frontiers in Genetics

METHODS
published: 26 October 2021
doi: 10.3389/fgene.2021.736390

A Novel Quality-Control Procedure to Improve the Accuracy of Rare Variant Calling in SNP Arrays

Ting-Hsuan Sun[1,2], Yu-Hsuan Joni Shao[3,4], Chien-Lin Mao[2], Miao-Neng Hung[2], Yi-Yun Lo[2], Tai-Ming Ko[1,5] and Tzu-Hung Hsiao[2,6,7,8]**

[1]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, [2]Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan, [3]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, [4]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan, [5]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, [6]Department of Public Health, Fu Jen Catholic University, New Taipei City, Taiwan, [7]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan, [8]Research Center for Biomedical Science and Engineering, National Tsing Hua University, Hsinchu, Taiwan

OPEN ACCESS

Edited by:
Gillian Belbin,
Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Xiang Zhan,
The Pennsylvania State University, United States
Shilin Zhao,
Vanderbilt University, United States

***Correspondence:**
Yu-Hsuan Joni Shao
jonishao@tmu.edu.tw
Tzu-Hung Hsiao
thsiao@vghtc.gov.tw

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 05 July 2021
Accepted: 21 September 2021
Published: 26 October 2021

Citation:
Sun T-H, Shao Y-HJ, Mao C-L, Hung M-N, Lo Y-Y, Ko T-M and Hsiao T-H (2021) A Novel Quality-Control Procedure to Improve the Accuracy of Rare Variant Calling in SNP Arrays. Front. Genet. 12:736390. doi: 10.3389/fgene.2021.736390

Background: Single-nucleotide polymorphism (SNP) arrays are an ideal technology for genotyping genetic variants in mass screening. However, using SNP arrays to detect rare variants [with a minor allele frequency (MAF) of <1%] is still a challenge because of noise signals and batch effects. An approach that improves the genotyping quality is needed for clinical applications.

Methods: We developed a quality-control procedure for rare variants which integrates different algorithms, filters, and experiments to increase the accuracy of variant calling. Using data from the TWB 2.0 custom Axiom array, we adopted an advanced normalization adjustment to prevent false calls caused by splitting the cluster and a rare het adjustment which decreases false calls in rare variants. The concordance of allelic frequencies from array data was compared to those from sequencing datasets of Taiwanese. Finally, genotyping results were used to detect familial hypercholesterolemia (FH), thrombophilia (TH), and maturity-onset diabetes of the young (MODY) to assess the performance in disease screening. All heterozygous calls were verified by Sanger sequencing or qPCR. The positive predictive value (PPV) of each step was estimated to evaluate the performance of our procedure.

Results: We analyzed SNP array data from 43,433 individuals, which interrogated 267,247 rare variants. The advanced normalization and rare het adjustment methods adjusted genotyping calling of 168,134 variants (96.49%). We further removed 3916 probesets which were discordant in MAFs between the SNP array and sequencing data. The PPV for detecting pathogenic variants with 0.01%<MAF≤1% exceeded 99.37%. PPVs for those with an MAF of ≤0.01% improved from 95% to 100% for FH, 42.11% to 85.19% for TH, and 18.24% to 72.22% for MODY after adopting our rare variant quality-control procedure and experimental verification.

Conclusion: Adopting our quality-control procedure, SNP arrays can adequately detect variants with MAF values ranging 0.01%~0.1%. For variants with MAF values of ≤0.01%,

experimental validation is needed unless sequencing data from a homogeneous population of >10,000 are available. The results demonstrated our procedure could perform correct genotype calling of rare variants. It provides a solution of pathogenic variant detection through SNP array. The approach brings tremendous promise for implementing precision medicine in medical practice.

Keywords: SNP array, rare variant, quality-control, genotyping, disease screening

INTRODUCTION

Globally, over 7,000 rare diseases affect 5–10% of the population (Richmond et al., 2021). Most of these diseases are caused by rare pathogenic variants, which have a minor allele frequency (MAF) in a population of <1% with high penetrance (Gautheron and Jéru, 2020). For example, inherited retinal degenerations are caused by mutations of 271 genes, including the EYS and ABCB4 genes (Chen et al., 2021; Retnet, (1996). Recent studies also discovered several pathogenic variants which are associated with common traits or complex diseases, such as hyperlipidemia, myocardial infarction, and diabetes (Lee et al., 2014; Riddle et al., 2020; Vrablik et al., 2020; Momozawa and Mizukami, 2021). Although detecting rare variants is important, it is still challenged because of the low MAF values. Large-scale genomic data are needed to identify pathogenic variants with a large effect and high penetrance.

Recently, several biobanks have been set up and have collected large-scale genetic data, including single-nucleotide polymorphism (SNP) arrays and next-generation sequencing (NGS) data (Chen et al., 2011; Bycroft et al., 2018; Kanai et al., 2018). Through such data, many rare pathogenic variants have been identified (Cirulli et al., 2020), (Blauwendraat et al., 2021) and have been widely used as disease-associated genetic markers, including monogenic (Firdous et al., 2018) and complex diseases (Marvel et al., 2017; Jurgens, 2020; Patel et al., 2020; Chen et al., 2021). High-density SNP arrays provide a rapid and efficient method to simultaneously genotype hundreds of thousands of specific variants (Kim and Misra, 2007; Visscher et al., 2017; Kim et al., 2018). The Taiwan Biobank has utilized SNP arrays to discover specific variants associated with hereditary diseases, drug metabolism, and drug responses of the Han Chinese population in Taiwan (Lin et al., 2019). Also, several companies, such as 23andme, provide direct-to-consumer genetic testing services which assess genetic risks for diseases or health conditions using SNP arrays (Tandy-Connor et al., 2018; Horton et al., 2019; Schleit et al., 2019). However, recent articles have pointed out the low accuracy and high false positive rates of SNP arrays for detecting rare variants (Wright et al., 2019; BMJ, 2021).

Variant calling of SNP arrays relies on clustering of probeset signals (Lamy et al., 2006). Clustering of rare variants becomes very difficult when only a limited number of alternative alleles exist (Weedon, 2019). Differences in signal distributions due to batch effects also cause misclustering. As shown as **Supplementary Figure S1** in "Supplemental materials", samples in a batch with high average signals of major alleles can be misclassified as alternative alleles. Noise signals induced in the experiment, such as by air bubbles or scratches, also cause incorrect calling (**Supplementary Figure S1B**). In addition, cross-hybridization reactions with non-target sequences can also induce false calling for probesets with low specificity for targeting sequences. The performance of probesets for rare variants is difficult to evaluate because of the low frequency of alternative alleles. Although several algorithms or procedures have been developed to improve the accuracy of genotyping of common variants (Hua et al., 2007; Xiao et al., 2007; Hunter-Zinck et al., 2020), methods that focus on rare variant calling for SNP arrays are still lacking. New strategies are needed to improve the accuracy of rare variants for further applications.

The objective of this study is to develop a rare variant quality-control (QC) procedure to improve the calling accuracy. We proposed a procedure combining advance normalization, rare het adjustment, and MAF comparisons to improve true positive rate. This approach was evaluated by Sanger validation or real-time Polymerase Chain Reaction (qPCR) and an external data set. As we demonstrate, our method provides a solution which makes SNP arrays feasible as screening tools for rare variants.

MATERIALS AND METHODS

Dataset of Single-Nucleotide Polymorphism Arrays

We used SNP arrays from the project of Taiwan Precision Medicine Initiative (TPMI) to conduct our data analysis. In total, 43,531 individuals were recruited from Taichung Veterans General Hospital (TCVGH; Taichung, Taiwan). Participants consented for blood to be drawn and SNP arrays to be performed, as well as for their clinical information to be linked. DNA of participants was extracted for genotyping on a custom Axiom array, TWB2.0, which was designed by the TWB based on 970 whole-genome sequence (WGS) data in the Taiwanese population (Lin et al., 2019).

In total, 714,461 probesets were designed on the TWB 2.0 Array plate (Santa Clara, CA, United States). It contained about 415,000 probesets for gene-wide association studies (GWASs) and imputation and also about 114,000 probesets for risk or pathogenic analysis selected from several sources, including ACMG, ClinVar, GWAS Catalog, HGMD, and the literature. We selected 267,247 probe-sets associated with rare variants (with an MAF of <1% in NGS data) to evaluate the calling accuracy of rare variants.

Genotype Calling, Advanced Normalization, and Rare Het Adjustment

Genotype calling was based on Affymetrix® Power Tools (APT, command-line software, Santa Clara, CA, United States). Data of 24 plates were grouped as a batch according to the processing date. After genotyping, we applied advanced normalization to adjust misclustering based on the batch effect. Advanced normalization was conducted with the advnorm package provided by Thermo Fisher Scientific (Santa Clara, CA, United States). We also applied a rare het adjustment to exclude probesets with different signals in replicated probes. The rare het adjustment was conducted with the axiomBestPractices-1.2.4 program and the command "-do-rare-het-adjustment" (see supplemental methods).

Comparing Allelic Frequencies Between the Single-Nucleotide Polymorphism Array and NGS Data

We collected 3,370 WGSs to estimate the MAFs for all variants. Data of 1,200 WGSs were accessed from TWB, and 2,170 of them were collected from our in-house program. Briefly, about 30× whole-genome sequencing was performed. Reads were aligned to the GRCh38 human genome using the GATK pipeline. Values of the MAF were calculated using VCFtools.

To compare allelic frequencies, MAFs between the SNP array and sequencing data were $\log_2$ transformed. $\log_2$ ratios of MAFs between the SNP array and sequencing data were calculated and utilized as parameters to estimate the MAP concordance. With the predict method Setting the upper and lower thresholds as 1.72 and −2, respectively, the probe-sets with mis-concordance of MAFs were identified and excluded.

Variant Validation Based on a Quantitative Polymerase Chain Reaction (qPCR) or Sanger Sequencing

We respectively selected 1,090, 55, and 132 probesets of familial hypercholesterolemia (FH), thrombophilia (TH), and maturity-onset diabetes of the young (MODY) for disease-oriented analyses. The MAF distributions of the probesets are shown in **Supplementary Table S2** in "Supplemental materials". All samples with heterozygous calls were validated by a qPCR or Sanger sequencing. We used the Applied Biosystems™ Primer Designer™ Tool (Applied Biosystems, Santa Clara, CA, United States) to pick specific primer pairs for Sanger sequencing, we designed primers with the Primer3 algorithm (Kumar and Chordia, 2015) and checked for sequence similarities throughout the human genome using the Primer-BLAST tool (Ye et al., 2012).

Independent Dataset for External Validation

Newly collected SNP array data of 5,358 samples was used to measure the reproducibility of our QC procedure. We used the same workflow to genotype calling. Samples with heterozygous calls were verified through qPCR or Sanger sequencing to evaluate calling accuracy.

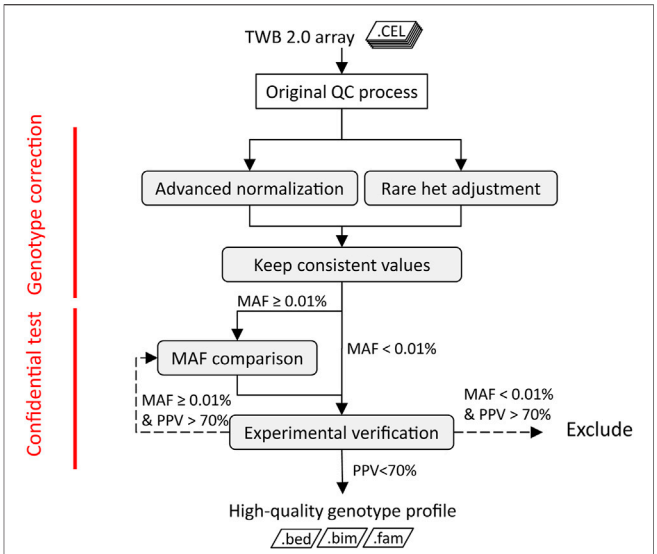


FIGURE 1 | Quality-control procedure of TWB 2.0 arrays of rare variants. The shaded boxes are steps of the rare variant quality-control procedure. The dotted line represents the management of variants with low PPV.

RESULTS

An Analytical Algorithm for Rare Variant Detection

We implemented a QC procedure that contains four key components to precisely detect rare variants in SNP arrays. As shown in **Figure 1A**, two algorithms, advanced normalization and rare het adjustment, were applied to post-QC data. The rare het adjustment checked signals of heterozygous calls from each replicate probe group and compared the signal distribution in a batch. Heterozygous calls were adjusted to "no call" if the signal distribution from the replicate probe group was uncertain. Advanced normalization detected clustering errors from specific plates in batches and reassigned those calls to the correct cluster. We consolidated results from the two algorithms and filtered conflicting results. In the third step, we compared the concordance of the MAF of each variant in the array with the corresponding variant in Taiwanese WGS data. Probesets with high deviations of MAF noted as low-concordance probes were excluded from the following analysis. Last, we assessed the performance of the genotyping results in disease screening and verified the results by a qPCR or Sanger. We designed this integrating approach to improve the quality of SNP array data in genotyping rare variants.

Genotype Correction Procedure Adjusts Incorrect Calls on the Array

We analyzed SNP array data from 43,433 individuals, which interrogated 267,247 rare variants. We assessed the crude number and rate of adjustments made by the two algorithms in correlation with the MAF. We adjusted 136,773 calls with the

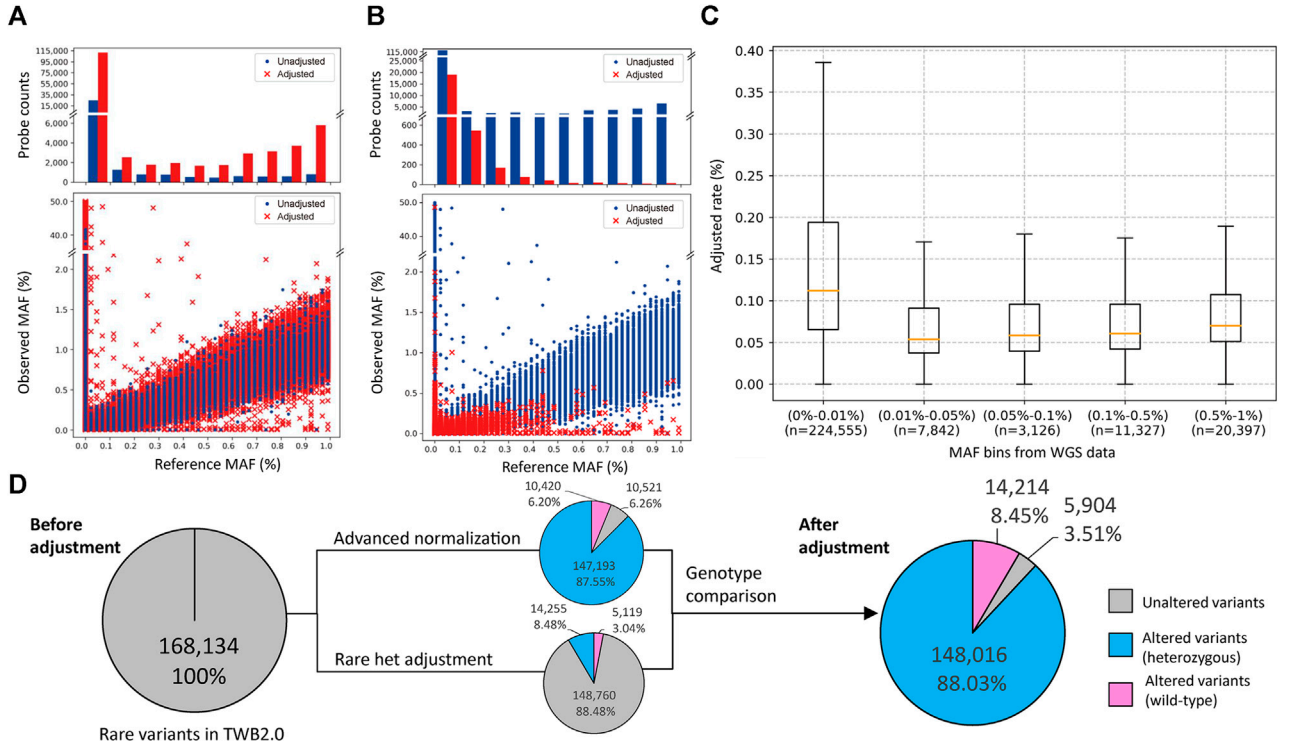


FIGURE 2 | Minor allele frequency (MAF) distribution and probeset count after performance of the rare variant quality-control procedure. **(A,B)** The former is the MAF distribution after using the advanced normalization method, and the latter is the MAF distribution after using the rare het adjustment method. The upper plot shows the number of probesets adjusted (red bar) or unadjusted (blue bar) by this algorithm. The bottom plot shows the distribution of MAFs between the single-nucleotide polymorphism (SNP) array and sequencing data. Adjusted data points are marked with a red "×", and unadjusted data points are marked with a blue dot. **(C)** Box plot of the adjusted rate per MAF bin based on sequencing data. **(D)** Number of rare variants with at least one heterozygous call in the original data, and after advanced normalization and rare het adjustment. Variants not adjusted by our algorithms are represented as a gray slice. Variants adjusted by our algorithms are represented as a blue slice. Variants adjusted by our algorithms are represented as a pink slice.

advanced normalization method (**Figure 2A**) and 19,347 calls with the rare het adjustment method (**Figure 2B**). **Figure 2C** shows distributions of variants that were adjusted by MAF. The adjustment rate was the highest in variants with MAFs of ≤0.01%.

In these genotype callings, 168,134 variants had at least one heterozygous call, and 99,110 variants were completely wild-type. The advanced normalization algorithm adjusted 10,420 variants (6.20%) to be the wild-type and modified the number of heterozygous calls in 147,193 variants (87.55%). The rare het adjustment algorithm adjusted 5,119 variants (3.04%) to be the wild-type and modified the number of heterozygous calls in 14,255 variants (8.48%). Taken together, our algorithm adjusted the genotyping calling of 162,230 variants (96.49%) in which 14,214 variants (8.45%) were adjusted to the wild-type and 148,016 variants (88.03%) were modified. Only 5,904 variants (3.51%) remained the same after these adjustments (**Figure 2D**).

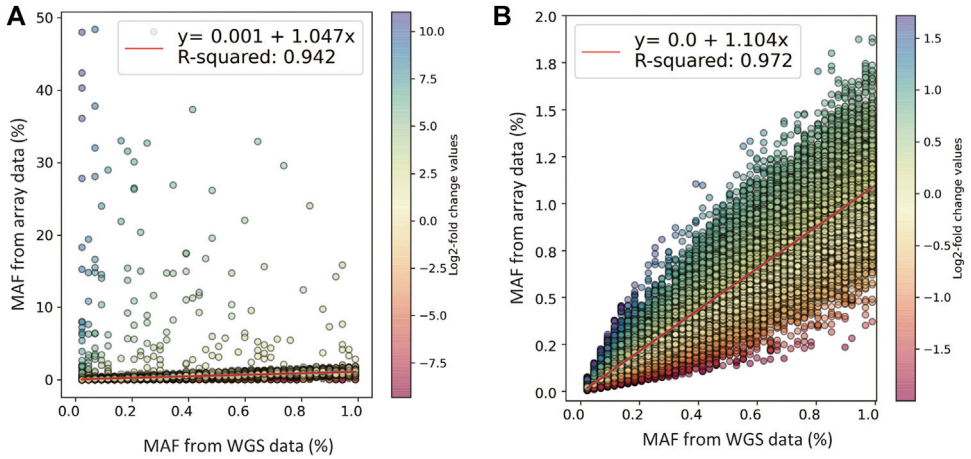
Discordance by Minor Allele Frequencys

We used log2 ratio of the computed MAFs as the parameter to compare between the SNP array and NGS (**Figure 3A**). By setting the upper threshold value to 1.72 and lower to −2, we were able to obtain the highest coefficient of determination in the MAF scatter plot. We removed 3,916 low-concordance probe-sets which were discordant in MAFs between the SNP array and NGS. The remaining probesets showed good concordance. The coefficient of the linear regression was 1.104, and the coefficient of determination was 0.972 (**Figure 3B**).

The number of discordant probesets in each MAF group is presented in **Figure 3C**. Compared to 0.6% of the probesets in the group of 0.5% < MAF≤1% which showed low concordance, 30.4% of the probesets in the group of 0.05% < MAF≤0.01% showed low concordance. This step excluded lots of non-working probesets. However, variants in the group of MAF<0.01% could not be applied due to limited MAF resolution of sequencing data.

Disease Screening and Experimental Validation of the True Positive Rate

To test the performance of genotype calling in disease screening after applying our analytical procedure, we investigated pathogenic variants of three hereditary diseases: FH, TH, and MODY. Totals of 1,090, 132, and 55 pathogenic variants on the SNP array were respectively associated with FH, MODY, and TH. There were 499 mutation carriers detected in FH, 76 in TH, and 148 in MODY in the original data. Numbers of carriers are presented in **Table 1** by 0.01% < MAF≤1% and MAF≤0.01%. We verified all samples with heterozygous genotypes through Sanger sequencing or a qPCR. Using our rare-variant QC procedure and



C

Categories	On TWB 2.0 *Number of variants*	Filtered by WGS *Number of variants (%)*	Output *Number of variants (%)*
(0.01 - 0.05%)	7,842	2,384 (30.4)	5,458 (69.6)
(0.05% - 0.1%)	3,126	673 (21.53)	2,453 (78.47)
(0.1% - 0.5%)	11,327	736 (6.5)	10,591 (93.5)
(0.5% - 1%)	20,397	123 (0.6)	20,274 (99.4)
All	42,692	3,916 (9.17)	38,776 (90.83)

FIGURE 3 | Quality-control assessments on the TWB 2.0 data after performance of rare variant quality-control procedures. **(A)** Comparison of minor allele frequencies (MAFs) between sequencing data and the TWB 2.0 before excluding low-concordance probes. **(B)** Comparison of MAFs between sequencing data and the TWB 2.0 after excluding low-concordance probes. **(C)** The distribution of excluded probes according to the MAF bins.

TABLE 1 | Performance of the TWB 2.0 in detecting rare pathogenic variants for familial hypercholesterolemia, thrombophilia, and maturity onset diabetes of the young in data with the original quality-control process, with rare variants quality-control procedure, and experimental verification.

	Familial hypercholesterolemia						Thrombophilia						Maturity-onset diabetes of the young					
—	0.01% < MAF≤1%			MAF≤0.01%			0.01% < MAF≤1%			MAF≤0.01%			0.01% < MAF≤1%			MAF≤0.01%		
Dataset	**TPs**	**FPs**	**PPV (%)**	**TPs**	**FPs**	**PPV (%)**	**TPs**	**FPs**	**PPV (%)**	**TPs**	**FPs**	**PPV (%)**	**TPs**	**FPs**	**PPV (%)**	**TPs**	**FPs**	**PPV (%)**
Original	474	5	98.96	19	1	95.00	19	0	100	24	33	42.11	—	—	—	27	121	18.24
With rare variants QC	470	3	99.37	17	0	100	19	0	100	23	4	85.19	—	—	—	26	10	72.22
External dataset (5,358 samples)	67	1	98.53	2	0	100	2	0	100	0	1	0	—	—	—	2	0	100

MAF, minor allele frequency; TPs, true positives; FPs, false positives; PPV positive predictive value.
The qPCR and Sanger sequencing results are the gold standard for TP and FP.

experimental verification, the PPV improved from 98.96 to 99.37% in variants with 0.01% < MAF≤1% and from 95 to 100% in variants with MAF≤0.01% in FH. The PPV remained 100% in variants with 0.01% < MAF≤1%, and it improved from 42.11 to 85.19% in variants with MAF≤0.01% in TH. In detecting pathogenic variants in MODY, all variants were in the group of MAF≤0.01%, and the PPV improved from 18.24 to 72.22% after our QC approach. We also examined 1.35% of negative calls and reached 100% of negative predictive value (NPV). The filtration trace in each step is shown in **Supplementary Table S3**.

In addition, we used an independent dataset of 5,358 samples to evaluate our procedures as an external validation. We reached 98.57, 100 and 66.67% of positive predictive value in familial hypercholesterolemia (FH), thrombophilia (TH), and maturity-onset diabetes of the young (MODY) respectively. (**Table 1**)

DISCUSSION

SNP array was widely used for variant calling. This proposed quality-control procedure can improve the accuracy of rare

variant calling to extend the application of SNP array. Although NGS has high accuracy of rare variant detection and considered as the gold standard, it requires high computational power and skilled bioinformaticians for variant calling. The computational process is massive when we have a large number of samples. SNP array is an alternative method to detect known-pathogenic variants. It is more convenient and efficient for variant calling due to the characteristic of probe hybridization comparing to WGS. This procedure makes SNP array suitable for pathogenic variant screening. Also, it enables us to utilize available biobank data for studying pathogenic variant in a population level.

We developed a robust QC pipeline which can effectively adjust false positive calling in rare variants and increase the positive detection rate. Two major algorithms, rare het adjustment and advance normalization, were used to correct false signals, while MAF comparisons tagged low-concordance probesets. Our data showed dramatic improvements in true positive rates. The PPVs for MODY and TH improved from 18 to 93% and from 57 to 100%, respectively. The demonstrated performance indicated that SNP array data combined with our QC algorithm could be directly applied to large-scale disease screening for the Taiwanese population.

The data of the original genotype calling algorithm showed a poor performance for rare variants in some cases. Positive rates of TH and MODY were as low as 56.58 and 17.53%, respectively. One of the potential reasons is that the original genotyping pipeline was designed for common variants (Bush and Moore, 2012). Most rare variants lack alternative alleles, causing an extremely skewed dataset for initial genotype gating and cluster splitting. This leads to incorrect genotype calling (King and Nicolae, 2014). Another reason for false calling could be induced by abnormal fluorescence signals due to bubbles or scratches. Although most probesets of the SNP array were designed for repetitive probes and randomly distributed at different locations to eliminate the effect, extremely high signals from bubbles or scratches will increase the average signal and raise false calls (BioRxiv, 2020). We introduced the rare het adjustment to eliminate this effect. It changed the unexpected result to "no call". Wrong clustering caused by batch effects is another source of incorrect calling. This kind of probeset often has a high intensity in the genotype cluster plot, leading to a missed split into the wrong cluster. To target this issue, advanced normalization was applied to identify batch effects and reassign genotype clusters. As the result we demonstrated, the two algorithms increased the PPVs of TH and MODY. By combining the two procedures, the performance of base calling of rare variants could be dramatically increased to 91.30 and 72.22%, respectively.

Incorrect calls from low-concordance probes are an important issue in rare variant detection. They can be caused by an improper probe design and non-specific hybridization. We utilized the procedure of MAF comparison to check the concordance of MAFs between array data and sequencing. Any probes with out-of-range MAF values were identified as low-concordance probes and excluded. This procedure marked 9.17% of probesets with low performance in the array. However, the procedure only works on variants with a frequency of >0.01% due the resolution of sequencing data. For probesets that interrogated variants with a frequency of <0.01%, experimental validation was used to test the concordance of the probesets.

Our data reveal the challenge of detecting rare mutations, especially of variants with a frequency of <0.01%. The positive rate decreased when the MAF decreased. For example, the positive detection rate of FH was up to 98.8% in the original data. MAFs of FH variants were mainly in the range of 0.5–0.01%. However, the positive rate of MODY was down to 17.53%, because the frequency of all variants was <0.01%. By considering the issue discussed above, our procedure demonstrated significant improvement in the positive prediction rate without losing many true positive calls, but one positive call was lost for TH and MODY.

We used the Log_2 ratios of MAFs between the SNP array and sequencing data as a parameter to estimate the MAF concordance. Ideally, the coefficient of regression line should be 1 if the MAFs of SNP array and sequencing are came from the same samples. However, the MAFs of SNP data and sequence data are derived from different cohort in our study. We observed the difference between array MAFs and WGS MAFs. Taking two variants we did experimental validation as example, the MAF of rs749038326 is 0.0842% in SNP array, but it is 0.0461% in the sequencing data. The MAF of another SNP, rs730882109, is 0.192% in the array data, but it is 0.0691% in the sequencing data. Few reasons can help to explain this phenomenon. First, the SNP data and WES data were derived from different cohorts. Different age distribution and disease condition of two cohorts may cause the discrepancy. In addition, the total number in the SNP data and the WES data are different. MAFs estimated from WGS did not provide enough resolution for rare variants.

As increasing numbers of novel variants are discovered from sequencing and WGS approaches, custom-designed genotyping arrays are an alternative strategy for investigating low-frequency and rare variants for large cohorts with the advantage of low costs (Hurd and Nelson, 2009; Berry et al., 2019). It The procedure we developed provides an excellent solution to overcome genotyping call issues in rare variants. In addition, as results we demonstrated, SNP arrays can be used for genetic disease screening. This has great potential for clinical utilization based on the advantage of low costs and low demands for computational power. The capacity of SNP arrays is up to millions of SNPs. All known pathogenic variants of genetic diseases in populations could be screened simultaneously. Our procedure provides a solution of correct genotype calling. Combined together, the approach brings tremendous promise for implementing precision medicine in medical practice.

DATA AVAILABILITY STATEMENT

The SNP array data in this study is from TPMI projects (https://tpmi.ibms.sinica.edu.tw/www/en/). A portion of the WGS data are from Taiwan BioBank (https://www.twbiobank.org.tw/new_web_en/about-export.php). The data analyzed in this study is subject to the following licenses/restrictions: One can apply to

access Taiwan Biobank Data. Requests to access these datasets should be directed to biobank@gate.sinica.edu.tw.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the This study was funded by Academia Sinica 40-05-GMM and AS-GC-110-MD02. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

T-HS, Y-HS, and T-HH conceived and designed the study. T-HS, Y-HS, and T-HH contributed to the conception of the research project. T-HS and C-LM analyzed the data. M-NH and Y-YL completed the experiments. T-HS prepared the draft manuscript. Y-HS and T-HH critically reviewed the manuscript. All authors discussed the results and contributed to the preparation of the final manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank all the participants and investigators from Taiwan Precision Medicine Initiative. This study was funded by Academia Sinica 40-05-GMM and AS-GC-110-MD02.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.736390/full#supplementary-material

REFERENCES

Berry, N. K., Scott, R. J., Rowlings, P., and Enjeti, A. K. (2019). Clinical use of SNP-microarrays for the detection of genome-wide changes in haematological malignancies. *Crit. Rev. Oncology/Hematology* 142, 58–67. doi:10.1016/j.critrevonc.2019.07.016

BioRxiv (2020). *Rare Heterozygous Adjusted Genotyping*. Available at: https://downloads.thermofisher.com/Axiom_Analysis/tech-note-Axiom%20-RHA_final_Rev_0.6.pdf.

Blauwendraat, C., Makarious, M. B., Leonard, H. L., Bandres-Ciga, S., Iwaki, H., Nalls, M. A., et al. (2021). A population scale analysis of rare SNCA variation in the UK Biobank. *Neurobiol. Dis.* 148, 105182. doi:10.1016/j.nbd.2020.105182

BMJ. Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ*, 2021. 372: p. n792.

Bush, W. S., and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *Plos Comput. Biol.* 8 (12), e1002822. doi:10.1371/journal.pcbi.1002822

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562 (7726), 203–209. doi:10.1038/s41586-018-0579-z

Chen, T.-C., Huang, D.-S., Lin, C.-W., Yang, C.-H., Yang, C.-M., Wang, V. Y., et al. (2021). Genetic characteristics and epidemiology of inherited retinal degeneration in Taiwan. *Npj Genom. Med.* 6 (1), 16. doi:10.1038/s41525-021-00180-1

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., et al. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* 40 (6), 1652–1666. doi:10.1093/ije/dyr120

Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* 11 (1), 542. doi:10.1038/s41467-020-14288-y

Firdous, P., Nissar, K., Ali, S., Ganai, B. A., Shabir, U., Hassan, T., et al. (2018). Genetic Testing of Maturity-Onset Diabetes of the Young Current Status and Future Perspectives. *Front. Endocrinol.* 9, 253. doi:10.3389/fendo.2018.00253

Gautheron, J., and Jéru, I. (2020). The Multifaceted Role of Epoxide Hydrolases in Human Health and Disease. *Int. J. Mol. Sci.* 22, 1. doi:10.3390/ijms22010013

Horton, R., Crawford, G., Freeman, L., Fenwick, A., Wright, C. F., and Lucassen, A. (2019). Direct-to-consumer genetic testing. *BMJ* 367, l5688. doi:10.1136/bmj.l5688

Hua, J., Craig, D. W., Brun, M., Webster, J., Zismann, V., Tembe, W., et al. (2007). SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics* 23 (1), 57–63. doi:10.1093/bioinformatics/btl536

Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B. R., Ji, S.-G., Sun, N., et al. (2020). Genotyping Array Design and Data Quality Control in the Million Veteran Program. *Am. J. Hum. Genet.* 106 (4), 535–548. doi:10.1016/j.ajhg.2020.03.004

Hurd, P. J., and Nelson, C. J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief. Funct. Genomics Proteomics* 8 (3), 174–183. doi:10.1093/bfgp/elp013

Jurgens, S. J. (2020). Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. *bioRxiv*, 2020, 2020 . .11.29.402495.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* 50 (3), 390–400. doi:10.1038/s41588-018-0047-6

Kim, J. M., Santure, A. W., Barton, H. J., Quinn, J. L., Cole, E. F., Visser, M. E., et al. (2018). A high-density SNP chip for genotyping great tit (Parus major) populations and its application to studying the genetic architecture of exploration behaviour. *Mol. Ecol. Resour.* 18 (4), 877–891. doi:10.1111/1755-0998.12778

Kim, S., and Misra, A. (2007). SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* 9, 289–320. doi:10.1146/annurev.bioeng.9.060906.152037

King, C., and Nicolae, D. (2014). GWAS to Sequencing: Divergence in Study Design and Analysis. *Genes* 5 (2), 460–476. doi:10.3390/genes5020460

Kumar, A., and Chordia, N. (2015). In silico PCR primer designing and validation. *Methods Mol. Biol.* 1275, 143–151. doi:10.1007/978-1-4939-2365-6_10

Lamy, P., Andersen, C. L., Wikman, F. P., and Wiuf, C. (2006). Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res.* 34 (14), e100. doi:10.1093/nar/gkl475

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95 (1), 5–23. doi:10.1016/j.ajhg.2014.06.009

Lin, J.-C., Chen, L.-K., Hsiao, W. W.-W., Fan, C.-T., and Ko, M. L. (2019). Next Chapter of the Taiwan Biobank: Sustainability and Perspectives. *Biopreservation and Biobanking* 17 (2), 189–197. doi:10.1089/bio.2018.0119

Marvel, S. W., Rotroff, D. M., Wagner, M. J., Buse, J. B., Havener, T. M., McLeod, H. L., et al. (2017). Common and rare genetic markers of lipid variation in subjects with type 2 diabetes from the ACCORD clinical trial. *PeerJ* 5, e3187. doi:10.7717/peerj.3187

Momozawa, Y., and Mizukami, K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. *J. Hum. Genet.* 66 (1), 11–23. doi:10.1038/s10038-020-00845-2

Patel, A. P., Wang, M., Fahed, A. C., Mason-Suares, H., Brockman, D., Pelletier, R., et al. (2020). Association of Rare Pathogenic DNA Variants for Familial Hypercholesterolemia, Hereditary Breast and Ovarian Cancer Syndrome,

and Lynch Syndrome With Disease Risk in Adults According to Family History. *JAMA Netw. Open* 3 (4), e203959. doi:10.1001/jamanetworkopen.2020.3959

Retnet (1996). Available from:. . https://sph.uth.edu/retnet/.

Richmond, P. A., Av-Shalom, T. V., Fornes, O., Modi, B., Elliott, A. M., and Wasserman, W. W. (2021). GeneBreaker: Variant simulation to improve the diagnosis of Mendelian rare genetic diseases. *Hum. Mutat.* 42 (4), 346–358. doi:10.1002/humu.24163

Riddle, M. C., Philipson, L. H., Rich, S. S., Carlsson, A., Franks, P. W., Greeley, S. A. W., et al. (2020). Monogenic Diabetes: From Genetic Insights to Population-Based Precision in Care. Reflections From a Diabetes Care Editors' Expert Forum. *Dia Care* 43 (12), 3117–3128. doi:10.2337/dci20-0065

Schleit, J., Naylor, L. V., and Hisama, F. M. (2019). First, do no harm: direct-to-consumer genetic testing. *Genet. Med.* 21 (2), 510–511. doi:10.1038/s41436-018-0071-z

Tandy-Connor, S., Guiltinan, J., Krempely, K., LaDuca, H., Reineke, P., Gutierrez, S., et al. (2018). False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet. Med.* 20 (12), 1515–1521. doi:10.1038/gim.2018.38

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101 (1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Vrablik, M., Tichý, L., Freiberger, T., Blaha, V., Satny, M., and Hubacek, J. A. (2020). Genetics of Familial Hypercholesterolemia: New Insights. *Front. Genet.* 11, 574474. doi:10.3389/fgene.2020.574474

Weedon, M. N. (2019). Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. *bioRxiv*, 696799.

Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., et al. (2019). Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.* 104 (2), 275–286. doi:10.1016/j.ajhg.2018.12.015

Xiao, Y., Segal, M. R., Yang, Y. H., and Yeh, R.-F. (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics* 23 (12), 1459–1467. doi:10.1093/bioinformatics/btm131

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13, 134. doi:10.1186/1471-2105-13-134

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.