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

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

Although protein-coding regions make up only ~1 % of the human genome, they harbor 85 % of all disease-associated mutations [1]. In this light, clinical use is encouraged of whole-exome sequencing and other sequencing methods that employ specially designed enrichment panels targeting potentially mutant exon regions [2]

However, there are some challenges to the clinical application of whole-exome sequencing, one of them being the appropriate depth of coverage, i.e. the number of times each nucleotide has been sequenced, usually designated as 10x, 20x, 50x, etc. [3]. Good coverage ensures better identification of sequencing errors, increasing sequencing accuracy. There are two factors determining the choice of coverage depth. The first one is time and costs that are directly proportional to the number of reads performed. The second is the minimal number of reads needed to achieve the desired error tolerance. No consensus has been reached yet regarding the second factor.

Using the short-read sequencing technology by Illumina, Bently et al. discovered in 2008 that almost every homozygous single nucleotide variant (SNV) can be detected at 15x coverage, while for accurate heterozygous SNV calling 33x coverage is required [4]. Subsequently, a 33x sequencing depth was adopted as standard coverage for SNV detection [5, 6]. In 2011 Ajay et al. reported that accurate detection of 95 % of SNVs, as well as short insertions and deletions, required 50x coverage. However, further experiments that employed new, improved reagents and software for data processing produced the same yield at 35x [7]. In 2014 Fang et al. published an article demonstrating that 60x coverage is needed for accurate detection of 95 % of insertions and deletions [8].

Such discrepancy indicates that recommended sequencing depth is not something easily determined, as the number of reads per region needed for accurate variant detection depends on the read quality, which, in turn, depends on the technique applied or reagents used or the quality of sample preparation. For example, amplification of GC-rich regions during polymerase reaction (PCR) can be a problem, resulting in poor sequencing quality, urging the researcher to increase the number of reads. Currently, there are reagent kits for PCR that can improve reaction quality and thereby the quality of sequencing. In 2013 Meyner et al. discovered that depending on the reagents used, 95 % of SNPs can be detected either at 20x or 46x coverage [9]. In 2014 the same authors reported 14x coverage as sufficient for accurate detection of 95 % of SNPs [10]. Besides, Li et al. demonstrated that coverage depth also depends on the number of individual samples to be sequenced [11]. For example, for detection of mutations with frequency <0.2 %, 4x sequencing of 3,000 samples yields the same result as 30x sequencing of 2,000 samples. To sum up, there are more factors affecting sequencing quality than it might seem, and the number of reads can be efficiently reduced upon estimating a contribution of each factor or based on the study goal.

In this work we show that Genotek01 enrichment panel allows to reduce the depth of coverage to 12x to achieve accurate calling of heterozygous variants and SNVs, with only 0.5 % difference between NGS and Sanger sequencing outcomes.

METHODS

DNA extraction, preparation and sequencing of DNA libraries

DNA was extracted from whole venous blood of patients with inherited diseases, using QIAmp DNA Mini Kit (Qiagen, Germany). Quality of genomic DNA was checked by agarose gel electrophoresis; among critical purity indicators was the absence of DNA degradation and RNA contamination. Concentration of the obtained DNA was measured by Qubit 3.0 Fluorometer (Life Technologies, USA). DNA libraries were prepared using NEBnext Ultra DNA library Prep Kit for Illumina (New England Biolabs, USA) using adaptor sequences for Illumina sequencing according to the manufacturer's protocol. Dual indexed libraries were obtained using NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) by the same manufacturer. Quality control of the obtained DNA libraries was performed using Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Target enrichment of the coding regions was carried out using MYbaits (MYacroarray, USA). For 100 bp paired-end sequencing, HiSeq 2500 System analyzer (Illumina, USA), HiSeq Rapid PE Cluster Kit v2 and HiSeq Rapid SBS Kit v2 (Illumina) were used following the manufacturer's protocol.

Sanger sequencing

Amplicons were fluorescently labeled using BugDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, USA) according to the manufacturer's protocol. Sanger sequencing was performed on ABI PRISM 3500 Genetic Analyzer (Applied Biosystems, USA) according to the manufacturer's protocol.

