*Briefings in Bioinformatics*, 00(00), 2020, 1–12

**doi: 10.1093/bib/bbaa084**
Problem Solving Protocol



# Using off-target data from whole-exome sequencing to improve genotyping accuracy, association analysis and polygenic risk prediction

Jinzhuang Dou†, Degang Wu†, Lin Ding, Kai Wang, Minghui Jiang, Xiaoran Chai, Dermot F. Reilly, E. Shyong Tai, Jianjun Liu, Xueling Sim, Shanshan Cheng and Chaolong Wang

Corresponding authors: Chaolong Wang, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. E-mail: chaolong@hust.edu.cn; Shanshan Cheng, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. E-mail: sscheng@hust.edu.cn
†These authors contributed equally to this work.

## Abstract

Whole-exome sequencing (WES) has been widely used to study the role of protein-coding variants in genetic diseases. Non-coding regions, typically covered by sparse off-target data, are often discarded by conventional WES analyses. Here, we develop a genotype calling pipeline named WEScall to analyse both target and off-target data. We leverage linkage disequilibrium shared within study samples and from an external reference panel to improve genotyping accuracy. In an application to WES of 2527 Chinese and Malays, WEScall can reduce the genotype discordance rate from 0.26% (SE$=6.4\times10^{-6}$) to 0.08% (SE$=3.6\times10^{-6}$) across 1.1 million single nucleotide polymorphisms (SNPs) in the deeply sequenced target regions. Furthermore, we obtain genotypes at 0.70% (SE$=3.0\times10^{-6}$) discordance rate across 5.2 million off-target SNPs, which had $\sim$1.2$\times$ mean sequencing depth. Using this dataset, we perform genome-wide association studies of 10 metabolic traits. Despite of our small sample size, we identify 10 loci at genome-wide significance ($P<5\times10^{-8}$), including eight well-established loci. The two novel loci, both associated with glycated haemoglobin levels, are *GPATCH8-SLC4A1* (rs369762319, $P=2.56\times10^{-12}$) and *ROR2* (rs1201042, $P=3.24\times10^{-8}$). Finally, using summary statistics from UK Biobank and Biobank Japan, we show that polygenic risk prediction can be significantly improved for six out of nine traits by incorporating off-target data ($P<0.01$). These results demonstrate WEScall as a useful tool to facilitate WES studies with decent amounts of off-target data.

**Key words:** whole-exome sequencing; linkage disequilibrium; low-coverage off-target data; genome-wide association study; polygenic risk score

**Jinzhuang Dou** and **Degang Wu** are postdoctoral fellows at School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
**Lin Ding, Kai Wang** and **Minghui Jiang** are graduate students at School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
**Xiaoran Chai** is a research officer at Singapore National Eye Centre, Singapore.
**Dermot F. Reilly** was a principal scientist at Merck Research Laboratories, Kenilworth, NJ, USA.
**E. Shyong Tai** is a professor at Saw Swee Hock School of Public Health, Duke-NUS Medical School, and Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
**Jianjun Liu** is the deputy executive director at Genome Institute of Singapore and a professor at Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
**Xueling Sim** is an assistant professor at Saw Swee Hock School of Public Health, National University of Singapore, Singapore.
**Shanshan Cheng** is an associate professor at the Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
**Chaolong Wang** is a professor at the Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.
**Submitted:** 17 January 2020; **Received (in revised form):** 9 April 2020

## Introduction

Although genome-wide association studies (GWAS) have successfully uncovered thousands of loci associated with human diseases or quantitative traits [1], most GWAS were based on array genotyping of common variants and, thus, had difficulty to detect rare causal variants [2]. To overcome this limitation, the whole-exome sequencing (WES) technique was developed around 2009 [3], when whole-genome sequencing (WGS) was prohibitively expensive. Despite the obvious limitation of WES for ignoring non-coding regions, it is sensible to prioritize protein-coding genes, which account for ~1.5% of the human genome, in order to reduce the sequencing cost. In the past decade, WES has achieved great success in identifying causal variants underlying both Mendelian disorders and complex traits [4–6]. Even though the sequencing cost has been dramatically reduced by now, WES remains a popular and useful tool in both clinical diagnosis and basic research.

Nevertheless, functional genomic studies have revealed important regulatory roles of non-coding variants [7–9]. The interpretation for numerous GWAS hits in non-coding regions has shifted from tagging the causal coding variants by linkage-disequilibrium (LD) to the direct causality of non-coding regulatory elements [1, 10]. Many complex diseases and traits were found to follow a polygenic architecture, where hundreds of genetic variants across the genome contribute to the phenotype. In particular, polygenic modelling of genome-wide single nucleotide polymorphisms (SNPs) has enabled explanation of a large proportion of the heritability of complex traits [11], improved disease risk prediction [12] and facilitated Mendelian randomisation (MR) analysis for causal inference between risk factors and complex diseases [13–15]. It has also been reported that polygenic risk can interact with monogenic risk variants for important genomic conditions, which may lead to better risk stratification for prevention [16]. These results collectively highlight the limitation of WES to capture the genuine disease variants residing in non-coding regions.

