز به دانلود در نرم‌افزار طراحی شده وجود دارد. در ابزار طراحی شده همچنین قابلیت مشاهده واریانت‌های گزارش شده توسط IGV به صورت گرافیکی برای کاربر فراهم شده است. در این مطالعه روند آنالیز مرحله سوم NGS برای بیماری ناشنوایی غیرسندرمیک با وراثت اتوزومی مغلوب اختصاصی شده است.

روش

در فرآیند حاشیه‌نویسی واریانت‌ها، ویژگی‌های مختلف هر واریانت به منظور تشخیص واریانت پاتوژن از میان پلیمورفیسم‌های خنثی به دقت بررسی می‌شود. مهم‌ترین ویژگی‌های یک واریانت شامل جایگاه واریانت مورد نظر در ژنوم (اینترونی، اگزونی یا UTR)، اثر آن بر توالی پروتئین (ایجاد کدون خاتمه، حذف کدون شروع و ایجاد کدون جدید)، اثر بر فرآیند ویرایش، ژنوتیپ واریانت (هتروزیگوت و هوموزیگوت)، میزان فراوانی آن در جمعیت و ارتباط ژن حاوی واریانت با ناشنوایی است. اطلاعات ذکر شده برای واریانت‌های شناخته شده در پایگاه‌های داده مختلفی فراهم شده است. از آنجایی که هدف از مطالعه یافتن واریانت پاتوژن است پس از گردآوری اطلاعات عمومی و اختصاصی هر واریانت، آن‌ها بر اساس میزان اثر مخربی که بر ساختار و توالی پروتئین می‌گذارند مرحله به مرحله فیلتر می‌شوند. به دلیل عدم وجود یک پروتکل جهانی برای ارزیابی ابزارها و الگوریتم‌های موجود، پیشنهاد یک روند پردازشی بهینه که هدف اصلی این مطالعه است می‌تواند موجب ارتقاء و افزایش سرعت آنالیز داده‌های NGS، افزایش دقت نتایج و کاهش گزارش نتایج مثبت و منفی کاذب شود.

