ВЛИЯНИЕ ВЫБОРА ЧИСЛА ПОКРЫТИЙ ПРИ СЕКВЕНИРОВАНИИ НА ТОЧНОСТЬ ОПРЕДЕЛЕНИЯ ЕДИНИЧНЫХ НУКЛЕОТИДНЫХ ВАРИАНТОВ

Д. И. Борисевич, А. Ю. Красненко, И. Ф. Стеценко, Д. А. Плахина, В. В. Ильинский ✉

ООО «Генотек», Москва

В настоящее время технология секвенирования нового поколения (NGS) широко применяется в клинической практике. Однако до сих пор стоимость одного исследования с использованием технологии NGS остается достаточно высокой, что ограничивает широкое применение данного метода. Одним из факторов, влияющих на стоимость, является выбор числа покрытий при секвенировании, то есть количество раз, которое был отсеквенирован каждый нуклеотид. С одной стороны, уменьшение числа покрытий значительно снижает стоимость и время, затрачиваемое на исследования, с другой стороны, при уменьшении данного показателя снижается качество получаемых результатов. До сих пор не существует однозначного мнения, какое минимальное число покрытий достаточно для получения достоверного результата. Целью данного исследования было определить минимальное число покрытий, достаточное для корректного определения гетерозигот и единичных нуклеотидных вариантов (SNV). В представленной работе, используя различные биоинформатические методы, было показано, что минимальное число покрытий соответствует 12X.

Ключевые слова: секвенирование нового поколения (NGS), число покрытий, мутация, рид, SNP, SNV

Благодарности: авторы благодарят Анну Давыдову из «Генотек» за помощь в написании статьи.

✉ **Для корреспонденции:** Ильинский Валерий Владимирович
Наставнический пер., д. 17, стр. 1, г. Москва, 105120; valery@genotek.ru

Статья получена: 22.06.2017 **Статья принята к печати:** 27.06.2017

THE IMPACT OF SEQUENCING DEPTH ON ACCURACY OF SINGLE NUCLEOTIDE VARIANT CALLS

Borisevich DI, Krasnenko AYu, Stetsenko IF, Plakhina DA, Ilinsky VV ✉

Genotek, Moscow, Russia

Today, next generation sequencing (NGS) is extensively used in the research setting. However, high costs of NGS testing still prevent its routine use in clinical practice. One of the factors affecting the cost of sequencing is the number of reads per site, i.e. the number of times each nucleotide gets sequenced. On the one hand, lower coverage makes the whole process much faster and less time-consuming. On the other hand, it results in poor data quality. No unanimous opinion has been reached yet as to what minimum depth of coverage can produce reliable results. The aim of this study was to determine the minimum number of reads sufficient for accurate base calling of heterozygous and single nucleotide variants (SNV). Using bioinformatics methods, we demonstrate that accuracy can be achieved at a minimum depth of 12X.

Keywords: Next-generation sequencing (NGS), sequencing depth, mutation, read, SNP, SNV

Acknowledgements: authors thank Anna Davydova of Genotek for her helpful comments.

✉ **Correspondence should be addressed:** Valery Ilinsky
per. Nastavnichesky, d. 17, str. 1, Moscow, Russia, 105120; valery@genotek.ru

Recieved: 22.06.2017 **Accepted:** 27.06.2017

Участки генома, кодирующие белки, составляют лишь 1 % всего человеческого генома, но именно в них сосредоточены 85 % мутаций, определяющих возникновнеие и развитие различных заболеваний [1]. В связи с этим экзомное секвенирование, а также секвенирование с использованием специально разработанных панелей для обогащения, фокусирующихся на тех участках экзонов, в которых могут быть обнаружены значимые мутации, нашли наибольшее применение в клинической практике [2].