A potential approach to overcome this limitation without additional sequencing cost is to utilize the off-target sequencing data. WES, like other target sequencing experiments, can produce off-target by-products due to the imperfect capture techniques in selecting target DNA fragments. It has been reported by several studies that many off-target sequencing reads are sparsely distributed across the vast non-coding regions in different WES datasets, totalling an amount at the same order of magnitude as those aligned to the target regions [17–20]. These by-product data, however, are typically not considered in conventional WES analysis. In contrast, we and others have previously shown that the shallow off-target sequencing data can be used to infer population structure and cryptic relatedness, and thus facilitate downstream analyses [18, 20–22]. It has also been demonstrated that through an imputation approach, off-target sequencing data can be used to detect genetic association with complex traits/diseases [17]. In fact, LD-based genotype calling has been widely applied in shallow WGS studies [23–25]. We expect such an approach will become more powerful with improvement of imputation algorithms and the increasing amount of WGS data available as reference panels [26–29].

In this study, we implement a flexible pipeline named WEScall to perform LD-based calling specifically for WES data with decent amounts of off-target reads. WEScall treats target and off-target regions separately given their dramatic difference in sequencing depth, and leverages LD information from both the study sample and an external reference panel to improve genotyping accuracy (Figure 1). We evaluate the performance of WEScall in a WES dataset consisting of 1299 Chinese and 1228 Malays in Singapore [22]. In addition, to demonstrate the benefits of incorporating off-target data, we analyse both coding and non-coding variants for association with 10 metabolic traits, and for polygenic risk prediction of nine metabolic traits whose GWAS summary statistics are available from UK Biobank (UKB) [30] and Biobank Japan (BBJ) [31]. Our pipeline is publicly available at https://github.com/dwuab/WEScall.

## Material and methods

### WEScall pipeline

Our WEScall pipeline was built upon the GotCloud pipeline (TopMed freeze 3 version) [25] with substantial modifications to tailor for WES data (Figure 1). We partitioned WES data into target and off-target regions, and analysed them with different procedures to ensure incorporating low-coverage off-target data would not compromise the analysis of the high-coverage target data.

WEScall takes standard BAM/CRAM files as the input and performs variant discovery in the targeted exonic regions using the 'vt discover' option in GotCloud. We do not perform variant discovery across the off-target regions because extremely low-coverage data can potentially lead to many false positives. Instead, we combine the list of 12 million common biallelic SNPs that have minor allele frequency (MAF) > 0.01 in the 1000 Genomes Project (1KGP) [26] with the novel variants detected in the target regions to form a union set of candidate variants for subsequent joint genotyping.

Next, we calculate genotype likelihoods at each candidate variant jointly across all samples using the 'vt joint_genotype_sequential' option, which takes mapping and base quality scores and sample-specific DNA contamination rates into account [32, 33]. To control for the memory usage and computational speed, our pipeline will automatically split the genome into small disjoint regions (1 Mb by default) and distribute parallel jobs for joint calling within each region. In this step, a series of sequencing features will be collected to reflect the quality of each candidate variant, including the average fraction of reference allele across all heterozygotes (ABE), Z-score for association between allele and strand (STZ), inbreeding coefficient (IBC), inflated rate of observing other alleles (IOR), Z-score for association between allele and sequencing cycle (CYZ) and phred-scale quality score for the assertion of alternative allele (QUAL). Detailed description of these features can be found on the TopMed pipeline website (Web resources).

We then follow the GotCloud pipeline to use a support vector machine (SVM) to filter low-quality variants in target regions [25]. We train the SVM model by assigning positive labels to the variants in target regions overlapped with the 1KGP Omni2.5M array dataset, and negative labels to variants in target regions meeting at least 3 of the following criteria: $ABE > 0.7$; $|STZ| > 5$; $IBC < -0.1$; $IOR > 2$; $CYZ < -5$; $QUAL < 5$. An optimal threshold of the SVM score will be automatically chosen by maximizing the Youden index of the SVM classifier. Variants in target regions with SVM scores below the threshold will be filtered. We do not apply the SVM to off-target variants because (1) we only consider known variants from the 1KGP for off-target regions, and (2) sequencing features of the off-target regions are systematically different from those of the target regions.

Finally, we use BEAGLE v4.1 to perform LD-based genotype calling, which integrates information from both individual