Bioinformatic analysis

The obtained reads were aligned to the reference genome *hg19* in BWA. PCR duplicates were removed using SAMtools rmdup, variant calling was performed using Genome Analysis Tool Kit (GATK). We detected 89 mutations: 10 homo- and hemizygous, 79 heterozygous, 80 point mutations (SNPs) and 9 short insertions and deletions (indels). We also genotyped 200 nt-long regions to the left and right of the detected mutations. All positions in those regions were analyzed and then validated by Sanger sequencing, the gold-standard for detecting short mutations. Chromatography data were processed uniformly. Mutation calling was done using the original Genotek software based on BioPython and R packages (sangerseqR, seqinR, Biostrings and Rsubread). Genotypes obtained through NGS were validated by Sanger sequencing, and sensitivity and specificity were then calculated.

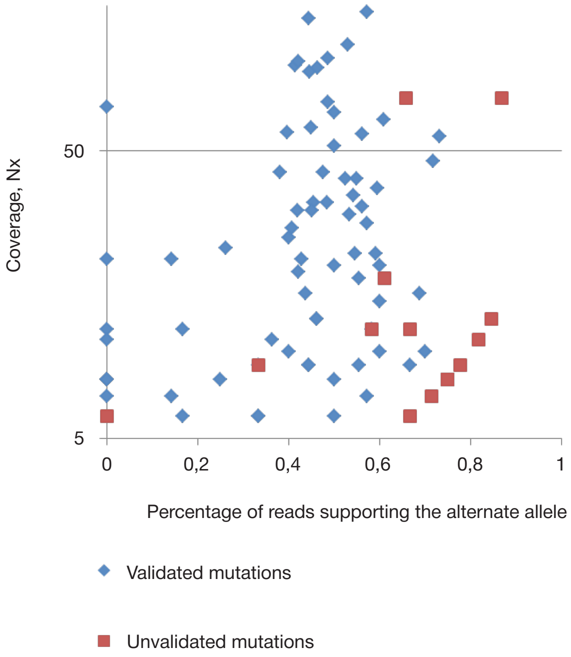

Fig. 1. Detection of Sanger-confirmed and unconfirmed mutations depending on the coverage depth and percentage of reads supporting the alternate allele. One point can represent more than one mutation

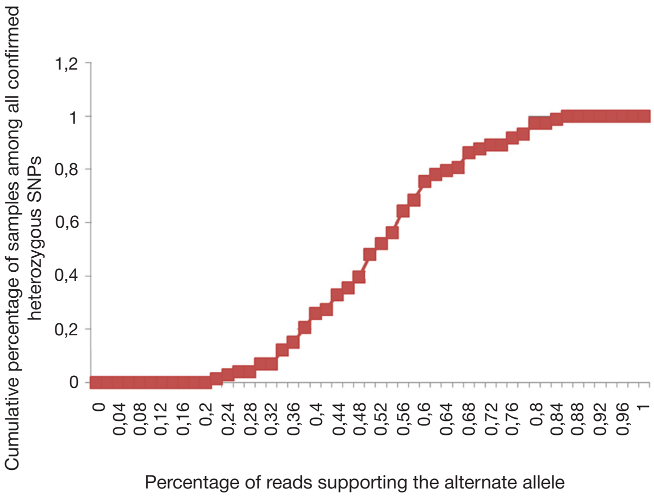

Fig. 2. Cumulative distribution of the percentage of samples with reads supporting the alternate allele X or less

Varying the number and percentage of reads for filtering out reference and alternate homozygous variants

Number of erroneous calls in a 12Mb library																						
	min. support	coverage depth	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
Number of detected variant calls 0/1 out of 1000																						
Threshold for heterozygous variants 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
Threshold for heterozygous variants 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
Threshold for heterozygous variants 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
Number of detected variant calls 1/1 with true genotype 0/1 per 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

RESULTS

Validation of mutations by Sanger sequencing

Sanger sequencing did not confirm 15 of 89 mutations detected by variant calling, meaning that they either had a different genotype (compared to the genotype identified by NGS) or were absent. Eight of 15 unconfirmed mutations were identified by NGS as heterozygous, but Sanger validation classified them as homozygous. Of note, NGS-detected heterozygosity was supported by only one read with the reference allele (see Fig. 1, the cluster of mutations in the lower right corner).

Simulation of various coverage depths

To determine the minimum depth of coverage, we ran a series of simulation tests decreasing the number of reads (bootstrapping) per mutation and the regions adjacent to it and also performed mutation calling. To estimate the error rate in the calls, we used Sanger-confirmed reference-matching homozygous positions.

Sequencing quality can be assessed using the Phred quality score (Q score) generated by the sequenator for each nucleotide [12]. However, this metric merely measures sequencing accuracy, which was insufficient for our purposes. We checked if each of the reads overlapping the position of interest supported the reference sequence, and if there were mismatches, we assumed an erroneous call.