Один из важнейших вопросов клинического использования экзомного секвенирования — выбор адекватного числа покрытий (количество раз, которое был отсеквенирован каждый нуклеотид; обычно обозначается как 10x, 20x, 50x и т. д.) [3]. Именно оно позволяет выявить возможные ошибки считывания нуклеотидов на машине и определить истинные позиции в геноме. При этом мы сталкиваемся с двумя разнонаправленными факторами, влияющими на определение количества покрытий. Первый фактор — это время и стоимость секвенирования, которые увеличиваются с ростом числа покрытий. Второй фактор — статистический: какое минимальное количество покрытий позволяет свести до уровня допустимой погрешности ошибку при выявлении мутаций, причем по этому фактору не существует единого мнения. С чем это связано?

Используя технологию коротких прочтений Illumina, Bentley и соавт. в 2008 г. определили, что почти все единичные нуклеотидные варианты (SNV) в гомозиготе обнаруживаются при покрытии 15x, тогда как для обнаружения

SNV в гетерозиготе необходимо покрытие 33х [4]. Основываясь на полученных данных, в большинстве последующих работ, связанных с обнаружением SNV, авторы использовали значение покрытия 33х в качестве стандартного [5, 6]. В 2011 г. Ajay и соавт. опубликовали статью, в которой показали, что для определения 95 % SNV, а также коротких инсерций и делеций необходимо устанавливать покрытие 50х. Однако последующие эксперименты с использованием усовершенствованных реагентов и программного обеспечения для обработки данных позволили авторам получить такой же результат, снизив значение покрытия до 35х [7]. В 2014 г. вышла статья Fang и соавт., в которой было найдено, что для обнаружения 95 % вставок и делеций необходимо устанавливать покрытие 60х [8].

Разброс в числе покрытий, представленный выше, показывает, что говорить об универсальности этого значения в настоящее время становится все сложнее, поскольку количество прочтений одного и ого же участка для обнаружения вариантов напрямую зависит от качества прочтения этого участка. На качество прочтения влияет не только технология секвенирования, используемые реагенты, но также и подготовка образца. Например, трудности амплификации GC-богатых участков при проведении полимеразной цепной реакции (ПЦР) приводят к ухудшению качества прочтения и, как следствие, к необходимости увеличения числа покрытий. В настоящее время существует химия для ПЦР, позволяющая улучшить качество реакции и тем самым дальнейшее качество прочтения при секвенировании. В 2013 г. Meynert и соавт. определили, что для обнаружения 95 % SNP требуется покрытие от 20х до 46х в зависимости от используемой химии [9]. Этими же авторами в 2014 г. было обнаружено, что для детекции 95 % SNP достаточной глубиной покрытия является 14 ридов [10]. Кроме того, Li и соавт. показали, что глубина покрытия также зависит и от количества индивидуальных образцов, используемых при секвенировании [11]. Так, при обнаружении мутации, встречающейся с частотой менее 0,2 %, секвенирование 3 000 индивидуальных образцов с покрытием 4х дает аналогичный результат, что и секвенирование менее 2 000 индивидуальных образцов при покрытии 30х. Таким образом, мы видим, что факторов больше, чем кажется на первый взгляд, и количество покрытий может быть эффективно уменьшено за счет ряда деталей при проведении исследования и исходя из его конкретных целей.

В представленной работе показано, что при использовании панели для обогащения Genotek01 минимальное покрытие, достаточное для корректного определения гетерозигот и SNV, соответствует 12х, при этом расхождение результатов секвенирования и результатов валидирования по методу Сэнгера составляет 0,5 %.