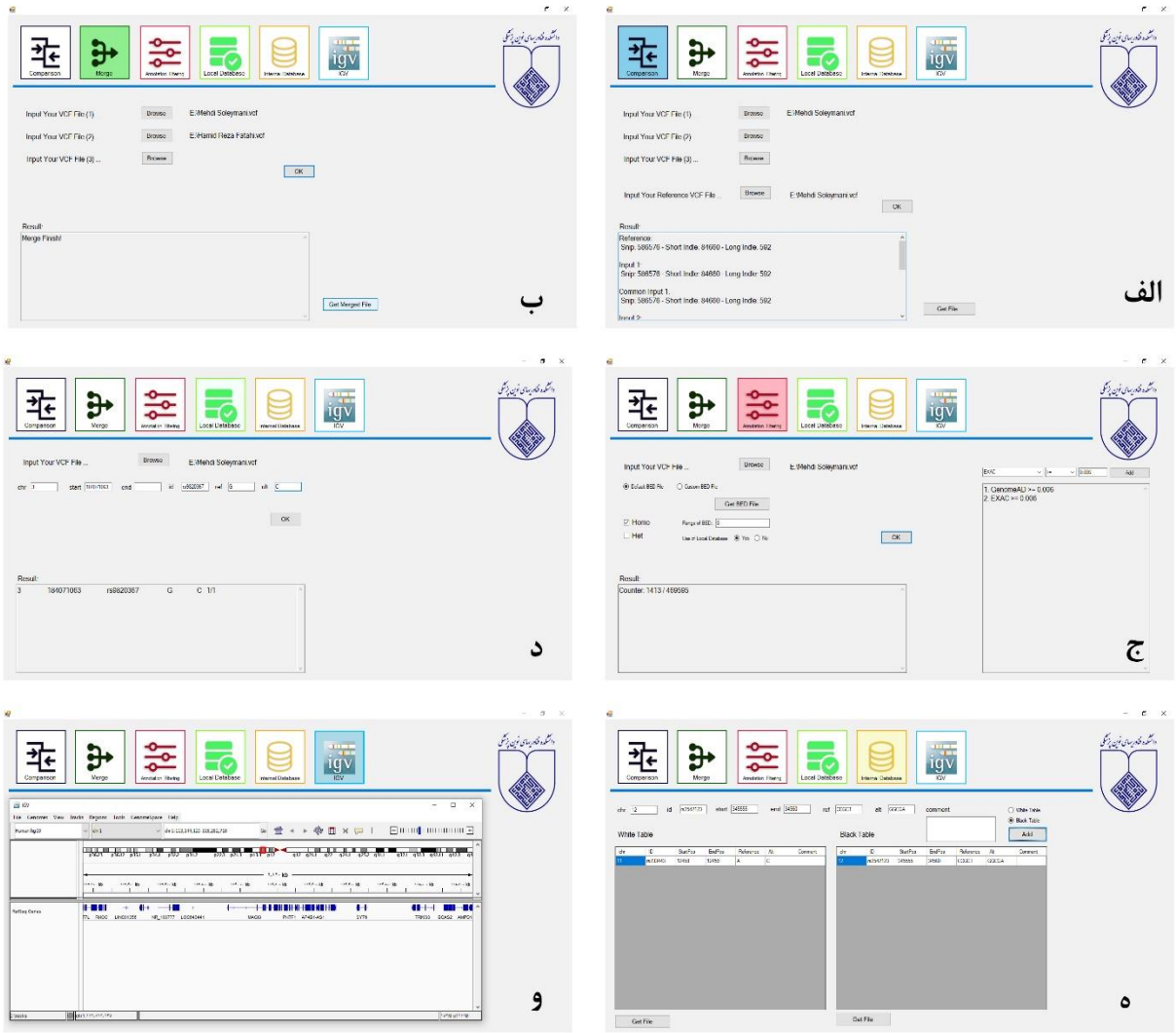
در این مطالعه به منظور بهینه‌سازی و تسریع مرحله حاشیه‌نویسی آنالیز NGS یک نرم‌افزار کاربردی با زبان برنامه‌نویس C# برای سیستم عامل ویندوز طراحی و پیاده‌سازی شده و با استفاده از داده‌های حاصل از توالی‌یابی کل اگزون‌های بیمار مبتلا به ناشنوایی غیرسندرمیک و با الگوی وراثتی اتوزومی مغلوب مورد ارزیابی قرار گرفته است. سخت‌افزار مورد استفاده در مطالعه حاضر یک لپ‌تاب مدل ASUS با پردازنده ۴ هسته‌ای ۲/۶ گیگاهرتز، ظرفیت حافظه RAM 4 گیگابایت و هارد با ظرفیت ۷۵۰ گیگابایت می‌باشد. چارچوب کلی نرم‌افزار طراحی شده شامل ورودی‌ها، خروجی‌ها و فرآیندهای در نظر گرفته شده در تحلیل فایل VCF در نمودار ۱ نشان داده شده است.



نمودار ۱: چارچوب کلی ورودی‌ها، خروجی‌ها و فرآیندهای در نظر گرفته شده در نرم‌افزار طراحی شده

در مورد ورودی‌های سیستم با توجه به نمودار ۱ اطلاعات هر فایل VCF به محض ورود به سیستم در دیتابیس محلی اضافه می‌شود. به منظور نمایش گرافیکی نتایج لازم است فایل BAM (Binary Alignment Map) متناظر با فایل ورودی نیز در اختیار نرم‌افزار قرار گیرد. یکی از تفاوت‌های آنالیز توالی‌یابی کل اگزوم با پنل‌های ژنی این است که در آنالیز کل اگزوم باید ژن‌های مرتبط با فنوتیپ مورد مطالعه مشخص شده باشند. در این مطالعه لیست ۷۵ ژن مرتبط با بیماری ناشنوایی ارثی غیرسندرمیک با الگوی وراثتی اتوزومی مغلوب به همراه شماره کروموزوم و موقعیت ابتدا و انتهای تمامی اگزون‌های آن‌ها تهیه شد. این فایل اصطلاحاً BED File نامیده می‌شود. این فایل به صورت پیش‌فرض در نرم‌افزار قابل دسترس است که می‌تواند با فایل انتخابی کاربر نیز جایگزین شود. موقعیت‌های ذکر شده در فایل BED از پایگاه داده Ensembl و بر اساس ژنوم مرجع hg19 استخراج شده است. همچنین به منظور اجرای فرآیند مقایسه در نرم‌افزار به یک فایل مرجع نیاز است که به فرمت VCF و به عنوان ورودی باید به نرم‌افزار ارائه شود.

فرآیندهای قابل اجرا در نرم‌افزار بر اساس منوی انتخاب شده توسط کاربر ورودی‌های لازم را توسط ماژول مربوطه دریافت و نتیجه مطلوب را در پنجره در نظر گرفته شده مطابق نمودار ۲ نمایش می‌دهند. در همه فرآیندها امکان ذخیره نتیجه حاصل به صورت یک فایل مستقل وجود دارد.



نمودار ۲: نمای ماژول‌های مختلف نرم‌افزار طراحی شده. الف) ماژول مقایسه کننده فایل‌های VCF، ب) ماژول یکپارچه‌سازی فایل‌های VCF، ج) ماژول حاشیه‌نویسی و تعیین پارامترهای فیلترینگ فایل VCF، د) ماژول ویرایش و نمایش پایگاه داده محلی، ه) ماژول ایجاد کننده پایگاه داده داخلی شامل واریانت‌های حائز اهمیت و متداول یا واریانت‌های مشکوک به خطا، و) ماژول نمایش گرافیکی نتایج

نتایج

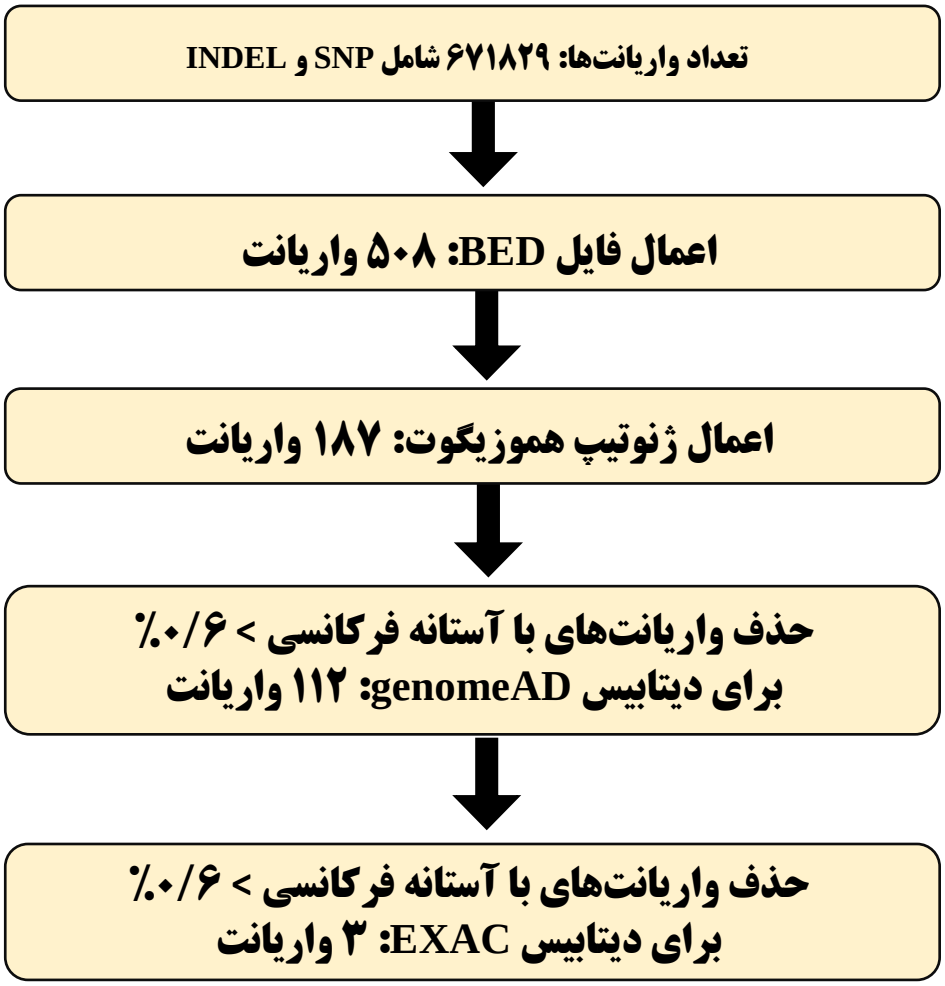
با توجه به هدف مطالعه، نرم‌افزاری برای پیدا کردن واریانت عامل بیماری در فردی مبتلا به ناشنوایی غیرسندرمیک با توارث اتوزومی مغلوب دارای ماژول‌های مقایسه کننده، ادغام کننده، حاشیه‌نویسی و فیلترینگ، نمایش گرافیکی نتایج، پایگاه داده داخلی به منظور حفظ اطلاعات ارزشمند فایل‌های آنالیز شده قبلی و یک پایگاه داده محلی به منظور تعیین وضعیت واریانت‌های خاص شناسایی شده توسط متخصص ژنتیک طراحی شد (نمودار ۲). قابلیت‌های این نرم‌افزار در مقایسه با سایر ابزارهای مشابه در جدول ۱ دسته‌بندی و گزارش شده است.

جدول ۱: مقایسه قابلیت‌های نرم‌افزار طراحی شده با موارد مشابه

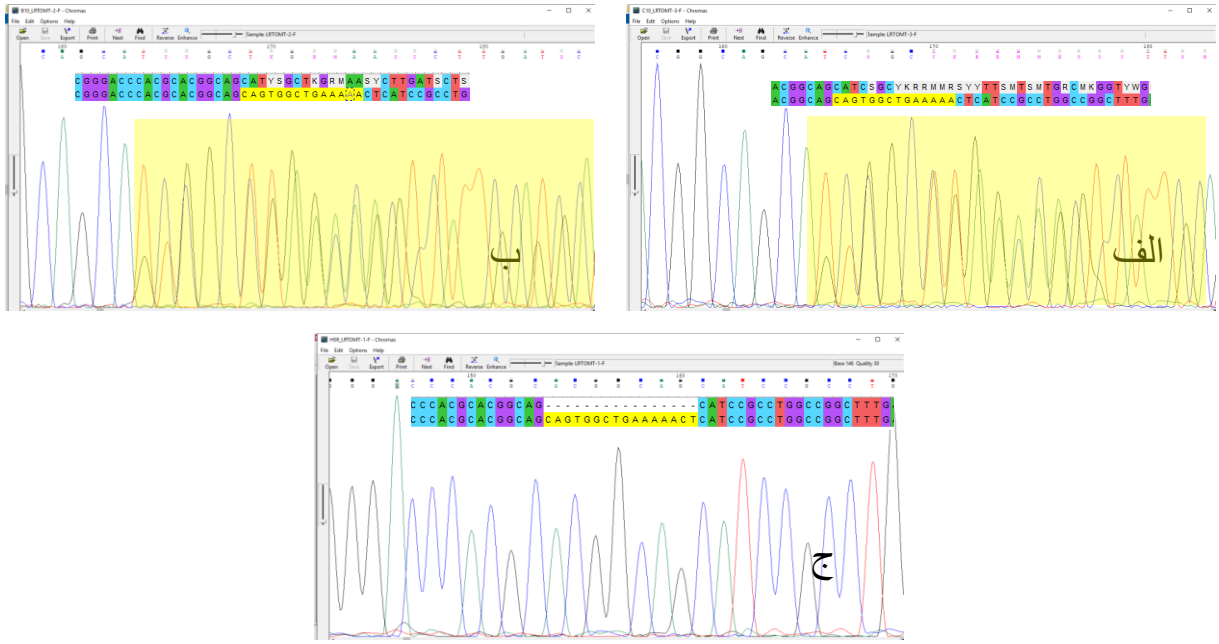
ویژگی‌های کاربردی	ابزار پیشنهادی	VarAFT	snpEff	ANNOVAR	DAmOLD	wANNOVAR	VARIFI
ساخت دیتابیس داخلی	*						
جستجوی اختصاصی بیماری	*					*	
نمایش گرافیکی نتایج	*	*					*
استفاده از فایل BED	*				*		*
تفکیک ماژول‌های مختلف	*		*				
قابلیت اجرای هم‌زمان ماژول‌ها			*				
سیستم عامل	ویندوز	مک، لینوکس، ویندوز	لینوکس، ویندوز	لینوکس، مک	تحت وب	تحت وب	تحت وب
خروجی	VCF	Excel	Text, VCF	Text, VCF	Text, HTML	Text, Excel	VCF
ورودی	VCFs, BED	VCF, Gvcf, ANN	Text, VCF, SAMtools Pileup format	GFF3, VCF, SOAPsnp, MAQ, CASAVA	VCF, BED	VCF, GFF3, ASM.tsv	BAM, FASTQ, BED
آخرین به‌روزرسانی	۲۰۲۰	۲۰۱۹	۲۰۱۸	۲۰۱۹	۲۰۲۰	۲۰۱۹	۲۰۱۹
رابط گرافیکی	*	*			*	*	*
نحوه دسترسی	محدود	آزاد	آزاد	آزاد	آزاد	آزاد	آزاد
نیاز به مهارت کامپیوتری			*	*			
ژنوم مرجع	Hg19	Hg19, Hg38	بیش از ۲۵۰۰ ژنوم را پشتیبانی می‌کند	Hg19, Hg38 Hg18	Hg19, Hg38 Hg18, mouse	Hg19, Hg38	Hg19
تعداد دیتابیس مورد استفاده	۲۴	۲۹	۲۹	۳۸	۳۷	۲۸	۸
حاشیه‌نویسی براساس ژن	*	*	*	*	*	*	*
واریانت های مورد آنالیز	SNPs, INDELs	SNPs, INDELs, CNVs	SNPs, INDELs, MNPs	SNPs, INDELs, Block substitutions	SNPs, INDELs	SNPs, INDELs	SNPs, INDELs

به منظور ارزیابی عملکرد نرم‌افزار آنالیز مرحله اول و دوم NGS توسط شرکت ماکروژن انجام شده و نتایج آن که یک فایل VCF و یک فایل BAM است به عنوان ورودی نرم‌افزار مورد استفاده قرار گرفته است. تعداد واریانت‌های فایل VCF بیمار مورد مطالعه ۶۷۱۸۲۹ واریانت شامل SNPها و Indelها بود. نتایج حاصل از فیلترینگ فایل ورودی در مراحل مختلف در نمودار ۳ نشان داده شده است. با توجه به نمودار بعد از انجام مرحله اولویت‌بندی خودکار واریانت‌ها براساس ژن‌های مرتبط با ناشنوایی غیرسندرمیک اتوزومی مغلوب تعداد واریانت‌ها ۵۰۸ شد. با توجه به شجره خانوادگی بیمار و وجود ازدواج فامیلی، در مرحله بعدی آنالیز، واریانت‌های هوموزیگوت انتخاب شدند و به این ترتیب تعداد واریانت‌ها به ۱۸۷ رسید. بعد از اعمال آستانه فراوانی جمعیتی ۰/۶٪ در پایگاه‌های داده GenomAD و ExAC تعداد واریانت‌های باقی‌مانده به

ترتیب به عدد ۱۱۰ و ۳ کاهش یافت. پس از بررسی جایگاه ژنومی ۳ واریانت نهایی، مشخص شد که ۱ واریانت در ناحیه اگزونی و ۲ واریانت در ناحیه اینترونی واقع شده است. از آنجایی که بر طبق راهنمای ACMG2015 بیش از ۸۵ درصد از واریانت‌های عامل بیماری در نواحی اگزونی واقع شده‌اند؛ حذف ۱۶ نوکلئوتیدی در اگزون شماره ۶ ژن LRTOMT برای مرحله نهایی یعنی آنالیز هم تفکیکی با روش توالی یابی سنگر (نمودار ۴)، به عنوان خروجی آنالیز نرم‌افزار طراحی شده، معرفی شد. جهش ذکر شده از نوع تغییر قالب می‌باشد که باعث تغییر توالی اسید آمینه‌ای و ایجاد کدون توقف زودرس می‌شود. مدت زمان آنالیز این داده توسط نرم‌افزار طراحی شده با استفاده از سخت‌افزار ذکر شده در قسمت روش به طور متوسط در چندین اجرا حدود ۱۵ دقیقه بود.



نمودار ۳: روند فیلترینگ با استفاده از نرم‌افزار



نمودار ۴: الکتروفروگرام مربوط به جهش c. 509_524del در الف) پدر ب) مادر ج) فرزند ناشنوا-

جهش ذکر شده به صورت هوموزیگوت در فرزند ناشنوا و به صورت هتروزیگوت در والدین مشاهده شده است. رنگ زرد جایگاهی را نشان می‌دهد که پس از حذف ۱۶ نوکلئوتید امتداد توالی‌ها در الکتروفروگرام فرد حامل (پدر و مادر) به هم می‌خورد.

بحث و نتیجه‌گیری

امروزه در زمینه تولید نرم‌افزارها و پلت‌فرم‌های تحت وب جهت ارتقاء و بهینه‌سازی حاشیه‌نویسی آنالیز NGS، مطالعات گسترده‌ای در دنیا در حال انجام است. پروژه‌های بزرگی در دهه گذشته در کشورهای مختلف دنیا مانند آمریکا و انگلستان به منظور جمع‌آوری لیست آلل‌های مرتبط با سلامتی و بیماری‌های انسانی آغاز شده است. نتایج چنین پروژه‌هایی به صورت پایگاه‌های داده رایگان مانند Clinvar، dbSNP و غیره در دسترس عموم قرار گرفته است. یکی دیگر از اهداف چنین پروژه‌هایی شناسایی واریانت‌های نادر و با فراوانی آللی پایین در جمعیت است؛ که برای تفسیر بالینی واریانت‌های موجود در ژنوم توالی‌یابی شده هر فردی بسیار مؤثر است [۱۴]. با وجود این که تاکنون پایگاه‌های داده بزرگی از نتیجه توالی‌یابی NGS هزاران فرد سالم و بیمار جمع‌آوری شده است؛ هنوز اطلاعات بشر در مورد تنوع ژنتیکی موجود در ژنوم انسان محدود است. همچنین پایگاه داده‌های ذکر شده برای کاربردهای بالینی ایجاد نشده‌اند و بیش‌تر برای کاربردهای پژوهشی توسعه یافته‌اند. این پایگاه‌های داده دارای خطاهایی مانند گزارش اشتباه یک واریانت به عنوان پاتوژن بدون شواهد کافی هستند. فیلتر کردن یا باقی ماندن یک واریانت به علت حاشیه‌نویسی اشتباه می‌تواند موجب گزارش نتیجه مثبت یا منفی کاذب شود. در پایان روند آنالیز حاشیه‌نویسی چندین واریانت جدید خواهیم داشت که اطلاعات آن‌ها در پایگاه‌های داده وجود ندارد و موجب اختلال در روند آنالیز می‌شود؛ بنابراین نیاز است واریانت‌های باقی‌مانده نهایی به صورت دستی توسط متخصص ژنتیک به طور دقیق بررسی شوند. با ذخیره ساختارمند نتایج چنین بررسی‌های وقت‌گیری به مرور زمان به یک پایگاه داده داخلی مورد اطمینان برای کاربردهای بالینی دست خواهیم یافت که می‌تواند به صورت یک منبع اطلاعاتی ارزشمند در اختیار سایر آزمایشگاه‌های بالینی و محققین این حوزه قرار گیرد. با توجه به این که سایر ابزارهای مشابه توسعه یافته در این حوزه مانند VARAFT،wANNOVAR و ANNOVAR چنین قابلیتی را ندارند، ایجاد چنین قابلیت ارزشمندی تنها از طریق بومی‌سازی نرم‌افزار طراحی شده مطابق مطالعه حاضر میسر شد. علاوه بر این فرکانس آللی یک

واریانت در کل جمعیت، یکی از معیارهای ضروری برای تفسیر بیماری‌زا بودن واریانت‌ها است؛ زیرا فرض بر این است که اگر بیماری ناشی از یک واریانت با فراوانی بالا باشد آن بیماری در جمعیت شایع خواهد بود؛ بنابراین در مورد بیماری‌های نادر و خیلی نادر مانند ناشنوایی ارثی واریانت‌های شایع در جمعیت حذف شد تا به واریانت عامل بیماری برسد. امکان تخمین صحیح فراوانی جمعیتی واریانت‌های مخصوص هر نژادی با استفاده از پایگاه داده محلی نرم‌افزار طراحی شده به وجود خواهد آمد.

نرم‌افزار طراحی شده قادر به انجام هم زمان حاشیه‌نویسی، اولویت‌بندی و فیلتر کردن خودکار واریانت‌ها بدون نیاز به داشتن مهارت‌های کدنویسی است. همچنین دارای ماژول‌های ادغام و یکپارچه کننده فایل‌های VCF و مقایسه فایل‌های VCF به دست آمده از پایپ لاین‌های مختلف است. به منظور سهولت استفاده و گرافیکی بودن، نرم‌افزار مورد نظر تحت سیستم عامل ویندوز طراحی شده است که این امر باعث کاهش سرعت آنالیز دیتاها می‌شود؛ زیرا در محیط ویندوز برخلاف محیط لینوکس از حداکثر قدرت پردازنده مرکزی نمی‌توان استفاده کرد. به دلیل این که ورودی و خروجی نرم‌افزار فایل VCF است، نتایج برای تمامی پلتفرم‌های توالی‌یابی NGS قابل استفاده خواهد بود.

در این مطالعه از بین انواع واریانت‌های ژنتیکی تنها SNPها و INDELها بررسی شده‌اند؛ بنابراین پیشنهاد می‌شود برای شناسایی سایر واریانت‌ها مانند تغییرات ساختاری، بازآرایی‌ها و ترانسلوکاسیون‌های کروموزومی الگوریتم‌های جداگانه‌ای پیاده‌سازی شود. فایل BED مورد استفاده در این مطالعه به دلیل ناکافی بودن اطلاعات جامعی در رابطه با بیماران ناشنوای ایرانی بر اساس مطالعات جمعیت‌های غیرایرانی ایجاد شده است؛ بنابراین بهتر است در آینده با بررسی‌های جامع‌تر در رابطه با بیماران ناشنوای ایرانی فایل BEDای متناسب نقاط جهش خیز جمعیت ایرانی ایجاد شود.

تعارض منافع

این مقاله مستخرج از طرح تحقیقاتی شماره ۳۹۷۵۶۵ با کد اخلاق IR.MUI.RESEARCH.REC.1397.408 می‌باشد که با حمایت مالی معاونت تحقیقات و فناوری دانشگاه علوم پزشکی اصفهان به انجام رسیده است. در این پژوهش هیچ‌گونه تضاد منافعی وجود ندارد.

References

1. Koohiyan M, Azadegan-Dehkordi F, Koohian F, Hashemzadeh-Chaleshtori M. Genetics of Hearing Loss in North Iran Population: An Update of Spectrum and Frequency of GJB2 Mutations. J Audiol Otol 2019;23(4):175-80. doi: 10.7874/jao.2019.00059

2. Falah M, Houshmand M, Balali M, Asghari A, Bagher Z, Alizadeh R, et al. Role of *GJB2* and *GJB6* in Iranian Nonsyndromic Hearing Impairment: From Molecular Analysis to Literature Reviews. Fetal Pediatr Pathol 2020;39(1):1-12. doi: 10.1080/15513815.2019.1627625

3. Korver AM, Smith RJH, Van Camp G, Schleiss MR, Bitner-Glindzicz MAK, Lustig LR, et al. Congenital hearing loss. Nat Rev Dis Primers 2017;3:16094. doi: 10.1038/nrdp.2016.94

4. Ahmed ZM, Riazuddin S, Ahmad J, Bernstein SL, Guo Y, Sabar MF, et al. P PCDH15 is expressed in the neurosensory epithelium of the eye and ear and mutant alleles are responsible for both USH1F and DFNB23. Hum Mol Genet 2003;12(24):3215-23. doi: 10.1093/hmg/ddg358

5. Behlouli A, Bonnet C, Abdi S, Bouaita A, Lelli A, Hardelin JP, et al. EPS8, encoding an actin-binding protein of cochlear hair cell stereocilia, is a new causal gene for autosomal recessive profound deafness. Orphanet J Rare Dis 2014; 9: 55. doi: 10.1186/1750-1172-9-55

6. Ben Said M, Grati M, Ishimoto T, Zou B, Chakchouk I, Ma Q, et al. A mutation in *SLC22A4* encoding an organic cation transporter expressed in the cochlea strial endothelium causes human recessive non-syndromic hearing loss DFNB60. Hum Genet. 2016;135(5):513–24. doi: 10.1007/s00439-016-1657-7

7. Booth KT, Azaiez H, Kahrizi K, Simpson AC, Tollefson WTA, Sloan CM, et al. PDZD7 and hearing loss: More than just a modifier. Am J Med Genet A 2015;167A(12):2957-65. doi: 10.1002/ajmg.a.37274

8. Schaafsma GCP, Vihinen M. VariOtator, a Software Tool for Variation Annotation with the Variation Ontology. Hum Mutat 2016;37(4):344-9. doi: 10.1002/humu.22954

9. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics 2019;35(11):1978-80. doi: 10.1093/bioinformatics/bty897

10. Krunic M, Venhuizen P, Müllauer L, Kaserer B, von Haeseler A. VARIFI-Web-Based Automatic Variant Identification, Filtering and Annotation of Amplicon Sequencing Data. J Pers Med 2019;9(1):10. doi: 10.3390/jpm9010010

11. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet 2012;49(7):433-6. doi: 10.1136/jmedgenet-2012-100918
12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164. doi: 10.1093/nar/gkq603
13. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6(2):80-92. doi: 10.4161/fly.19695
14. Biesecker LG. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. Genet Med 2012;14(4):393-8. doi: 10.1038/gim.2011.78

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2021; 7(4):435-444

Implementation and Optimization of Annotation and Interpretation Step of Next-Generation Sequencing Data for Non-Syndromic Autosomal Recessive Hearing Loss

Shahhoseini Mahdi [1], Molavi Newsha[2], Tabatabaiefar Mohammad Amin[3], Sehhati Mohammadreza[4]*

• **Received:** 4 Feb 2020 • **Accepted:** 21 Jun 2020

Introduction: The precision and time required for analysis of data in next-generation sequencing (NGS) depends on many factors including the tools utilized for alignment, variant calling, annotation and filtering of variants, personnel expertise in data analysis and interpretation, and computational capacity of the lab and its optimization is a challenging task.
Method: An application software was designed and implemented in C# for optimizing the third step of NGS data analysis. In this study, annotation, filtering, and interpretation of NGS data were specifically optimized for non-syndromic autosomal recessive hearing loss disease.
Results: Whole-exome sequencing data of a patient with a pathogenic mutation confirmed by familial genetic analysis, which contained a total number of 671829 variants after primary analysis, were evaluated by the implemented software. After filtering the variants based on a predefined BED file, 508 variants remained. According to the patient's pedigree, in the next step of analysis, homozygote variants were selected and only 187 variants remained. After applying the population frequency threshold of 0.6% on gnomeAD and ExAC databases, the number of variants reached 110 and 3, respectively. The identified pathogen was approved by the results of Sanger sequencing done for family co-segregation. This analysis took about 15 minutes on a moderate PC.
Conclusion: The designed software is a fully graphical one that has the capability of comparing, viewing, filtering, and merging input files without any coding. Moreover, it can construct a local database from the analyzed files and apply region constraints and user-defined thresholds on various fields of the database.

Keywords: Next-Generation Sequencing, Annotation, Variant Effect, Variant Filtering

• **Citation:** Shahhoseini M, Molavi N, Tabatabaiefar MA, Sehhati MR. Implementation and Optimization of Annotation and Interpretation Step of Next-Generation Sequencing Data for Non-Syndromic Autosomal Recessive Hearing Loss. Journal of Health and Biomedical Informatics 2021; 7(4): 435-44. [In Persian]

1. M.Sc. in Biomedical Engineering- Bioelectric, Bioinformatics Dept., Faculty of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
2. M.Sc. in Human Genetics, Genetics and Molecular Biology Dept., Faculty of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
3. Ph.D. in Medical Genetics, Associate Professor, Genetics and Molecular Biology Dept., Faculty of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
4. Ph.D. in Biomedical Engineering- Bioelectric, Assistant Professor, Bioinformatics Dept., Faculty of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
***Corresponding Author:** Mohammadreza Sehhati
Address: Bioinformatics Dept., Faculty of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Hezar Jarib Street, Isfahan, Iran
• **Tel:** 03137923854 • **Email:** mr.sehhati@gmail.com