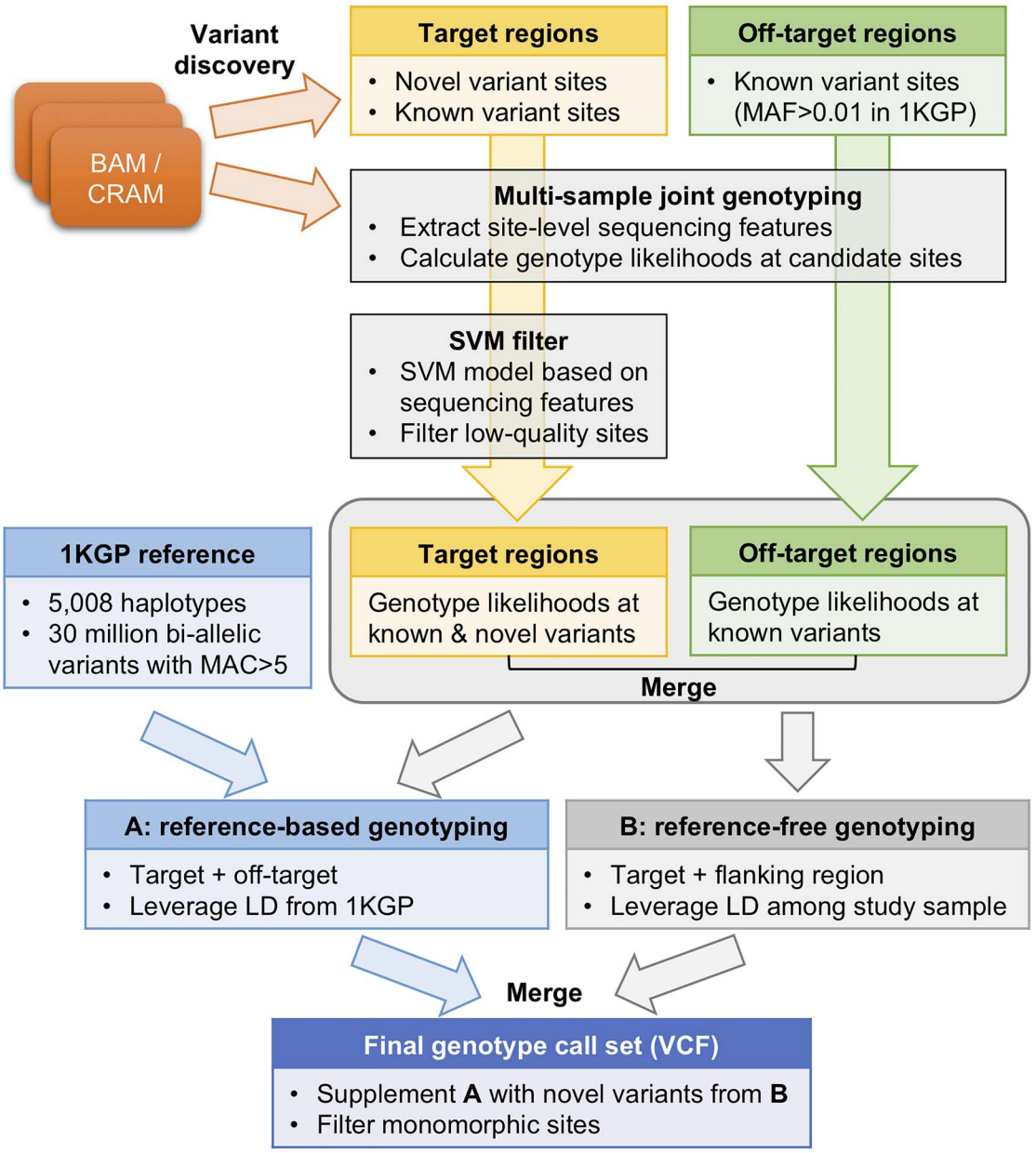


**Figure 1**. Workflow of the WEScall pipeline.

sites through genotype likelihoods and the neighbouring sites through LD [23]. We use the 1KGP dataset as an external reference panel to improve the genotyping accuracy. Nevertheless, this reference-based approach will discard any variants not present in the 1KGP dataset. We therefore supplement an additional round of reference-free genotype calling, which only leverages LD information within the study samples. The final genotype call set is obtained by combining novel variants from the reference-free calling and known variants from the reference-based calling. Sites that are monomorphic among the study sample are excluded. We have implemented our pipeline to automatically split the genome into small overlapping chunks, perform genotype calling in massive parallelization for individual chunks and merge the results. The default settings are 10 000 markers per chunk with 1000 markers overlapping with adjacent regions.

In addition, we have implemented a module for quality control (QC) of the genotype call set. Specifically, we exclude sites that have dosage $r^2$ (DR2 output by BEALGE v4.1) <0.8, fail to pass the test for Hardy–Weinberg equilibrium (HWE, $P < 10^{-5}$), or have unusually high sequencing depths (>300× for target regions and >50× for off-target regions) because they are likely to reside in repetitive regions. In addition, we exclude SNPs within 5 bp of insertions/deletions (INDELs) reported by 1KGP and remove sites for which >5% of the samples have maximum genotype probability <0.9. Empirically, we found these filters useful to improve the genotyping quality. Users can set different filtering thresholds when using this QC module. In general, because most of these QC metrics have bimodal distributions with low-quality SNPs having, e.g. very low dosage $r^2$, small HWE *P*-values or unusually high sequencing depths, quality of the final call set will not be sensitive to different filtering thresholds around the

above settings. However, depending on the properties of the samples and WES experimental design, users might consider relaxing the HWE threshold in the presence of strong population structure, or adjusting the sequencing depth filtering criteria according to average sequencing depth (e.g. filtering sites with depth >10 times of the mean depth).

