European Journal of Human Genetics (2020) 28:1459–1466
https://doi.org/10.1038/s41431-020-0656-2

ARTICLE

Medium-coverage DNA sequencing in the design of the genetic association study

Chao Xu[1,2] · Ruiyuan Zhang[1] · Hui Shen[1] · Hong-Wen Deng[1,3]

Received: 20 December 2019 / Revised: 29 April 2020 / Accepted: 12 May 2020 / Published online: 26 May 2020

Abstract

DNA sequencing is a widely used tool in genetic association study. Sequencing cost remains a major concern in sequencing-based study, although the application of next generation sequencing has dramatically decreased the sequencing cost and increased the efficiency. The choice of sequencing depth and the sequencing sample size will largely determine the final study investment and performance. Many studies have been conducted to find a cost-effective design of sequencing depth that can achieve certain sequencing accuracy using minimal sequencing cost. The strategies previously studied can be classified into two groups: (1) single-stage to sequence all the samples using either high (>~30×) or low (<~10×) sequencing depth; and (2) two-stage to sequence an affordable number of individuals at a high-coverage followed by a large sample of low-coverage sequencing. However, limited studies examined the performance of the medium-coverage (10–30×) sequencing depth for a genetic association study, where the optimum sequencing depth may exist. In this study, using a published simulation framework, we comprehensively compared the medium-coverage sequencing (MCS) to the single- and two-stage high/low-coverage sequencing in terms of the power and type I error of the variant discovery and association testing. We found, given certain sequencing effort, MCS yielded a comparable discovery power and better type I error control compared with the best (highest power) scenarios using other high- and low-coverage single-stage or two-stage designs. However, MCS was not as competent as other designs with respect to the association power, especially for the rare variants and when the sequencing investment was limited.

Introduction

DNA sequencing is a widely used tool in genetic association studies, which aimed to identify potential genes or regions that contribute to specific diseases and traits. Sequencing cost remains a major concern in DNA sequencing study [1], although the application of next generation sequencing (NGS) has dramatically decreased the sequencing cost and increased the sequencing efficiency [2]. The main factor that determines sequencing cost is sequencing depth/coverage, which is defined as the average number of a nucleotide in the genome has been sequenced in one experiment [1]. Here, we used the terms coverage and depth interchangeably for the definition aforementioned, although coverage has also been used to indicate the breadth of a sequenced genome.

A number of studies have been conducted to find a cost-effective sequencing design that can achieve certain sequencing accuracy using minimal sequencing cost [3–6]. The strategies previously studied can be classified into two groups: single-stage design and two-stage design. The single-stage design is to sequence all the samples using either high (>~30×) or low (<~10×) sequencing depth. The two-stage design is first sequencing an affordable small number of individuals at a high coverage to create a reference panel, then followed by a large sample of low-

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0656-2) contains supplementary material, which is available to authorized users.

✉ Hong-Wen Deng
hdeng2@tulane.edu

[1] Center for Bioinformatics and Genomics, Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA 70112, USA

[2] Department of Biostatistics and Epidemiology, The University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

[3] School of Basic Medical Science, Central South University, 410013 Changsha, China

coverage sequencing and perform dense imputation based on the locally constructed reference. It has been shown that using the local reference panel can help identify rare variants that may be missed in other populations [7–9]. There were plenty of studies focused on single-stage design. For example, Flannick et al. [3] developed a statistical framework to compare the sensitivity and specificity of low-coverage (0.5–4×) sequencing and SNP (single nucleotide polymorphism) array (100 k to 2.5 M) individually and jointly. Pasaniuc et al. [4] studied the performance of extremely low-depth (0.1–0.5×) sequencing in genome-wide association study. In addition, Li et al. [10] systematically compared the performance of both low and high depth (2–30×) sequencing for studies of complex traits. Regarding the two-stage design, Xu et al. [11] simulated the entire study process of sequencing-based association study and showed that the two-stage sequencing was an effective approach under certain conditions, such as for the discovery of rare variants.

However, limited studies examined the performance of medium-coverage (10–30×) sequencing for a genetic association study, where the optimum sequencing depth may exist. There were several discussions in previous studies indicating that medium-coverage may have a good performance in NGS study. Shen et al. [5] demonstrated that medium depth could achieve optimal detection power under case-only sequencing design. Nielsen et al. [12] concluded that under a large sample size, medium or low coverage might be more cost-effective than the other study designs. Xu et al. [11] specified a similar trend that the medium depth may be the optimal design in real application by comparing low-, high-coverage, and two-stage sequencing in NGS study. Recently, Gilly et al. [13] compared the genotype accuracy at depths 15×, 22.5×, and 30× by downsampling reads from a cohort of 100 samples. Their result demonstrated that the 15× was possible to achieve near-perfect sensitivity and quality for rare SNP calling and genotyping compared with 30× sequencing.

