

Обобщение внутрилабораторных критериев качества геномных вариантов, валидируемых по результатам полногеномного секвенирования

Антоненко А.^{1*}, Белов Р.¹, Ревкова М.¹, Доморацкая Е.¹, Барциц А.¹,
Панферова А.¹, Губона М.¹, Уланова П.¹, Дибирова Х.¹, Соколова Н.¹,
Дорошук Н.¹, Макарова М.¹, Грознова О.², Криницына А.¹, Беленикин М.^{1**}

¹ ООО «Эвоген», Москва, Россия

² Научно-исследовательский клинический институт педиатрии им. акад. Ю.Е. Вельтищева
ФГБОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия

* antonenko_a.n@mail.ru; ** evogenlab@yandex.ru

Ключевые слова: полногеномное секвенирование; WGS; валидация вариантов NGS; секвенирование по Сэнгеру

Мотивация и цель: Стремительное развитие технологий высокопроизводительного секвенирования (NGS), алгоритмов биоинформатики и обработки данных NGS сопряжено с необходимостью валидации выявленных потенциальных причинных (каузативных) вариантов альтернативными методами. Наиболее часто валидация найденных вариантов проводится с помощью капиллярного секвенирования (КС) методом Сэнгера, так как данный метод считается так называемым золотым стандартом при проведении клинических исследований. Количество научных статей по сравнению результатов NGS для экзомов и генных панелей в сравнении с КС непрерывно растет. И все больше накапливается доказательств об избыточности валидации во многих случаях. Так, валидация КС «высококачественных вариантов» NGS (глубина прочтения не менее 20х, представленность варианта не менее 20 %, высокая оценка качества вызова варианта) уже не считается необходимой [1]. А итоговое увеличение стоимости исследования, трудозатрат и сроков выдачи заключений при валидации всех результатов NGS с помощью КС приводит скорее к снижению эффективности тандема NGS+КС. При экзомных исследованиях валидацию потенциально каузативных вариантов часто рекомендуется проводить выборочно, например, в случае небольших найденных инсерций/делеций, а также однонуклеотидных вариантов, которые по результатам биоинформатического анализа имеют качество варианта ниже порогового [2]. Однако численные значения таких пороговых значений в литературных данных приводятся редко, поскольку предполагается индивидуальный их расчет в каждой лаборатории (например, [3]: 5х, представленность варианта от 35 %, значения QUAL от 139). В отличие от результатов экзомных и таргетных NGS исследований, информация о сравнении результатов полногеномного секвенирования (WGS) и валидации КС в литературе представлена очень ограниченно. В отличие от подходов таргетного секвенирования, методология проведения WGS с использованием безамплификационного (PCR-free) подхода предполагает максимальную однородность результатов секвенирования и позволяет получить данные, которые априори имеют наибольшую равномерность охвата по всей длине генома, наименьшее число экспериментальных артефактов и меньшие

требования по необходимой глубине прочтений [4]. Постоянное увеличение объема и качества WGS данных ставит вопрос о необходимости обновления рекомендаций по применению КС в качестве валидационного инструмента. Так, в рекомендациях [5] 2022 г. акцент делается на необходимости позитивной переоценки использования WGS при диагностике и обновлении ранее опубликованных руководств по проведению NGS с указанием, что варианты, соответствующие определенным внутри конкретной лаборатории критериям и показателям качества обработки данных, могут не требовать дополнительной валидации. Целью нашего исследования явилось обобщение внутрилабораторных критериев качества, по достижении которых валидация варианта с помощью КС может считаться избыточной или подходить близко к этому порогу.

Методы и алгоритмы: Всю экспериментальную работу и биоинформатический анализ данных WGS и КС проводили на базе собственной лаборатории. Пробо-подготовку образцов для WGS на платформах DNBseq-T7/G400 (MGI) в режиме PE150 проводили по безамплификационному PCR-free протоколу с предварительной ферментативной фрагментацией. Валидацию всех вариантов с помощью КС выполняли на секвенаторе ABI3500 по стандартным протоколам производителя.

Результаты: На данный момент лаборатория провела наибольшее в РФ число WGS, однако для анализа данных мы использовали 20K+ образцов, из которых ~4500 составляли образцы с выявленными орфанными/редкими вариантами и наследственными опухолевыми синдромами, валидированными по методу Сэнгера. Из валидированных КС образцов: 97.9 % были подтверждены рутинным образом (результаты КС полностью совпали с результатами NGS), 1.1 % требовали дополнительного уточнения и экспертного анализа, в результате которого при валидации было получено подтверждение (достигнуто согласование данных NGS и результатов КС), 1 % составили технические отказы в сложных областях. Полученные результаты позволили сформулировать следующие критерии для возможности непроведения валидации вариантов с помощью КС: глубина прочтения варианта 10x и выше; представленность аллельного варианта 15–20 % и выше; вариант располагается в области, не имеющей псевдогенов; в ближайшем окружении к валидируемому варианту отсутствуют множественные и сложные для амплификации и аннотации инсерции/делеции, протяженные тандемные и однонуклеотидные повторы.

Выводы: 1. Пороговое значение качества вариантов зависит от индивидуальных деталей алгоритма и биоинформатического пайплайна. 2. Валидация методом Сэнгера оказалась полезной для ~1–2 % наиболее сложных вариантов (например, инделов до 30 п. н., если дополнительные варианты расположены в непосредственной близости от валидированного варианта, областей с повторами, GC-регионов, псевдогенов, вариантов с покрытием <10x), однако большинство из этих вариантов все еще можно вызвать *in silico* с помощью дополнительной биоинформатической обработки данных NGS. 3. Однако КС по-прежнему сохраняет актуальность для валидации самих клинических кейсов, т. е. подтверждения действительной каузативности найденных в результате биоинформационной обработки вариантов и проведения сегрегационного анализа. Это приводит к тому, что наиболее эффективным способом более информативного и обширного сегрегационного анализа видится проведение WGS с пониженной глубиной прочтения (~10x) для родственников в дополнение к стандартному 30x секвенированию пробанда.

Summary of intralaboratory quality criteria for genomic variants validated by whole-genome sequencing results

Antonenko A.^{1*}, Belov R.¹, Revkova M.¹, Domoratskaya E.¹, Bartcic A.¹, Panferova A.¹, Gubona M.¹, Ulanova P.¹, Dibirova H.¹, Sokolova N.¹, Doroshuk N.¹, Makarova M.¹, Groznova O.², Krinitsina A.¹, Belenikin M.^{1**}

¹ Evogen LLC, Moscow, Russia

² Russia Veltishev Research and Clinical Institute for Pediatrics and Pediatric Surgery of the Pirogov Russian National Research Medical University, Moscow, Russia

* antonenko_a.n@mail.ru; ** evogenlab@yandex.ru

Key words: whole-genome sequencing, WGS; validation of NGS variants; Sanger sequencing

Motivation and Aim: The rapid development of high-throughput sequencing (NGS) technologies, bioinformatics algorithms, and NGS data processing is associated with the need to validate identified potential causative variants using alternative methods. The most common way to validate the identified variants is by capillary sequencing (CS) using the Sanger method, as this method is considered the so-called ‘gold standard’ for clinical research. The number of scientific articles comparing NGS results for exomes and gene panels versus CS is continuously increasing. And evidence is accumulating about the redundancy of validation in many cases. Thus, CS validation of ‘high-quality variants’ of NGS (read depth of at least 20x, variant representation of at least 20 %, high quality score of variant calling) is no longer considered necessary [1]. And the resulting increase in the cost of the study, labour costs and time of issuing conclusions when validating all NGS results with the help of CS leads rather to a decrease in the efficiency of the NGS+CS tandem. In exome studies, validation of potentially causative variants is often recommended to be performed selectively, for example, in the case of small insertions/deletions found, as well as single-nucleotide variants that, according to the results of bioinformatic analysis, have a variant quality below the threshold [2]. However, the numerical values of such thresholds are rarely reported in the literature, as they are supposed to be calculated individually in each laboratory (e. g. [3]: 5x, variant representation from 35 %, QUAL values from 139). Unlike the results of exome and targeted NGS studies, there is very limited information in the literature on the comparison of the results of whole genome sequencing (WGS) and CS validation. In contrast to target sequencing approaches, the methodology of WGS using a non-amplification (PCR-free) approach implies maximum homogeneity of sequencing results and allows obtaining data that a priori have the greatest uniformity of coverage over the entire genome length, the lowest number of experimental artefacts and lower requirements for the necessary depth of reads [4]. The continuous increase in the volume and quality of WGS data raises the question of the need to update the recommendations for the use of CS as a validation tool. For example, the 2022 guidelines [5] emphasise the need to positively reassess the use of WGS in diagnostics and update previously published NGS guidelines to state that variants that meet laboratory-specific criteria and quality indicators for data processing may not require additional validation. The aim of this research was to summarise the intra-laboratory quality criteria at which validation by CS of a WGS-derived variant may be considered redundant or close to this threshold.

Methods and Algorithms: All experimental work and bioinformatics analyses of WGS and CS data were performed in our own laboratory. Sample preparation for WGS on DNBseq-T7/G400 (MGI) platforms in PE150 mode was performed using a non-amplification PCR-free protocol with preliminary enzymatic fragmentation. Validation of all variants by CS was performed on an ABI3500 sequencer according to the producer's standard protocols.

Results: To date, the laboratory has performed the largest number of full genome sequencing in the Russian Federation, but we used 20K+ samples for data analysis, of which ~4500 were samples with identified orphan/rare variants and inherited tumour syndromes validated using the Sanger method. Of the CS validated samples: 97.9 % were validated routinely (CS results were in complete agreement with NGS results), 1.1 % required additional clarification and expert analysis resulting in validation (agreement between NGS data and CS results was achieved), 1 % were technical failures in complex areas. The results obtained allowed us to formulate the following criteria for the possibility of not validating variants using CS: variant read depth – 10x and higher; allelic variant representation 15–20 % and higher; the variant is located in a region without pseudogenes; there are no multiple and difficult for amplification and annotation insertions/deletions, extended tandem and single nucleotide repeats in the immediate vicinity of the variant to be validated.

Conclusion: (1) The threshold for variant quality depends on the individual details of the algorithm and bioinformatic pipeline. (2) Sanger validation has proven useful for ~1–2 % of the most difficult variants (e. g. indels up to 30 bp if additional variants are located in close proximity to the validated variant, regions with repeats, GC regions, pseudogenes, variants with <10x coverage), but most of these variants can still be induced in silico with additional advanced bioinformatic NGS data processing. (3) However, CS still remains relevant for validating the clinical cases themselves, i. e. confirming the actual causality of the variants found by bioinformatics processing and performing segregation analysis. This leads us to believe that the most effective way to perform a more informative and extensive segregation analysis is to perform WGS with reduced read depth (~10x) for relatives in addition to the standard 30x sequencing of the proband.

Список литературы/References

1. Arteche-Lopez A. et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. *Sci Rep.* 2021;11:5697. doi 10.1038/s41598-021-85182-w
2. Strom S.P. et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genet Med.* 2014;16(7):510-515. doi 10.1038/gim.2013.183
3. Bauer P. et al. Development of an evidence-based algorithm that optimizes sensitivity and specificity in ES-based diagnostics of a clinically heterogeneous patient population. *Genet Med.* 2019;21(1):53-61. doi 10.1038/s41436-018-0016-6
4. Meynert A.M. et al. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics.* 2014;15:247. doi 10.1186/1471-2105-15-247
5. Souche E. et al. Recommendations for whole genome sequencing in diagnostics for rare diseases. *Eur J Hum Genet.* 2022;30:1017-1021. doi 10.1038/s41431-022-01113-x