МАТЕРИАЛЫ И МЕТОДЫ

Выделение ДНК, приготовление и секвенирование ДНК-библиотек

ДНК выделялась из цельной венозной крови, полученной от пациентов, страдающих наследственными заболеваниями, с помощью набора QIAmp DNA Mini Kit (Qiagen, Германия). Контроль качества геномной ДНК проводили с помощью агарозного гель-электрофореза. Важным показателем было отсутствие деградации ДНК и загрязнения РНК. Измерение концентрации выделенной ДНК проводили на приборе Qubit 3.0 Fluorometer (Life Technologies, США). ДНК-библиотеки готовили с помощью набора NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, США) с использованием адаптеров для секвенирования на платформе Illumina, согласно протоколу производителя. Двойное баркодирование выполнялось с помощью ПЦР с набором NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) той же фирмы. Контроль качества полученных библиотек фрагментов ДНК был проведен с помощью Agilent 2100 Bioanalyzer (Agilent Technologies, США). Для таргетного обогащения кодирующих регионов генома использовался набор MYbaits (MYcroarray, США). Секвенирование проводилось на геномном анализаторе HiSeq 2500 System (Illumina, США) методом парных прочтений длиной 100 нуклеотидов с использованием наборов HiSeq Rapid PE Cluster Kit v2 и HiSeq Rapid SBS Kit v2 (Illumina) по протоколу производителя.

Секвенирование по Сэнгеру

Для проведения процедуры мечения ампликонов флуоресцентными метками использовали наборы BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, США) по протоколу производителя. Секвенирование по Сэнгеру проводилось на генетическом анализаторе ABI PRISM 3500 Genetic Analyzer (Applied Biosystems, США) по протоколу производителя.

Биоинформатический анализ

Полученные чтения выравнивали на референсный геном *hg19* с помощью программы BWA. Дедупликацию ридов выполняли командой rmdup в SAMtools, коллинг (variant calling) осуществляли с помощью пакета GATK (Genome Analysis ToolKit). Было обнаружено 89 мутаций: 10 гомо- и гемизигот, 79 гетерозигот; 80 точечных мутаций (SNP) и 9 коротких инсерций и делеций (инделов). Были также определены генотипы в регионах в 200 нуклеотидов влево и вправо от мутации. Все позиции из регионов, проанализированных в ходе анализа секвенирования, были валидированы с использованием секвенирования по Сэнгеру, которое считается золотым стандартом для детекции коротких мутаций. Хроматограммы были обработаны единообразно. Коллинг мутаций из хроматограмм был осуществлен с помощью собственного ПО, разработанного «Генотек» на основе BioPython и пакетов R: sangerseqR, seqinr, Biostrings и Rsubread. Нами было проведено сравнение генотипов, полученных методом высокопроизводительного секвенирования (NGS), с результатами секвенирования по Сэнгеру и рассчитаны чувствительность и специфичность.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Валидация мутаций секвенированием по Сэнгеру

Секвенированием по Сэнгеру не были подтверждены 15 из 89 обнаруженных с помощью коллинга мутаций, т. е. они или имели другой генотип, чем был определен секвенированием, или отсутствовали. При этом 8 из 15 неподтвержденных мутаций имели гетерозиготный генотип, в то время как валидация по Сэнгеру определила наличие гомозиготной мутации. Стоит отметить, что в данном случае гетерозигота по данным NGS была поддержана только одним ридом с референсным аллелем (на рис. 1 кластер мутаций в правом нижнем углу графика).

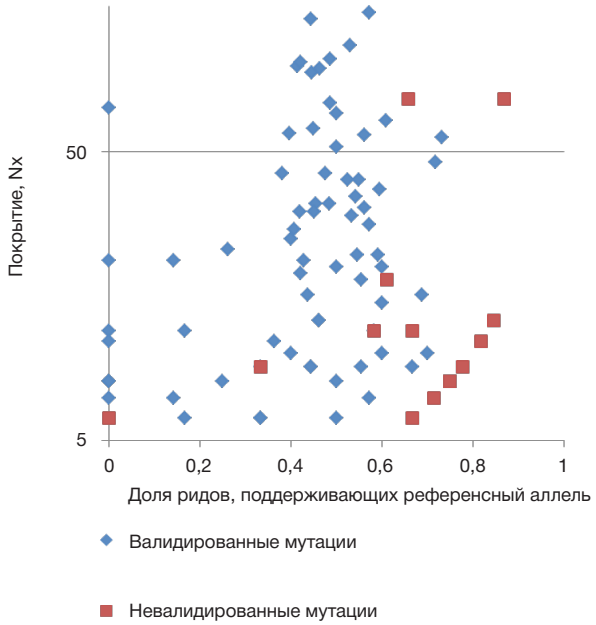

Рис. 1. Распределение валидированных и невалидированных Сэнгером мутаций по покрытию и доле ридов, поддерживающей альтернативный аллель. Одной точке может соответствовать более одной мутации