## Application to WES data from the Singapore Living Biobank Project

WES data from the Living Biobank Project have been described previously [22]. Briefly, 2671 samples were whole-exome sequenced on the Illumina HiSeq2000 platform (125 bp paired end). The target regions were captured using the Nimblegen SeqCap EZ Exome v3 kits (Roche cat no: 06465692001). We aligned sequencing reads to the human reference genome (GRCh37) using BWA-MEM (v0.7.12) [34], followed by removal of PCR duplicates using Sambamba (v0.6.4) [35] and base quality score recalibration using GATK (v3.6) [36]. The resultant BAM files were checked for DNA contamination using VerifyBamID [37]. Samples with contamination rate >0.08 or mean sequencing depth across target regions <10× were removed. The final sample size was 2527, including 1299 Chinese and 1228 Malays.

We applied WEScall to generate four LD-based call sets for comparison: (1) without using off-target data or an external reference panel; (2) including the 1KGP reference panel; (3) including off-target data and (4) including both off-target data and the 1KGP reference panel. Genotypes obtained from each call set, without QC filtering, were compared against Illumina OmniExpress array genotyping data of 2451 individuals to derive genotype discordance rate for SNPs in target and off-target regions, respectively. For the array genotyping data, we excluded SNPs with call rate <0.95, HWE $P < 10^{-5}$ in either Chinese or Malays, or MAF <0.01, resulting in 595 668 autosomal SNPs for the comparison. Sequencing depth, defined as the average number of sequencing reads mapped to each variant, has major impacts on the genotyping quality, especially when the sequencing depth is low. We therefore compared genotype discordance rates for samples at different sequencing depths averaged across variants, and for variants at different sequencing depths averaged across samples.

Furthermore, we compared population structure and kinship estimates derived from our sequencing call set to those derived from array genotypes. We generated a Singapore ancestry map by applying principal components analysis (PCA) on array genotypes from the Singapore Genome Variation Project (SGVP) [38]. We then projected our study samples onto the SGVP map using the 'trace' program in the LASER package with default settings [21]. The SGVP dataset consisted of genotypes across 1 141 519 autosomal SNPs with MAF > 0.05 for 96 Chinese, 89 Malays and 83 Indians sampled in Singapore [38]. Among SNPs in SGVP, 505 641 were overlapped with our OmniExpress array data, and 733 494 were overlapped with our sequencing call set. Kinship coefficients between pairs of study samples were estimated using the SEEKIN-het estimator [22], which accounts for population structure and admixture based on the first two PCs from the LASER analysis. We classified a pair of individuals as the $k$th-degree relatives if their estimated kinship coefficient fell between $2^{-k-3/2}$ and $2^{-k-1/2}$ ($k = 0, 1, 2$ and 3) [39]. Sequencing-based and array-based estimates, including the first two PCs from LASER and the kinship coefficients, were compared using the Pearson's correlation.

## Genetic association with metabolic traits

Based on the final call set, we performed genetic association analysis with 10 metabolic traits, including body mass index (BMI), waist-to-hip ratio (WHR), systolic blood pressure (SBP), diastolic blood pressure (DBP), total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), fasting blood glucose (FBG) and glycated haemoglobin (HbA1c). TG was log-transformed to reduce the skewness of its distribution. Outliers >5 standard deviations away from the mean of each trait were removed. Three pairs of duplicates/-monozygotic twins were removed due to inconsistent age or sex. We also removed 98 individuals on diabetic medications or with undiagnosed diabetes defined as HbA1c ≥ 6.5 (%, or 48 mmol/mol) or FBG ≥ 7 (mmol/l) for all association analyses. In addition, for association analyses with lipid traits (TC, LDL, HDL and TG), we excluded 68 individuals on cholesterol medications. Supplementary Data, Table S7 summarised the sample QC procedure for each trait. For 66 individuals on blood pressure medications, we adjusted their blood pressures by adding 10 (mmHg) to SBP and 5 (mmHg) to DBP. We performed single-variant association tests using the linear mixed model implemented in GEMMA, adjusting for age, age$^2$, sex and the first two ancestry PCs [40]. The cryptic relatedness among subjects was accounted by random effects, for which the between-subject correlation was specified as twice of the estimated kinship coefficient. $P$-values and effect sizes were obtained by Wald tests. By convention, we considered variants with $P < 5 \times 10^{-8}$ as genome-wide significant and $P < 10^{-6}$ as suggestive. We picked the most significant variant as the index variant of a locus and grouped the significant/suggestive variants within 500 kb as one locus. Novel association loci were defined as those locating >500 kb apart from known loci of the corresponding trait in the GWAS Catalog [41]. We used LocusZoom to visualize genome-wide significant loci [42].