We analyzed 372,443 nucleotides. Of them 276 did not match the reference sequence, while others did. Thus, the calculated error rate was 0.0741 %, equivalent to Q31 on the Phred quality score.

For 69 positions with confirmed heterozygous mutations, the percentage of reads supporting the alternate allele was estimated (Fig. 2).

Based on these data and the frequency of erroneous calls, we calculated the frequency of combinations at various coverage depths, ranging from 2x to 50x, and the number of reads supporting the alternate allele, ranging from 0 to the maximum. The obtained data were used to calculate frequency of 2 error types: a truly heterozygous variant identified as homozygous reference and a truly heterozygous variant identified as homozygous alternate at different cut-off levels for reference and alternate homozygous variants. To filter out homozygous reference calls, we varied the number of reads from 2 to 10. To filter out homozygous alternate calls, we varied the percentage of reads supporting the alternate allele between 70, 80 and 90 % (see the Table). We found that for short mutations (SNPs and indels) the accuracy of the applied method was as high as 99.7 %, with sensitivity of 98 % at 12x coverage. Lower coverage led to a considerable decrease in sensitivity (decreasing sigmoidal character) and therefore cannot be recommended. While planning a lab experiment, an average number of reads per base should be determined to achieve 12x coverage of the target region. Therefore, we plotted a correlation between an average depth of coverage and the percentage of the target region covered by 12 reads (Fig. 3).

It was found that to cover >90 % of the target regions at least 12x depth, 40x coverage by deduplicated reads is required.

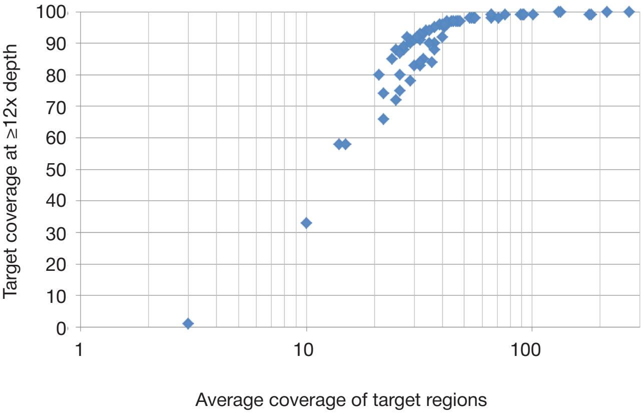

Fig. 3. Percentage of target regions sequenced at 12x depending on the average coverage of target regions. Each point represents one sample

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

RESULTS

Sanger sequencing did not confirm 15 of 89 mutations detected by NGS; 8 of 15 unvalidated mutations were homozygous and not heterozygous as suggested by NGS. Such outcome is largely dependent on the error model employed by GATK, the software used to obtain a set of variants for the studied genome, which interprets single reads with reference or non-reference alleles differently during variant calling. GATK employs the reference confidence model in combination with cohort analysis [13, 14]. Therefore, if the obtained sequence matches the non-reference allele, GATK treats the nucleotide variants from this read as sequencing errors and ignores them when calculating a genotype. If the obtained sequence matches the reference allele, GATK considers the probability of error to be low and returns a heterozygous (not a homozygous) genotype. Besides, in our study the majority of mutations unconfirmed by Sanger sequencing were detected at a low depth of coverage (≤10x). The obtained results confirm that accurate mutation calls require deep sequencing in order to avoid single sequencing errors that could distort the obtained data [15]. The depth of coverage per base is a probabilistic value and can be calculated with reliable precision. We showed that the error rate for the data obtained by HiSeq 2500 System corresponds to the instrumental error. We also calculated the minimal coverage (12x) required for accurate sequencing. This value is lower than the one proposed by Bentley et al. [4], which may be due to the improved equipment and new reagents used in our study and, therefore, fewer sequencing errors. State-of-the-art bioinformatic methods also allow for better error filtering without loss of sensitivity.

CONCLUSIONS

Our work demonstrates that to achieve at least 90 % coverage of the target genome at >12x, 40x coverage by deduplicated reads is required. This value depends on the enrichment reagents and protocol applied, read types and lengths. Besides, depending on the protocols for library preparation and nucleic acid extraction, the degree of duplication in the obtained sequences may vary, which must be accounted for when calculating the desired number of nucleotides per sample.

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.

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.