Given the lack of research on the performance of medium-coverage sequencing (MCS) in genetic association study, we comprehensively compared the medium depth (10–30×) sequencing with high/low depth sequencing in terms of variant discovery and association testing power through a simulation study.

Materials and methods

Simulation data

We generated sets of MCS data using a simulation framework published by a previous study [11]. Independent case-control scenarios were considered having equally distributed samples of European populations based on an additive genetic model. Given a certain sequencing coverage and sample size, samples of sequencing data with SNPs in a region of 100 kb were generated based on the reference genome chromosome 22. The length of 100 kb spans long enough to cover an LD block for European populations. The commonly used Illumina pair-end sequencing reads with 125 bp read length were generated by ART [14]. SNPs with minor allele frequency (MAF) ≥ 0.05, 0.01 ≤ MAF < 0.05, and MAF < 0.01 were defined as common, low-frequency, and rare SNPs, respectively. Among the simulated SNPs, 15 causal SNPs were randomly selected with the effect size (deleterious only) determined by controlling the variance it explained to be ~1% respectively for each selected SNP. In the generated simulation data, the actual variance explained by all causal variants corresponded to 10.9%. We set the disease prevalence to be 9.3%, which is similar to the prevalence of type 2 diabetes in the United States [15]. The disease associations for simulated SNPs were simulated by Hapgen2 [16] and tested by PLINK (v1.07) [17], PLINK/SEQ (v0.10, https://atgu.mgh.harvard.edu/plinkseq/), and SKAT [18]. The procedure was repeated 1000 times for each scenario. The complete simulation code is available at https://github.com/xu1912/spS-Gas.

We considered 16 MCS scenarios, which are presented in Table 1. We selected four levels of sequencing depth to cover the typical settings for medium coverage: 12×, 16×, 20×, and 24×. The sample size ranged between 200 and 4000 to make the sequencing effort comparable with the previously published results, which included 12 high-coverage, 14 low-coverage, and 51 two-stage (high coverage to impute in low coverage) scenarios.

The generated data were compared with publicly available results of high-/low-coverage sequencing scenarios, which were produced by the same simulation framework and settings. We kept all the parameters the same except for the coverage and sample size. More details for the simulation framework and settings can be found in the previous article [11].

Model evaluation

We evaluated the various scenarios with respect to the power, type I error, and FDR of variant discovery and

Table 1 Medium-coverage sequencing scenarios considered.

Depth	Sample size
12	1000, 2000, 4000
16	500, 1000, 2000, 3000
20	200, 400, 600, 1200, 1600, 2400
24	500, 1000, 2000

Table 2 Power and type I error rate of variant discovery and association in medium-coverage scenarios.

Sample size	Coverage	Effort (kpd)	Scenario	Discovery					Association				
				Total	Rare	Low frequency	Common	Overall type I error rate	Total	Rare	Low frequency	Common	Overall type I error rate
200	20	4	200@20	60.20%	38.35%	89.25%	94.64%	7.47E−05	3.24%	0.55%	3.08%	7.57%	0.013
400	20	8	400@20	68.72%	51.77%	91.07%	95.20%	8.47E−05	7.11%	1.11%	7.11%	16.99%	0.029
500	16	8	500@16	70.48%	54.62%	91.09%	95.27%	8.79E−05	8.69%	1.49%	9.54%	20.67%	0.036
500	24	12	500@24	71.76%	56.72%	91.34%	95.28%	8.73E−05	8.81%	1.40%	10.04%	21.18%	0.035
600	20	12	600@20	73.35%	59.23%	91.55%	95.43%	8.96E−05	10.20%	1.87%	12.27%	23.98%	0.042
1000	12	12	1000@12	78.82%	68.27%	91.74%	95.27%	1.15E−04	15.75%	4.10%	22.05%	35.27%	0.066
1000	16	16	1000@16	77.74%	66.23%	92.05%	95.65%	9.70E−05	15.89%	4.11%	21.91%	34.92%	0.065
1000	24	24	1000@24	78.82%	68.01%	91.94%	95.68%	9.63E−05	16.31%	4.18%	23.20%	35.89%	0.068
1200	20	24	1200@20	80.10%	70.09%	92.10%	95.73%	9.77E−05	18.30%	5.29%	26.40%	39.64%	0.077
2000	12	24	2000@12	82.14%	73.33%	92.57%	95.92%	1.04E−04	26.12%	11.04%	40.90%	50.44%	0.104
1600	20	32	1600@20	82.36%	73.71%	92.34%	95.86%	1.01E−04	22.88%	8.00%	34.63%	47.04%	0.094
2000	16	32	2000@16	83.39%	75.35%	92.63%	95.89%	1.03E−04	26.93%	11.57%	40.85%	51.40%	0.107
2000	24	48	2000@24	84.26%	76.82%	92.53%	95.89%	1.03E−04	27.01%	11.54%	41.40%	51.99%	0.107
2400	20	48	2400@20	85.31%	78.45%	92.71%	95.98%	1.04E−04	30.93%	15.18%	48.33%	56.00%	0.118
3000	16	48	3000@16	85.30%	78.50%	92.58%	95.88%	1.06E−04	35.13%	19.41%	54.43%	59.41%	0.131
4000	12	48	4000@12	87.52%	81.98%	93.05%	96.13%	1.08E−04	43.19%	28.25%	64.13%	65.44%	0.155