## Polygenic risk prediction

Because polygenic risk prediction requires GWAS summary statistics from large-scale studies, we downloaded data for HbA1c, FBG, TC, HDL, LDL, TG, SBP, DBP and BMI from UKB [30] and BBJ [31] (Web resources) and performed fixed-effect inverse-variance meta-analysis to obtain more robust effect size estimates for each trait [43]. Summary statistics of WHR were not available in either UKB or BBJ, while summary statistics for FBG were only available in UKB. Given that the European sample size (from UKB) was much larger than the East Asian sample size (from BBJ), we chose a European LD reference panel based on 503 Europeans from 1KGP. We calculated polygenic risk score (PRS) using SNPs shared by UKB, BBJ and our Living Biobank WES dataset. We chose the clumping and thresholding (P + T) method implemented in the PRSice 2.0 program [44], where we set the association threshold to $P < 10^{-6}$ and clumped SNPs with LD $r^2 > 0.2$ within a 250 kb window. The predictive value of PRS was measured by the adjusted $R^2$ under a linear model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$ vector of trait values for $n$ individuals; $\mathbf{X}$ is an $n \times 7$ matrix with columns being the intercept, PC1, PC2, age, age$^2$, sex and PRS; $\boldsymbol{\beta}$ is a $7 \times 1$ vector of the corresponding coefficients and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector with each element $\epsilon_i \sim N(0, 1)$ for $i = 1, 2, \ldots n$. We compared PRS calculated using SNPs within target regions (a) and from both target and off-target regions (b). To test if $R^2$ from (b) is significantly higher than $R^2$ from (a), we performed 10 000 bootstraps by sampling the same number of individuals from the original data with replacement and

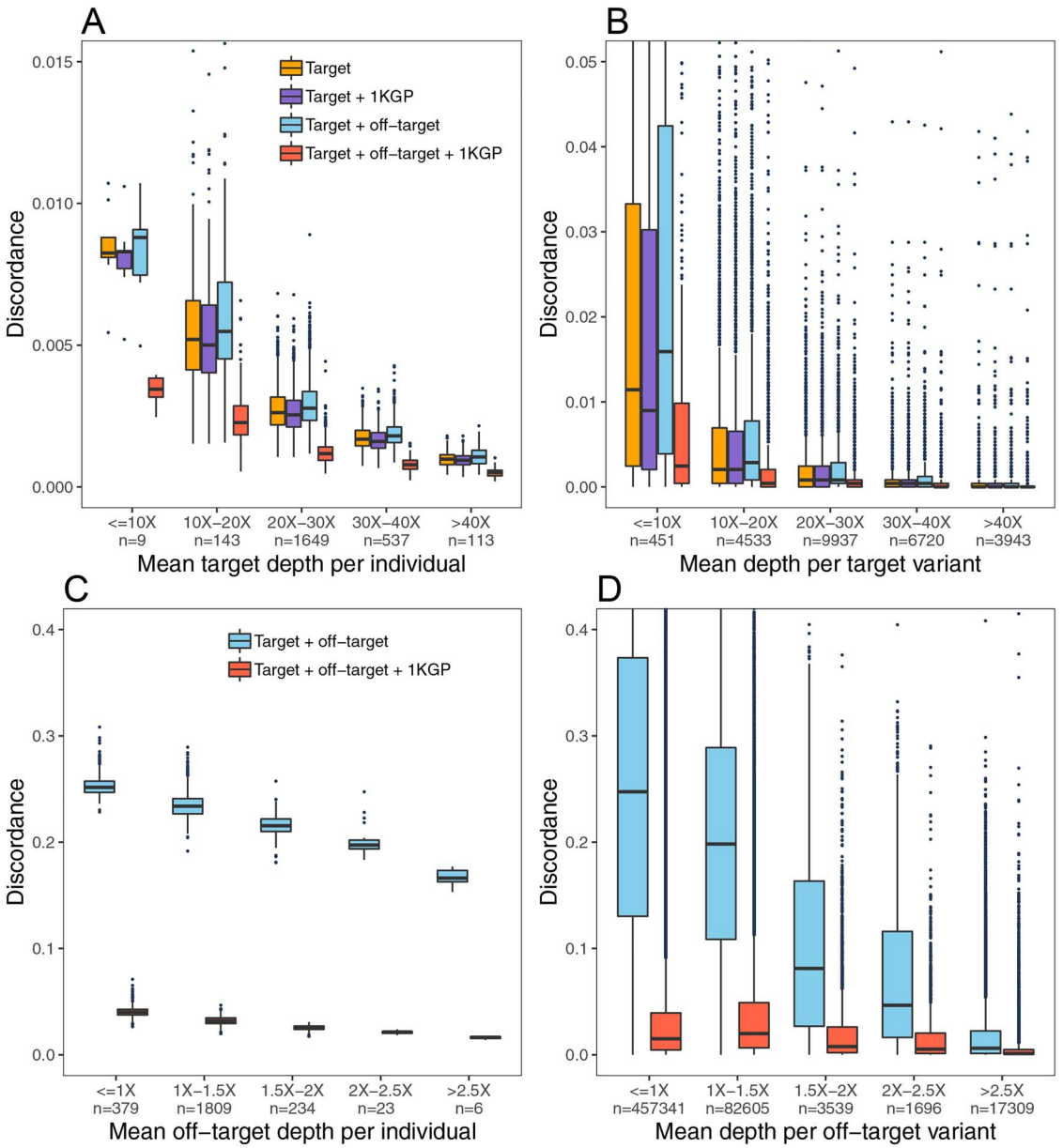


**Figure 2**. Genotype discordance rate between WES and array data as a function of sequencing depth. (**A**) Discordance per individual averaged across 25 349 SNPs in the target regions. (**B**) Discordance per targeted SNP averaged across 2451 individuals. (**C**) Discordance per individual averaged across 308 455 SNPs in the off-target regions. (**D**) Discordance per off-target SNP averaged across 2451 individuals. Numbers of individuals or SNPs per sequencing depth bin are indicated along the x-axis.

recalculating the adjusted $R^2$ for both (a) and (b) to obtain the distribution of difference in $R^2$.