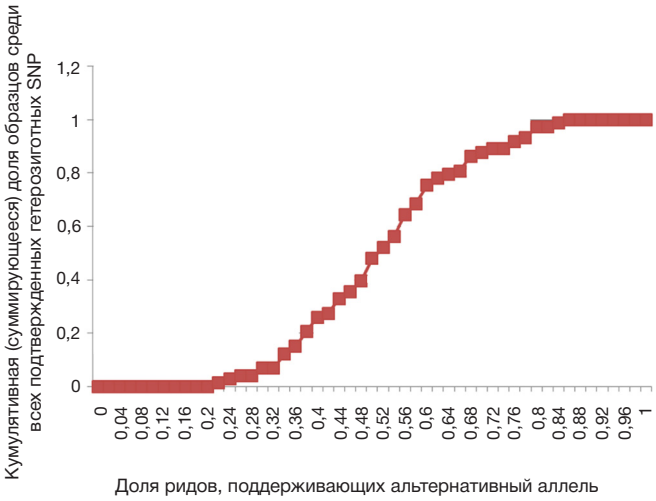

Рис. 2. Кумулятивное распределение доли образцов, имеющих долю ридов, поддерживающих альтернативный аллель X или менее

Симулирование различного покрытия

Для определения минимального порога секвенирования многократно симулировали понижение покрытия (bootstrap) для каждой из доступных мутаций и окружающего их региона, а также был проведен коллинг мутаций в таких данных. Частоту ошибок коллинга оценивали, используя позиции, которые в изучаемых образцах были подтверждены секвенированием по Сэнгеру как имеющие генотип референсных гомозигот.

Оценка качества секвенирования возможна с помощью метрики Phred quality score (Q score), выдаваемой секвенатором для каждого нуклеотида [12]. Однако данное качество является лишь оценкой точности секвенирования и не дает надежных значений для имеющихся данных. Мы проверяли, совпадает ли каждый имеющийся рид, перекрывающий такую позицию, по генотипу с референсным, и если находили отличие, то считали это ошибкой коллинга.

Были проанализированы 372 443 нуклеотида. Из них 276 были нуклеотидами, отличающимися от референсного, а остальные совпадали с ним. Таким образом, частота ошибок составила 0,0741 %, что эквивалентно примерно Q31 по Phred quality score.

Для 69 позиций, в которых имелись подтвержденные гетерозиготные мутации, мы оценили распределение доли ридов, поддерживающих альтернативный аллель (рис. 2).

Используя полученное распределение ридов из гетерозиготы и частоту ошибок коллинга, была рассчитана частота возникновения комбинаций для различных покрытий от 2x до 50x для различного числа ридов от 0 до максимально возможного, поддерживающих альтернативный аллель. Полученные частоты применили для расчета частоты двух типов ошибок: определение гетерозиготной позиции как

Варьирование числа и доли ридов для отсечки референсных и альтернативных гомозигот

Число ошибочных коллов на 12Мб библиотеке																						
мин.поддержка	покрытие -->	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
≥ 2		6	19	39	65	98	137	183	236	294	360	432	510	595	687	784	889	999	1116	1240	1370
≥ 3		0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	3	3	4	5	6
≥ 4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 5		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 6		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 8		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
≥ 10		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Число обнаруженных коллов 0/1 из 1000																					
Порог гетерозигот 0,9	≥ 2	0	375	625	782	875	930	961	978	978	989	994	996	998	1000	1000	1000	1000	1000	1000	1000
	≥ 3		0	250	469	640	765	851	908	934	962	978	987	993	996	998	999	999	1000	1000	1000
	≥ 4			-1	157	328	492	633	744	817	882	924	952	970	982	989	994	996	998	999	999
	≥ 5				0	93	219	359	498	612	721	803	865	909	941	962	975	985	990	994	996
	≥ 6					0	55	141	252	366	495	610	707	787	849	895	928	952	968	979	987
	≥ 7						0	31	88	161	269	384	498	604	696	773	834	881	916	942	961
	≥ 8							0	18	44	108	191	289	394	500	598	685	760	820	868	905
	≥ 9								0	0	28	70	131	211	304	402	500	593	676	748	808
	≥ 10									0	1	16	44	89	151	227	315	407	500	588	668
Порог гетерозигот 0,8	≥ 2	0	375	625	626	781	875	930	960	934	962	978	986	992	983	989	994	996	998	994	996
	≥ 3		0	250	313	546	710	820	890	890	935	962	977	987	979	987	993	995	998	994	996
	≥ 4			0	1	234	437	602	726	773	855	908	942	964	965	978	988	992	996	993	995
	≥ 5				0	-1	164	328	480	568	694	787	855	903	924	951	969	981	988	988	992
	≥ 6					0	0	110	234	322	468	594	697	781	832	884	922	948	966	973	983
	≥ 7						0	0	70	117	242	368	488	598	679	762	828	877	914	936	957
	≥ 8							0	0	0	81	175	279	388	483	587	679	756	818	862	901
	≥ 9								0	0	1	54	121	205	287	391	494	589	674	742	804
	≥ 10									0	0	0	34	83	134	216	309	403	498	582	664
Порог гетерозигот 0,7	≥ 2	0	375	375	626	781	711	821	890	817	881	924	951	909	941	961	929	951	969	942	960
	≥ 3		0	0	313	546	546	711	820	773	854	908	942	904	937	959	928	950	969	942	960
	≥ 4			0	1	234	273	493	656	656	774	854	907	881	923	950	923	947	967	941	959
	≥ 5				0	0	0	219	410	451	613	733	820	820	882	923	904	936	959	936	956
	≥ 6					0	0	1	164	205	387	540	662	698	790	856	857	903	937	921	947
	≥ 7						0	0	0	0	161	314	453	515	637	734	763	832	885	884	921
	≥ 8							0	0	0	0	121	244	305	441	559	614	711	789	810	865
	≥ 9								0	0	0	0	86	122	245	363	429	544	645	690	768
	≥ 10									0	0	0	0	0	92	188	244	358	469	530	628
Число обнаруженных коллов 1/1 при настоящем генотипе 0/1 на 1000																					
≥ 0,9		9	6	5	4	3	3	3	2	13	11	9	8	7	6	5	5	4	4	9	8
≥ 0,8		9	6	5	90	51	32	23	17	56	42	33	26	21	47	38	32	27	24	42	37
≥ 0,7		9	6	189	90	51	178	114	78	170	125	94	73	130	104	84	134	111	93	136	116

референсной гомозиготы; и определение гетерозиготной позиции как альтернативной гомозиготы при различных порогах отсечки референсных и альтернативных гомозигот.

Для отсечки референсных гомозигот варьировали фиксированное число ридов (от 2 до 10). Для отсечки альтернативных гомозигот варьировали долю ридов, поддерживающих альтернативный аллель, между 70, 80 и 90 % (таблица).

Установлено, что для коротких мутаций (SNP и инделов) точность метода достигает 99,7 %, чувствительность — 98 % при покрытии в точке 12х. Меньшее покрытие приводит к значительному снижению чувствительности в сигмоидном характере и поэтому не может быть рекомендовано.

Для планирования эксперимента в лаборатории важно знать, с каким средним покрытием нужно секвенировать образец, чтобы таргетные регионы были покрыты на 12х. Для расчета этой величины мы построили корреляцию между средним покрытием образца и долей таргетного региона, покрытого 12 ридами (рис. 3).

Было обнаружено, что для покрытия таргетных регионов не менее чем на 12х на 90 % и более необходимо минимум 40х среднее покрытие дедуплицированными ридами.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Секвенированием по Сэнгеру не удалось подтвердить 15 из 89 мутаций, при этом 8 из 15 неподтвержденных мутаций имели гомозиготный, а не гетерозиготный генотип. Наличие данных мутаций определяется моделью ошибок GATK, программой, позволяющей получить набор вариантов в исследуемом геноме, которая во время коллинга по-разному трактует одиночные риды с нереференсным и референсным аллелями. Дело в том, что в программе GATK используется модель «уверенности в референсе» (reference confidence model) в сочетании с когортным анализом [13, 14]. Поэтому в случае получения одиночного рида, совпадающего с нереференсным аллелем, GATK считает варианты нуклеотидов в данном риде ошибкой секвенирования и не учитывает их при расчете генотипа. А в случае вариантов, найденных в одиночном риде, совпадающем с референсным аллелем, она считает ошибку маловероятной и выдает гетерозиготный (а не гомозиготный) генотип. Кроме того, большинство мутаций, которые не были подтверждены секвенированием по Сэнгеру, имели низкое покрытие (≤ 10х). Полученные результаты подтверждают, что для корректного коллинга мутаций необходимо секвенирование с достаточно большим покрытием, чтобы одиночные ошибки секвенирования не искажали результаты [15].

Необходимое покрытие в точке является вероятностной величиной и может быть рассчитано с достоверной точностью. Нами было показано, что частота ошибок в результатах, получаемых с использованием HiSeq 2500 System, соответствует погрешности прибора, и рассчитано минимально необходимое покрытие в точке для практического применения — 12х. Эта величина меньше в сравнении с результатами Bentley и соавт. [4]. Данное улучшение может быть вызвано меньшим числом ошибок за счет использования усовершенствованного прибора

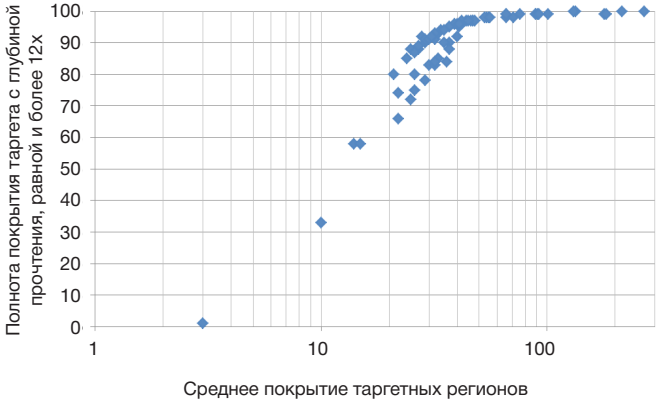

Рис. 3. Доля таргетных регионов, покрытых 12х, в зависимости от среднего покрытия таргетных регионов. Каждая точка соответствует одному образцу

22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1506	1649	1798	1954	2115	2284	2458	2639	2826	3020	3219	3426	3638	3857	4082	4313	4550	4794	5043	5299	5562	5830	6104	6386	6672	6965	7265	7570	7882
7	8	9	11	12	14	15	17	19	21	23	26	28	31	34	37	40	43	46	50	54	58	62	67	71	76	81	86	92
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
998	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
992	995	997	998	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
974	983	989	993	995	997	998	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
933	953	968	978	986	990	994	996	997	998	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
857	895	924	946	962	974	982	988	992	995	996	998	999	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
738	798	846	885	916	939	956	969	979	985	990	993	995	997	998	999	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
998	999	999	998	999	999	1000	1000	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
998	999	999	998	999	999	1000	1000	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
998	999	999	998	999	999	1000	1000	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
996	998	998	998	999	999	1000	1000	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
990	994	996	996	998	998	1000	1000	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
972	982	988	991	994	996	998	999	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
931	952	967	976	985	989	994	996	996	998	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
855	894	923	944	961	973	982	988	991	995	996	998	999	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
736	797	845	883	915	938	956	969	978	985	990	993	995	997	998	999	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
974	983	967	979	986	973	983	988	979	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
974	983	967	979	986	973	983	988	979	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
974	983	967	979	986	973	983	988	979	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
972	982	966	979	986	973	983	988	979	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
966	978	964	977	985	972	983	988	979	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
948	966	956	972	981	970	981	987	978	986	990	994	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
907	936	935	957	972	963	977	984	976	984	989	993	988	992	995	990	993	996	992	994	996	997	996	997	998	997	997	999	997
831	878	891	925	948	947	965	976	971	981	986	992	987	991	994	990	993	996	992	994	996	997	996	997	998	997	997	999	997
712	781	813	864	902	912	939	957	958	971	980	987	983	989	993	989	992	995	992	994	996	997	996	997	998	997	997	999	997
8	7	7	6	6	5	5	5	8	8	7	7	6	6	6	6	5	5	8	7	7	7	6	6	6	6	5	5	7
32	28	25	40	35	32	29	26	38	35	32	29	27	37	34	31	29	27	36	34	31	29	27	35	33	31	29	27	34
100	86	120	105	93	123	109	98	126	113	102	92	116	105	96	118	108	100	120	111	103	95	113	105	98	115	108	100	117

и новых реагентов для секвенирования. Более точные современные методы биоинформатики также позволяют увереннее фильтровать ошибки секвенирования без потери чувствительности.

ВЫВОДЫ

Нами было показано, что для достижения не менее чем 90 % покрытия таргетных регионов на 12x или более, необходимо получить количество данных, эквивалентное среднему покрытию таргета 40x после дедупликации. Данная величина может зависеть от используемого набора и протокола для обогащения, типа и длины ридов. К тому же в зависимости от протокола создания библиотек и получения исходного биоматериала обнаруживается разная степень дупликации при одинаковом объеме данных, что также должно быть учтено при расчете необходимого числа нуклеотидов на образец.

Литература

1. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014; 59 (1): 5–15.
2. de Bruin C, Dauber A. Insights from exome sequencing for endocrine disorders. Nat Rev Endocrinol. 2015; 11 (8): 455–64.
3. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15 (2): 121–32.
4. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CJ et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 56 (7218): 53–9.
5. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 2009; 19 (9): 1622–9.
6. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L et al. The diploid genome sequence of an Asian individual. Nature. 2008; 456 (7218): 60–5.
7. Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res. 2011; 21 (9): 1498–1505.
8. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barrón LT, Rosenbaum J et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 2014; 6 (10): 89.
9. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 2013; 14: 195.
10. Meynert AM, Ansari M, Fitzpatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 2014; 19 (15): 247.
11. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011; 21 (6): 940–51.
12. Illumina. Technical Note: Sequencing. Quality Scores for Next-Generation Sequencing [Internet]. [cited 2017 Jun] Available from: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
13. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant. 2015; 8 (6): 831–46.
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20 (9): 1297–303.
15. Kim K, Seong MW, Chung WH, Park SS, Leem S, Park W et al. Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. Genomics Inform. 2015; 13 (2): 31–9.

References

1. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014; 59 (1): 5–15.
2. de Bruin C, Dauber A. Insights from exome sequencing for endocrine disorders. Nat Rev Endocrinol. 2015; 11 (8): 455–64.
3. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15 (2): 121–32.
4. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CJ et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 56 (7218): 53–9.
5. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 2009; 19 (9): 1622–9.
6. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L et al. The diploid genome sequence of an Asian individual. Nature. 2008; 456 (7218): 60–5.
7. Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res. 2011; 21 (9): 1498–1505.
8. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barrón LT, Rosenbaum J et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 2014; 6 (10): 89.
9. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 2013; 14: 195.
10. Meynert AM, Ansari M, Fitzpatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 2014; 19 (15): 247.
11. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011; 21 (6): 940–51.
12. Illumina. Technical Note: Sequencing. Quality Scores for Next-Generation Sequencing [Internet]. [cited 2017 Jun] Available from: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
13. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant. 2015; 8 (6): 831–46.
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20 (9): 1297–303.
15. Kim K, Seong MW, Chung WH, Park SS, Leem S, Park W et al. Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. Genomics Inform. 2015; 13 (2): 31–9.