association under the constraint of study-wide sequencing investment. Given the same length of sequenced genome content, the sequencing investment can be represented by the sequencing effort, which was defined as the product of the sequencing coverage and sample size of that study [10, 19]. The person depth (pd) defined as 1× coverage for one person was used as the unit of the sequencing effort. One kpd means a sequencing effort of 1000 pd. Among all the scenarios, we picked several levels along the increase of the sequencing effort at 4, 8, 12, 16, 24, 32, and 48 kpd. Each effort level included multiple scenarios by different combinations of the coverage and sample size. At each level, we compared the highest power from medium-coverage scenarios to the best (highest power) from the other (low-, high-coverage, and two-stage sequencing) scenarios by the optimum power ratio ($\mathrm{OPR} = \frac{\max\{\text{Power of MC designs}\}}{\max\{\text{Power of Other designs}\}}$). Using the same scenarios having the highest power, we computed the type I error ratio (TR) of MCS and other sequencing options.

The power we evaluated contained two aspects: the variant discovery power and the association testing power, so did the type I error. The variant discovery power was defined as the proportion of the identified variants among the total variants. The variant discovery type I error was defined as the proportion of the falsely identified variants among the total non-variant sequences. The association power was defined as the proportion of statistically significant variants identified among the total true causal variants. A logistic regression model was employed to perform an association test of the variants with Bonferroni correction for multiple testing. In addition, the rare variants were examined by the region-based association test SKAT as follows: We divided each of the simulated 100 kb sequences into nine regions with a length of 20 kb by a sliding window of 10 kb. Only rare variants were considered for each region. Multiple testing was adjusted by Bonferroni correction. The test result for a specific region using SKAT was checked against whether the region contained any rare causal variants.

Results

We first assessed the performance of MCS in SNP discovery using the proposed simulation methods and compared it to single-stage high-coverage, low-coverage, and two-stage scenarios. For MCS, with the same sequencing effort, sequencing more subjects at a lower coverage usually resulted in a higher discovery power, especially for the rare variants (Table 2 and Fig. 1). For example, given the effort of 48 kpd, the rare variant discovery power increased from 76.82% to 81.98% when the sample size increased from 2000 to 4000 and coverage decreased from 24× to 12×