## Results

### Evaluation of genotyping accuracy

Different from traditional genotype calling for WES data, our WEScall pipeline makes utilization of LD information contributed by off-target data and an external reference panel such as the 1KGP. In our Living Biobank WES dataset, the mean sequencing depth was ~32× and ~1.2× across target and off-target regions, respectively. In particular, the off-target sequencing depth was higher near the target regions but stabilized at ~0.9× across regions >300 bp away from the target regions (Supplementary Data, Figure S1). As shown in Figure 2 and Supplementary Data, Table S1, inclusion of both off-target data and the 1KGP reference panel improved the genotyping accuracy for target regions compared to the other three call sets. The overall genotype discordance rate was reduced from 0.26% ($SE = 6.4 \times 10^{-6}$) for call set 1, which did not use off-target data or any external reference data, to 0.12% ($SE = 4.4 \times 10^{-6}$) for call set 4, which used both off-target data and the 1KGP reference panel. The improvement was more evident for samples with low target sequencing depth ≤10×, whose genotype discordance rate was reduced from 0.84% ($SE = 1.9 \times 10^{-4}$) for call set 1 to 0.34% ($SE = 1.2 \times 10^{-4}$) for call set 4. For samples sequenced at >40×, the improvement was smaller, reducing genotype discordance rate by 2-fold from 0.10% ($SE = 1.9 \times 10^{-5}$) for call set 1 to 0.05% ($SE = 1.3 \times 10^{-5}$) for call set 4. It is worth noting that the discordance rates, however, were similar for call sets

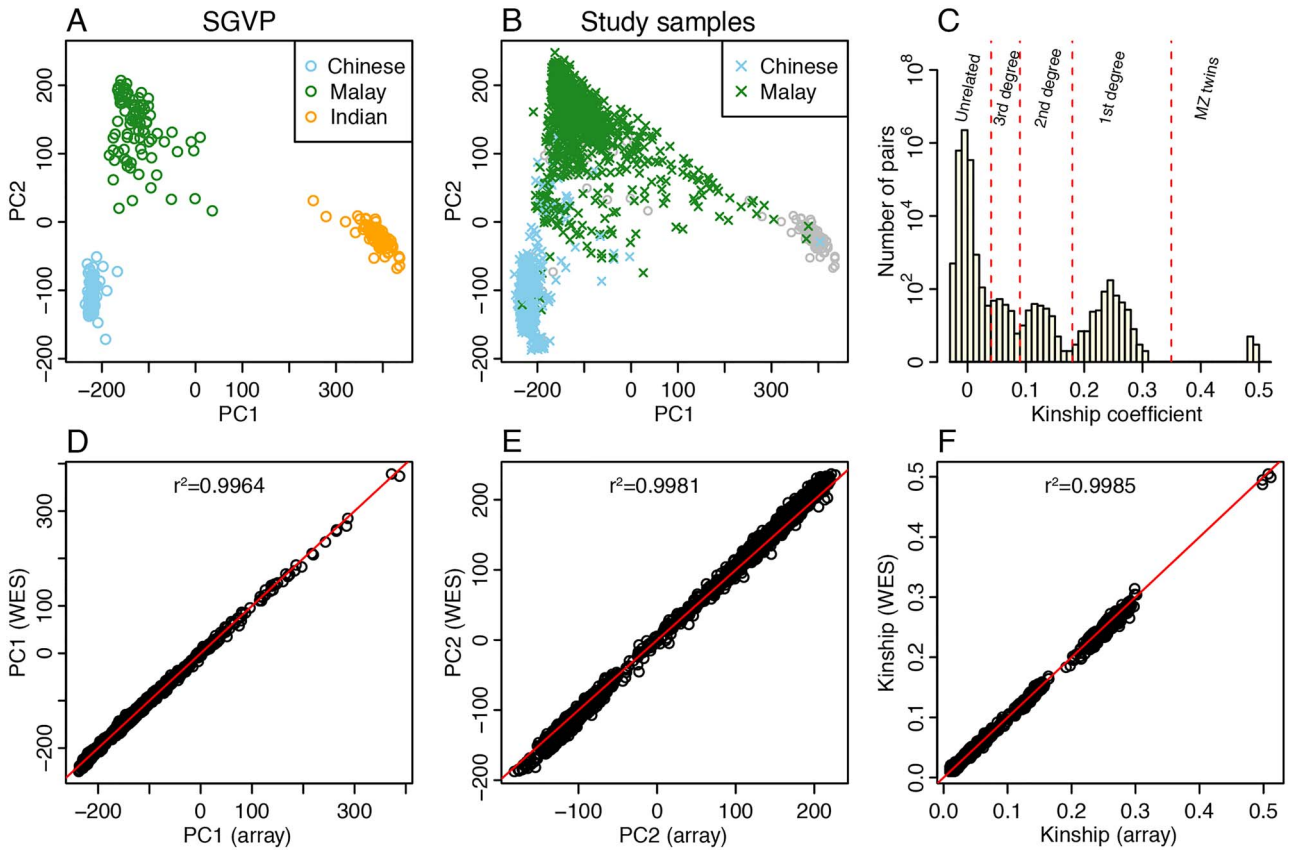


**Figure 3**. Estimation of ancestry and kinship coefficients based on SNPs from both target and off-target regions. **(A)** Reference ancestry space derived from PCA on the genotypes of Chinese, Malays and Indians from the SGVP panel. **(B)** Projection of WES samples into the SGVP reference space based on 733 494 SNPs overlapping with SGVP. WES samples are coloured by self-reported ethnicity, and the SGVP reference individuals are represented by grey symbols. **(C)** Cryptic relatedness among 2527 WES samples. We classified two individuals as the kth-degree relatives if their estimated kinship coefficient fell between $2^{-k-3/2}$ and $2^{-k-1/2}$. **(D–F)** Comparison of estimated PC1, PC2 and kinship coefficients derived from array genotypes and WES-based genotypes. In each panel, $r^2$ denotes squared Pearson's correlation between the estimates derived from array data and from WES data. In panel F, due to the enormously large number of sample pairs, we only included those with WES-based kinship estimates >0.01 in the figure and the calculation of $r^2$.

1–3. Similar patterns were observed when we evaluated genotype discordance rates at the SNP level across different sequencing depths (Figure 2B; Supplementary Data, Table S1).

Next, we evaluated the genotyping accuracy for off-target regions in call sets 3 and 4. Genotype discordance rates were calculated based on 562 490 off-target SNPs shared by both call sets and the array genotyping dataset. Most samples had off-target sequencing depth between 1× and 1.5×. For these samples, the off-target genotype discordance rate was as high as 23.44% ($SE = 1.3 \times 10^{-5}$) without using an external reference panel and was reduced to 3.21% ($SE = 5.5 \times 10^{-6}$) when we used the 1KGP data to help with phasing (Figure 2C; Supplementary Data, Table S2). Comparing across different sequencing depths in call set 4, the discordance rate ranged from 4.08% ($SE = 1.4 \times 10^{-5}$) for samples with <1.0× off-target data to 1.62% ($SE = 6.9 \times 10^{-5}$) for those with >2.5× off-target data. At variant level, most variants were covered by <1.0× sequencing data and had discordance rate at 3.27% ($SE = 5.3 \times 10^{-6}$) in call set 4 (Figure 2D). For variants sequenced to 2.0–2.5×, the discordance rate was reduced to 1.88% ($SE = 6.7 \times 10^{-5}$).

Finally, we performed a series of QCs to filter low-quality variants from call set 4 (Supplementary Data, Figure S2). These filters reduced the total number of SNPs from 1 470 782 to 1 093 210 in the target regions, and from 11 767 985 to 5 200 391 in the off-target regions. After QC, the genotype discordance rate was decreased from 0.12% ($SE = 4.4 \times 10^{-6}$) to 0.08% ($SE = 3.6 \times 10^{-6}$) for target SNPs and from 3.26% ($SE = 4.8 \times 10^{-6}$) to 0.70% ($SE = 3.0 \times 10^{-6}$) for off-target SNPs (Supplementary Data, Figure S3, Tables S3 and S4).

## Population structure and cryptic relatedness

We further evaluated the quality of genotypes derived from our WEScall pipeline (call set 4) in the inference of population structure and cryptic relatedness. As shown previously [22, 38], we constructed a Singapore ancestry map by applying PCA on the SGVP dataset (Figure 3A). We projected all our 2527 WES samples onto the SGVP map, confirming that most of our samples were Chinese and Malays with the exception of a few clustering with Indians and some potentially admixed samples (Figure 3B) [21]. After accounting for the population structure and admixture, we identified six pairs of monozygotic twins, 462 pairs of first-degree relatives, 161 pairs of second-degree relatives and 165 pairs of third-degree relatives based on the estimated kinship coefficients (Figure 3C). The estimated PCs and kinship coefficients showed remarkable concordance with those derived from array genotyping data of 2451 individuals (Figure 3D–F, squared Pearson's correlation $r^2 = 0.9964$ for PC1, 0.9981 for PC2 and 0.9985 for kinship coefficients), confirming the high quality of our call set.

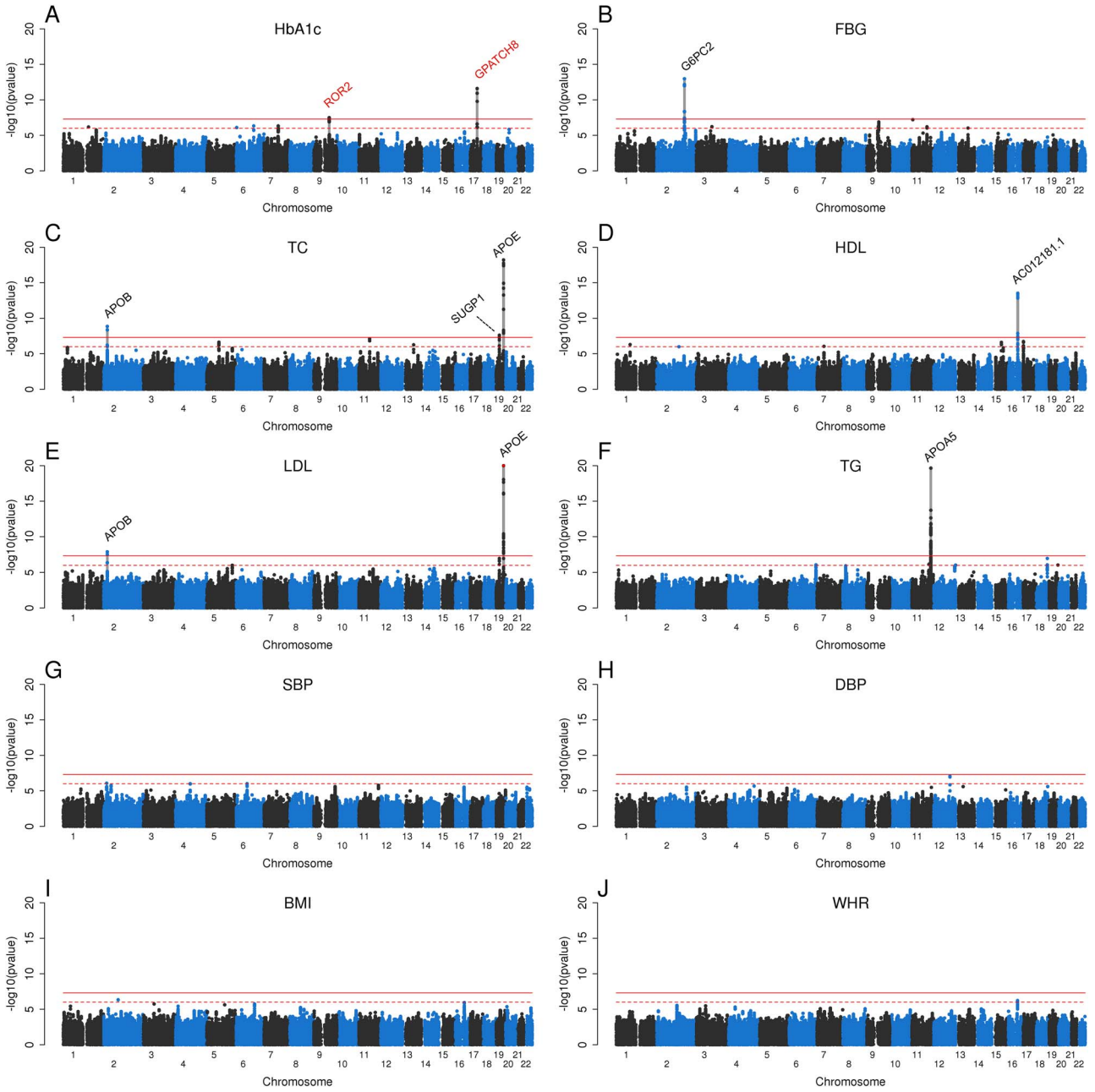


**Figure 4**. Manhattan plots of single-variant association analysis *P*-values for 10 metabolic traits. Red solid line: *P*-value threshold of $5 \times 10^{-8}$; red dotted line: *P*-value threshold of $10^{-6}$. The genome-wide significant loci are highlighted by the grey bar. Novel loci are labelled by red text. For the *APOE* locus in (E), *P*-values $<10^{-20}$ were set to $10^{-20}$ and were highlighted in red.

## Off-target data help identify genetic association signals

We performed GWAS of 10 metabolic traits using our final call set of 6 352 105 autosomal variants. The sample size ranges from 1963 to 2398 across 10 traits (Supplementary Data, Figure S4, Table S7). We controlled for both population structure and cryptic relatedness using our estimated ancestry PCs and kinship coefficients under a linear mixed model. The values of the genomic inflation factor ($\lambda_{GC}$) were between 0.989 and 1.043 for all 10 traits (Supplementary Data, Figure S5). Glycaemic traits (HbA1c and FBG) and lipid traits (TC, HDL, LDL and TG) showed strong association signals deviating from the null distribution, while SBP, DBP, BMI and WHR did not have evident signals (Figure 4; Supplementary Data, Figure S5).

In total, we identified 143 significant associations for glucose and lipid traits, which were further collapsed into 10 loci (Figure 4; Table 1). Eight of these loci were previously known to associate with corresponding traits (i.e. <500 kb from reported signals in the GWAS Catalog) [41], all of which can be identified by the analysis of either target or off-target variants. In our samples, *APOE* [MIM: 107741] and *APOB* [MIM: 107730], which were well-known LDL-associated genes, were also the top genes associated with TC in same directions (Table 1), consistent with the positive genetic correlation between TC and LDL identified by GWAS summary statistics [45]. We also identified two novel loci associated with HbA1c at the genome-wide significance level, which were

**Table 1.** Loci significantly associated with metabolic traits ($P < 5 \times 10^{-8}$)

| