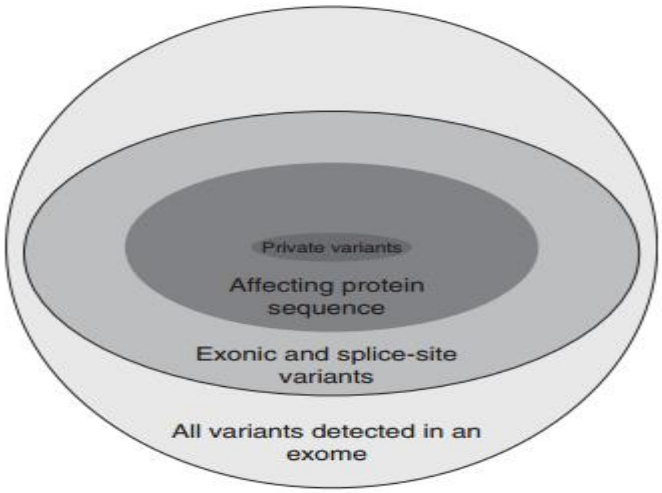
استراتژی های شناسایی ژن های مسبب بیماری در توالی یابی اگزوم (WES)

تعدادی واریانت هایی که در مطالعات توالی یابی exome شناسایی شده اند، بسیار متفاوت است که به مواردی همچون مجموعه غنی سازی Exome استفاده شده ، پلت فرم توالی یابی مورد استفاده و الگوریتم های مورد استفاده برای نقشه برداری و باز خوانی واریانت ها بستگی دارد. به طور معمول بین ۲۰۰۰۰ تا ۵۰۰۰۰ واریانت در هر اگزوم توالی یابی شده شناسایی می شود.

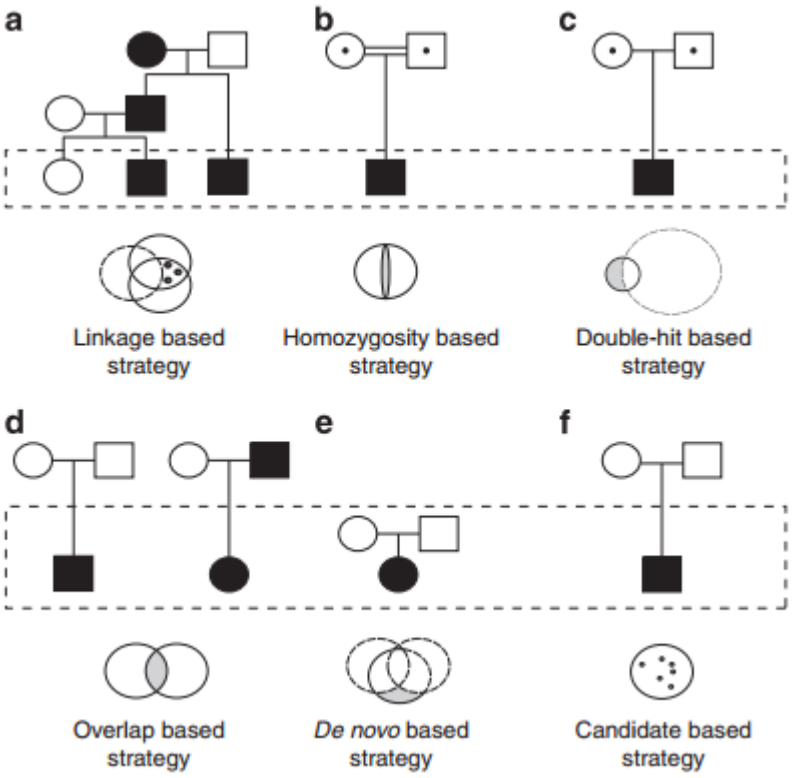
به منظور کاهش تعداد واریانت های مثبت کاذب، واریانت ها ابتدا براساس معیارهای کیفیت فیلتر می شوند، از جمله تعداد کل خوانش های مستقل نشان دهنده یک واریانت (مثلا حداقل پنج خوانش مستقل) و درصد خوانش های نشان دهنده واریانت (به عنوان مثال، واریانت حداقل در ۲۰٪ خوانش ها در مورد واریانت های هتروزیگوت وجود داشته باشد یا در مورد واریانت های هوموزیگوت حداقل ۸۰٪ خوانش ها واریانت را داشته باشند). پس از آن، واریانت های موجود در خارج از نواحی کد کننده، می توانند فیلتر شوند و همچنین واریانت های مترادف (جهش هایکه باعث تغییر کدون یک اسید آمینه به کدون دیگر همان اسید آمینه می شوند) در نواحی کد کننده نیز بر اساس این فرضیه که این واریانت ها تاثیر حداقلی بر روی پروتئین دارند، فیلتر می شوند. موارد ذکر شده باعث می شود که تعداد واریانت های مسبب بیماری بالقوه به حدود ۵،۰۰۰ کاهش یابد. مهمترین کاهش واریانت ها ناشی از حذف واریانت های شناخته شده (معمولا با استفاده از dbSNP، مطالعات منتشر شده یا پایگاه داده های داخلی) است. این مرحله به طور معمول تعداد جهش های کاندیدای بالقوه را در حدود ۹۰-۹۵٪ کاهش می دهد. پس از این، به طور معمول بین ۱۵۰ تا ۵۰۰ واریانت غیر مترادف یا واریانت های جایگاه پیرایش اختصاصی به عنوان واریانت های پاتوژن اولویت بندی می شوند.



اولویت بندی واریانت های NGS. اندازه بیضی های محصور نشان دهنده تعداد نسبی واریانت ها است که پس از هر مرحله اولویت بندی باقی می ماند.

تاکید بر این نکته مهم است داشته باشیم اولویت بندی ممکن است واریانت پاتوژنیک را نیز خارج کند.

واضح است که این اولویت بندی اولیه چیزی بیشتر از یک انتخاب اولیه نیست و به ندرت واریانت پاتوژن را به تنهایی شناسایی می کند و استراتژی های دیگری برای پیدا کردن جهش مسبب بیماری در میان این ۱۵۰-۵۰۰ واریانت اختصاصی مورد نیاز است. ما به طور خلاصه در مورد هر یک از این استراتژی ها صحبت می کنیم:



استراتژی لینکاژ (Linkage): برای یک خانواده با یک اختلال ارثی مونوژنیک، چندین فرد مبتلای خانواده را می توان برای شناسایی واریانت مشترک، توالی یابی کرد. علاوه بر این، اعضای غیر مبتلای خانواده را نیز می توان برای خارج کردن واریانت های خوش خیم اختصاصی توالی یابی کرد. با انتخاب تعداد بیشتر اعضای مبتلای خویشاوند دور از فامیل، میزان واریانت های خوش خیم مشترک به حداقل می رسد. برای مثال، دو sib مبتلا در حدود ۵۰ درصد DNA خود باهم اشتراک دارند. با ترکیب داده های توالی این افراد، کاهش واریانت های اختصاصی که باید مورد توجه قرار گیرد امکان پذیر خواهد بود.

اولین کاربرد این روش توسط Ng و همکاران بود که از دو برادر - خواهر مبتلا استفاده کرد تا واریانت های مشترک را کاهش دهد و توانست میزان ژن های کاندید را به ۹ ژن کاهش دهد. Krawitzet یک نسخه پیچیده تر از این رویکرد را با استفاده از داده های توالی exome استفاده کرد. در این روش haplotype ها برای تمام واریانت های مشترک در ۳ برادر-خواهر مبتلا را تعیین کردند و سپس واریانت ها را بر اساس یکسان بودن از نظر تبار (identical by descent) انتخاب کردند. به این ترتیب آنها تعداد ژنهای کاندید را از ۱۴ به ۲ ژن کاهش دادند.

استراتژی هوموزیگوسیتی(Homozygosity strategy): در مورد یک اختلال نادر ارثی که مظنون به همخونی است، پیش فرض های اولیه این است که این بیماری توسط یک واریانت هوموزیگوت به ارث رسیده از هر دو والد ایجاد می شود و این واریانت درون یک منطقه هوموزیگوتی بزرگ قرار دارد. بنابراین، واریانت های هوموزیگوت را می توان از منظر حضور آنها مناطق هوموزیگوت ژنوم بیمار اولویت بندی کرد. این مناطق می توانند توسط میکروارا های SNP شناسایی شوند و در طی روند اولویت بندی مورد استفاده قرار گیرند، اما Beckeret اخیرا نشان داد که داده های exome خود ممکن است شامل تعداد کافی SNP های آموزنده برای ایجاد نقشه های هوموزیگوسیتی قابل اعتماد باشد. تفاوت اصلی این استراتژی در مقایسه با استراتژی لینکاژ این است که در این استراتژی، واریانت ها هر چند هوموزیگوت هستند ولی تنها زمانی انتخاب می شوند که در درون یک ناحیه بزرگ هموزیگوتی قرار داشته باشند. به این ترتیب، این روش مقدار واریانت های مورد نظر برای پیگیری جهت شناسایی ژن بیماری را به میزان کافی در موارد فردی کاهش می دهد و دیگر نیازی به سکانس اعضای خانواده اضافی، حداقل برای شناسایی واریانت اولیه نیست. با این

konkooran.ir

وجود، این روش تنها می تواند لوکوس های هوموزیگوت را در مناطق با چگالی هدف کافی حاوی SNP های informative را تشخیص دهد.

استراتژی دو ضربه (Double-hit strategy): هنگامی که تنها یک بیمار بدون هیچ یک از اعضای خانواده در دسترس است و اختلال مشکوک به توارث مغلوب است (اما بدون هیچ گونه نشانه ای برای همخونی)، ممکن است توالی یابی اگزوم تنها همین فرد صورت گیرد و ژن هایی که واریانت های هوموزیگوت و هتروزیگوت مرکب را حمل می کنند انتخاب می شوند، زیرا به طور متوسط تعداد نسبتا کمی از این واریانت های غیر هم معنی اختصاصی به صورت ازدواج غیر خویشاوندی به این صورت حضور دارند.

استراتژی همپوشانی (Overlap strategy): در نبود هتروژنیتی ژنتیکی، می توان جهش تک ژنی را در چندین بیمار غیر خویشاوند که فنوتیپ مشابهی دارند جستجو کرد. این استراتژی بویژه برای شناسایی ژن مسبب بیماری در اختلالات غالب بسیار کاربردی است، زیرا ژن های بسیار بیشتری با واریانت های غیر مترادف هتروزیگوت نسبت به ژن های با واریانت های غیر مترادف هوموزیگوت یا هتروزیگوت مرکب وجود دارند. هر چه تعداد بیمارن مورد بررسی افزایش یابد با ترکیب داده ها تعداد زیادی از ژن ها خارج شده ژن های کاندیدای کمتری برای پیگیری وجود خواهد داشت. اولین مطالعه برای پیدا کردن ژن جهش یافته در یک اختلال غالب با استفاده از رویکرد همپوشانی بر روی داده های توالی یابی exome چهار فرد غیرخویشاوند انجام شد.

استراتژی De Novo: استراتژی همپوشانی ذکر شده در بالا فقط برای بیماری های نادر که عمدتا مونوژنیک هستند، کاربرد دارد. در مورد اختلالات شایع که از نظر ژنتیکی بسیار هتروژن هستند، اهداف جهش یافته (یعنی مقدار ژنوم اشغال شده توسط ژن هایی که اگر جهش یافته باشند منجر به بیماری می شوند)، نسبت به شانس منطقی برای یافتن دو بیمار مبتلا دارای جهش در ژن یکسان بسیار بیشتر است. با این حال داشتن هدف جهشی بزرگ شانس رخداد جهش denovo در طی میوز را در یکی از این ژن ها و ایجاد بیماری افزایش می دهد. به خصوص هنگامی که یک اختلال بیشتر به صورت اسپورادیک اتفاق می افتد و با کاهش باروری مرتبط است، به عنوان مثال در مورد معلولیت های ذهنی، علت اصلی می تواند رخداد جهش های de novo قلمداد شود. جهش های De Novo در این بیماران را می توان با استفاده از رویکرد توالی یابی اگزوم مبتنی بر خانواده شناسایی کرد. با توالی یابی اگزوم بیمار و همچنین والدین وی (Trio-analyse)، کاندید های جهش جدید را می توان با فیلتر کردن تمام واریانت های ارثی انتخاب کرد. این استراتژی تعداد محدودی واریانت پاتوژن های بالقوه را ارائه می دهد، به طور متوسط فقط ۰-۳ جهش جدید. یافته های قابل توجهی در مورد ناتوانی های ذهنی، انواع گوناگون اوتیسم و اسکیزوفرنی بدست آمده است که نشان می دهد جهش های de novo ممکن است بخش قابل توجهی از اشکال اسپورادیک اختلالات تکوین عصبی شایع را شامل شود. باید توجه داشت همانند استراتژی های دیگر، در این استراتژی نیز مواردی که به این ترتیب اولویت بندی می شوند، به احتمال زیاد مسبب ایجاد بیماری هستند، اما وقوع جهش de Novo به خودی خود مدرک کافی برای بیماری زایی نیست و نیازمند بررسی follow-up برای بررسی عود (یعنی تایید وقوع و بیماری زایی در افراد دیگر) یا تائید عملکردی (functional proof) بیماری زایی است. علاوه بر این، تمرکز بر روی واریانت هایی که که تنها در کودک مبتلا وجود دارد و در والدین وجود ندارد، نه تنها می تواند در اثر جهش های جدید باشد، بلکه در طی توالی یابی و نقشه برداری به صورت آرتیفکت بوجود آید، به همین دلیل این استراتژی مستلزم تعیین دقیق توالی و پوشش برابر در هر سه نمونه مورد بررسی می باشد. بنابراین غنی سازی و تجمیع نمونه های سه گانه (Trio) در یک آزمایش

Konkooran.ir

معقول به نظر می رسد. در نهایت، استراتژی de novo نیازمند سه آزمایش در هر بیمار دارد و بنابراین تنها زمانی انجام می شود که هیچ یک از استراتژی های دیگر احتمال موفقیت نداشته باشند و نمونه های والدین در دسترس باشند.

استراتژی کاندید (Candidate strategy): در مواردی که یک فرد مبتلا، بدون در دسترس بودن سایر اعضای خانواده یا دیگر افراد مبتلا داشته باشیم، گزینه ها محدود می شوند. در این حالت اولویت بندی ممکن است بر اساس تاثیر پیش بینی شده واریانت بر روی عملکرد و ساختار پروتئین باشد، که یکی از اولویت ها جهش های ایجاد کننده کدون خاتمه زودرس، جهش های تغییر چهارچوب قرائت ترجمه و جهش در سایت های اسپلایسینگ است. در مورد واریانت های missense، دو معیار اصلی جهت اولویت بندی در نظر گرفته می شود، یکی تاثیر بر ساختار پروتئین (برای مثال، نمره Grantham) و دیگری حفاظت تکاملی نوکلئوتید واریانت (به عنوان مثال نمرات phyloP یا GERP). جهش های پاتوژنیک به طور قابل توجهی بیشتر از واریانت های خوش خیم محافظت شده هستند. از طرفی اطلاعات مربوط به عملکرد ژن و ارتباط آن با فنوتیپ و آنچه که در مورد پاتوفیزیولوژی آن شناخته شده یا قابل پیش بینی است، می تواند در اولویت بندی بیشتر و بهتر مورد استفاده قرار گیرد. روش های انتخاب نوع کاندیداها در این مقیاس نسبتا جدید هستند. هنگامی که هیچ واریانت کاندیدی (یعنی جهش های ایجاد کننده پروتئین بن بریده) وجود نداشته باشد، اغلب مطالعات به سمت پیش بینی کننده های کامپیوتری و محاسباتی گرایش پیدا می کنند تا تاثیر واریانت های missense بر ساختار یا عملکرد پروتئین را تعیین کنند.

Konkooran.ir

Table 1 Overview of applicability, assumptions, advantages, and disadvantages for exome sequencing strategies

Strategy	*Applicability*	*Assumptions*	*Advantages*	*Disadvantages*
Linkage (Figure 3a)	Multiple affected within a single family.	Fully penetrant mutation segregating with the disorder.	Additional individuals can be sequenced to limit the search space further.	Might require large sequencing efforts, affected share a lot of private variants.
Homozygosity (Figure 3b)	Single affected from consanguine parents.	Homozygous mutation within a homozygous stretch.	Only a single patient is required. Only a single experiment is required.	The disorder might not necessarily be caused by a mutation in a homozygous region. Homozygosity based on exome data might suffer from limited resolution.
Double-hit (Figure 3c)	Single affected with a recessive disorder.	A single rare homozygous or two rare compound heterozygous mutations.	Only a single patient is required. Only a single experiment is required.	Relies on the availability of (ethnically matched) control data. Depends on expected
Overlap (Figure 3d)	Multiple affected with a dominant disorder.	The disorder is completely (or mostly) monogenic and all patients suffer from the same disorder.	Only (three) single patients are required. Only a single experiment is required.	Relies heavily on accurate phenotyping and the assumption of a single involved locus.
De novo (Figure 3e)	Single sporadic affected.	The mutation occurs *de novo* in the patient.	Only a single patient is required. Does not necessarily depend on control data or other prioritization assumptions.	Relies on accuracy of sequencing technology.
Candidate (Figure 3f)	Single affected with a dominant disorder without additional family members.	The causative gene or mutation shares features with known genes/mutations.	Only a single patient is required. Only a single experiment is required. Does not necessarily depend on control data or other prioritization assumptions.	Biased approach, relying on current biological knowledge.