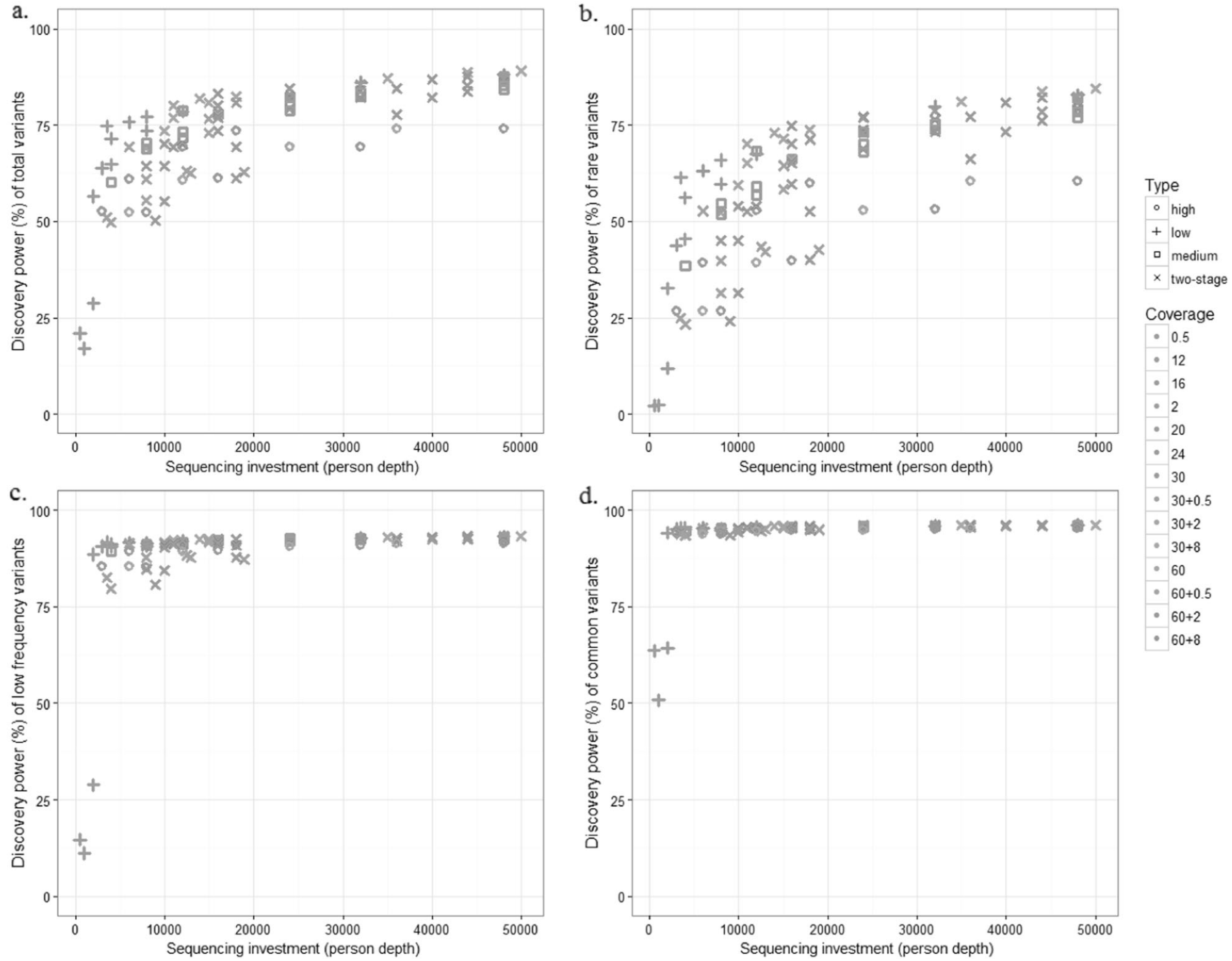

Fig. 1 Sequencing investment and discovery power of variants. The variant discovery power of high/low/medium coverage, and two-stage sequencing scenarios (denoted by symbols) using different sequencing coverages (denoted by colors) for **a** total variants; **b** rare variants (MAF < 0.01); **c** low-frequency variants (0.01 ≤ MAF < 0.05); and **d** common variants (MAF ≥ 0.05). The *x*-axis presents the sequencing investment in person depth. The *y*-axis presents the discovery power.

(Table 2). Relative to the other types of sequencing coverage under the same sequencing effort, the MCS did not exhibit greater advantage in the power for variant detection. The MCS only led at effort level of 12 kpd by a narrow margin (Fig. 1). However, we can find a two-stage scenario (100@30 + 4000@2) reaching a higher power with less sequencing effort than the MCS at 12 kpd. At other effort levels, the medium-coverage resulted in detection power close to the best (highest power) of other scenarios, most of which were low-coverage as depicted in Table 3. We defined optimum power ratio (OPR) to compare the best power performance of MCS to other sequencing designs at the same sequencing effort. In Table 3, most of the OPRs are > 0.9 except at the 4 kpd scenario. Similar results were observed for the detection of rare variants (Supplementary Table 1). At 4 kpd, the maximum detection power of MCS was only 68.30% of that using other coverages (38.35% vs 56.15% from low-coverage). On the other hand, MCS had a better type I error control for most of the effort slices. Especially, when the effort exceeded 10 kpd, the MCS outperformed other sequencing options by a low type I error and comparable power in variant detection (Supplementary Table 1). For instance, among 48 kpd scenarios, the optimum power using MCS was very close to that using other coverages (87.52% vs 88.01% from low-coverage) with a less corresponding type I error (0.01% vs 0.02%). At other slices with OPR > 0.9, the type I error of MCS was also much lower relative to the other scenarios with a higher power, such as the 8, 16, and 24 kpd with the type I error ratio (TR) of 18.45%, 33.33%, and 1.36% respectively (Table 3). Further, at 12 kpd scenarios, MCS not only yielded higher power but also much less type I error with TR of ~0.01 (0.01% vs 0.77%).

For the association analysis, we first evaluated the power of MCS in testing association for common variants. The sample size played a more important role in association than

Table 3 Optimum detection and association power comparison between MCS and other sequencing designs.

Sequencing effort (kpd)	Detection (overall)				Association (common variant)				Association (low-frequency variant)			
	MCS	Best of others*	OPR	TR	MCS	Best of others*	OPR	TR	MCS	Best of others*	OPR	TR
4	60.20%	71.48% (L)	0.84	0.75	7.60%	30.04% (L)	0.25	0.24	3.08%	17.14% (L)	0.18	0.19
8	70.48%	77.13% (L)	0.91	0.18	20.69%	52.61% (L)	0.39	0.33	9.54%	38.65% (L)	0.25	0.25
12	78.82%	78.54% (L)	1.00	0.01	35.29%	57.20% (L)	0.62	0.54	22.05%	42.98% (L)	0.51	0.45
16	77.74%	83.05% (L)	0.94	0.33	35.96%	54.20% (T)	0.66	0.56	22.38%	39.66% (T)	0.56	0.50
24	82.14%	84.47% (T)	0.97	0.01	50.44%	57.16% (T)	0.88	0.82	40.90%	42.76% (T)	0.96	0.74
32	83.39%	86.05% (L)	0.97	0.89	51.40%	62.66% (L)	0.82	0.74	40.85%	48.88% (L)	0.84	0.65
48	87.52%	88.01% (L)	0.99	0.55	65.44%	70.04% (L)	0.93	0.87	64.13%	57.61% (L)	1.11	0.79

MCS medium-coverage sequencing, *OPR* optimum power ratio, *TR* corresponding type I error ratio.

[a] Best of others is the other sequencing scenarios having the highest power. The letter in () indicated the best of other sequencing option: L—low-coverage sequencing; T—two-stage sequencing.

the choice of sequencing depth for the design using medium-coverage. Given a certain sequencing effort, the larger sample size always produced higher power (Supplementary Table 1). When collapsing the coverage axis (as depicted in Fig. 2a), an increase of sample size led to a power gain (Fig. 2a and c). However, if we collapsed the sample size axis (as depicted in Fig. 2b), the benefit of using a higher coverage was very limited (Fig. 2b, c). In addition, MCS was not as competent as other sequencing coverages in identifying disease-associated variants. The OPR was only 24.71% (7.42% vs 30.04% from low-coverage) when the sequencing effort was fixed at 4 kpd. With the sequencing effort increased from 4 to 48 kpd, the OPR elevated from 0.25 to 0.92 (Table 3). We observed a similar trend for the type I error of association testing using MCS relative to other sequencing coverages. The higher sequencing effort, the closer type I error using MCS to other sequencing options, while the type I error of MCS was always less (Table 3). In Fig. 3a, we plotted the common variant association power versus type I error for the scenarios with optimum power using MCS and other sequencing depth under the same sequencing efforts. The points from optimum MCS and other sequencing coverages likely lie in the same line, which also provided a general idea of how the sequencing-based genetic association study performed given our model assumption. Similar results were also observed for the association test of low-frequency variants (Table 3 and Fig. 3b). However, MCS was able to achieve high power to identify low-frequency variants with less type I error than the other sequencing options. When the sequencing effort is large, at 48 kpd, MCS yielded higher power (64.13% from 4000@12 vs 57.61% from low-coverage 6000@8) and 21% less type I error based on the comparison of same scenarios (17.18% vs 21.71%).

Further, we examined the performance of MCS on the rare variants association testing using SKAT. For the MCS scenarios at the same sequencing efforts, we observed the larger sample size, the higher association power (Table 4). Increasing the sample size was more beneficial than increasing the sequencing coverage, not only in common and low-frequency variants testing, but also for the rare variants. In the scenarios with optimum power using other sequencing coverages, the testing power quickly rose to ~90% when the sequencing effort was over 8 kpd. However, the best power from MCS was far less than others at the same investment, most of which were low-coverage and two-stage design as depicted in Supplementary Table 2. The OPRs were all less than 0.5 until the sequencing effort increased to 32 kpd (Supplementary Table 2). Different from the general association test for common and low-frequency variants, the rare variants were grouped into regions and tested by regions. In Fig. 4, the medium coverage may reach the same power of other designs with a less type I error. For instance, the top right end of the MCS designs (4000@12) compared with the bottom left end of the best power of others (low-coverage 2000@2). But MCS was not efficient in rare variant association study with respect to the sequencing investment.

Discussion

In this study, we investigated the efficiency of sequencing studies using medium-coverage compared with single-stage low-/high-coverage and two-stage sequencing designs. Assuming a disease prevalence of ~9.3% and 15 causal variations, our result is generalizable to the study of complex diseases, like type 2 diabetes, but not rare diseases or single-gene disorders. Our result showed that, given certain sequencing investment, the variant discovery power of MCS was just close to the best power of the other schemes. However, the medium-coverage yielded less type I errors in variant discovery compared with the best (highest power) of others. Further, the variant association testing power of

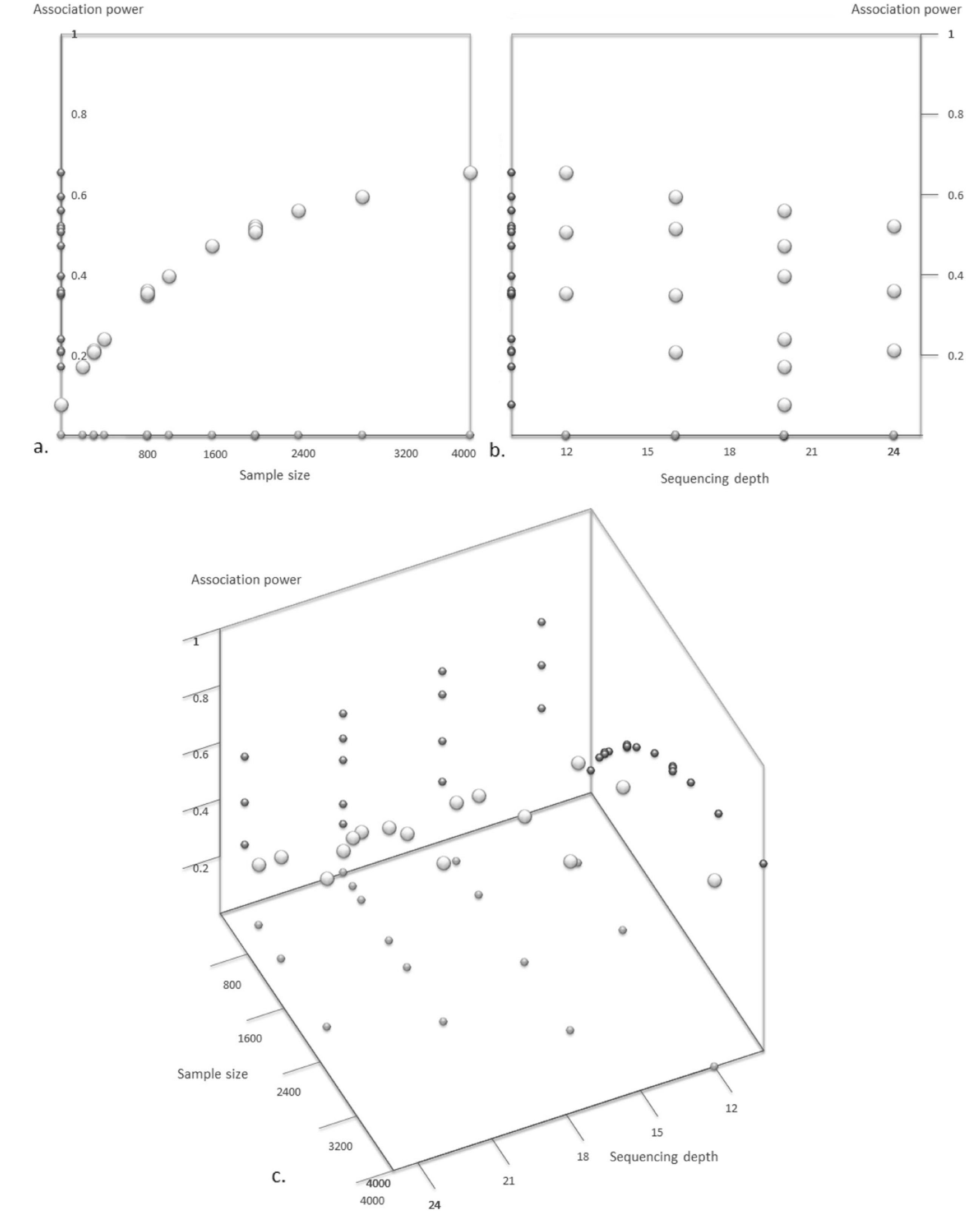

Fig. 2 Common variant association power versus sequencing depth and sample size. a Collapsed 2D view of association power vs sample size; **b** Collapsed 2D view of association power vs sequencing depth; **c** 3D plot of association power versus sequencing depth and sample size. The green dots represent values in 3D space. The purple, red, and orange dots are projections into three subplots (2D), respectively.

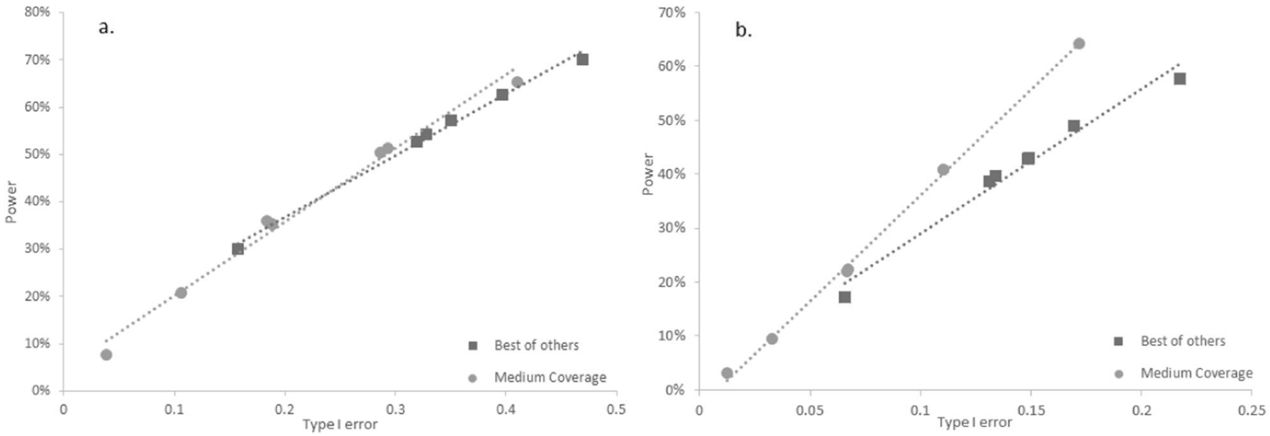

Fig. 3 Common and low-frequency variant association power and type I error. **a** Common variant; **b** Low-frequency variant.

Table 4 Power and type I error rate of region-based rare variant association in MCS.

Sample size	Coverage	Effort (kpd)	Scenario	Power	Type I error
4000	12	48	4000@12	77.20%	0.447
3000	16	48	3000@16	67.10%	0.385
2400	20	48	2400@20	56.50%	0.305
2000	24	48	2000@24	48.30%	0.262
2000	16	32	2000@16	47.90%	0.258
1600	20	32	1600@20	36.40%	0.169
2000	12	24	2000@12	46.20%	0.247
1200	20	24	1200@20	24.60%	0.117
1000	24	24	1000@24	17.70%	0.082
1000	16	16	1000@16	19.00%	0.097
1000	12	12	1000@12	17.70%	0.098
600	20	12	600@20	5.10%	0.027
500	24	12	500@24	2.70%	0.014
500	16	8	500@16	2.70%	0.016
400	20	8	400@20	0.90%	0.005
200	20	4	200@20	0.00%	0.000

MCS medium-coverage sequencing.

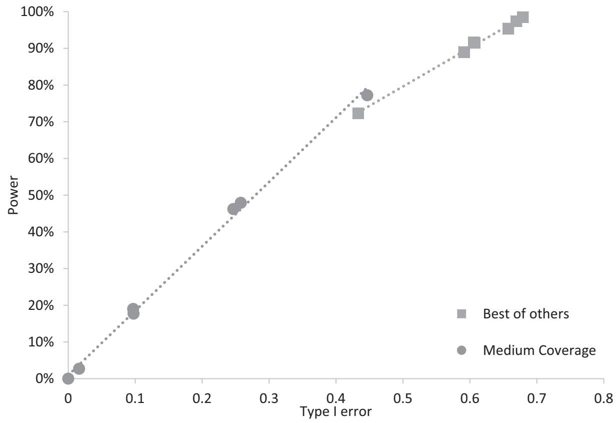

Fig. 4 Rare variant association power and type I error. Region-based rare variant association power and type I error for the scenarios with optimum power at series of sequencing efforts using medium-coverage sequencing (MCS) and other sequencing depths.

using medium-coverage was not comparable with the best power of others for most of the cases.

Based on our findings, we can make an addendum to existing guidelines on designing a cost-effective sequencing study given a certain sequencing investment. If the study goal is only detecting variants/mutations and there is sufficient sequencing effort (≥12 kpd), medium-coverage is applicable in addition to low-coverage and appropriate two-stage design. Medium-coverage can reach a tightly close power but will be less error-prone relative to low-coverage and two-stage design. However, if the study goal is not only detecting but also identifying disease-associated variants, MCS may not be appropriate. Instead, low-coverage and two-stage schemes should be adopted. In summary, we provided a recommendation table for the selection of low-/medium-/high-coverage and two-stage sequencing when designing a genetic association study (Supplementary Table 3).

Different from the study of SNP, the detection of structural variations (SVs) using NGS mostly relies on the sequencing depth, such as the copy number variations (CNVs) and indels. Gilly et al. found that genotype accuracy is substantially more dependent on sequencing depth for indels than for SNPs [13]. In a recent study, the performance of several CNV detection tools varied with the sequencing depth, with high-coverage resulted in high sensitivity and specificity [20]. We expected the sequencing coverage would play a more important role than that in the present study of SNP. The high-coverage sequencing may be more appropriate for scenarios targeted on SVs. The cost-efficiency of various sequencing coverage in the study of indels, CNVs, and other SVs needs further investigation.

By comparing MCS scenarios together with single-stage low-/high-coverage and two-stage sequencing scenarios, we have investigated nearly all the practically used sequencing depth. Some general guidelines can be concluded to design a sequencing-based association study. First, the sample size is more important than the sequencing depth for the association test. Second, increasing the sequencing coverage does help in decreasing the error of variant detection. Above all, one limitation of our study is that the default parameter setting in the simulation was used without fine-tuning and thus the observed type I error may not represent the true number in real-world analysis. However, our study still faithfully presented the performance of MCS and other sequencing coverage.

Acknowledgements The work was partially supported by grants from the National Institutes of Health (R01 AR059781, R01 MH104680, R01 AR069055, U19 AG055373, and P20GM109036), Edward G. Schlieder Endowment, and startup funds from Tulane University. This research was supported in part using high performance computing (HPC) resources and services provided by Technology Services at Tulane University, New Orleans, LA.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
3. Flannick J, Korn JM, Fontanillas P, Grant GB, Banks E, Depristo MA, et al. Efficiency and power as a function of sequence coverage, SNP array density, and imputation. PLoS Comput Biol. 2012;8:e1002604.
4. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44:631–5.
5. Shen Y, Song R, Pe'er I. Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association. Bioinformatics. 2011;27:1995–7.
6. Wu Y, Zheng ZL, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 2017;18:86.
7. Danjou F, Zoledziewska M, Sidore C, Steri M, Busonero F, Maschio A, et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat Genet. 2015;47:1264.
8. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur J Hum Genet. 2014;22:1321–6.
9. Kreiner-Moller E, Medina-Gomez C, Uitterlinden AG, Rivadeneira F, Estrada K. Improving accuracy of rare variant imputation with a two-step imputation approach. Eur J Hum Genet. 2015;23:395–400.
10. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011;21:940–51.
11. Xu C, Wu K, Zhang JG, Shen H, Deng HW. Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. Genet Epidemiol. 2017;41:187–97.
12. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
13. Gilly A, Suveges D, Kuchenbaecker K, Pollard M, Southam L, Hatzikotoulas K, et al. Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. Nat Commun. 2018;9:4674.
14. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28:593–4.
15. Centers for Disease C. Prevention: National Diabetes Statistics Report: estimates of diabetes and its burden in the United States. Atlanta, GA: US Department of Health and Human Services; 2014.
16. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011;27:2304–5.
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
19. Wetterstrand KA. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP); 2016. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
20. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol. 2019;15:e1007069.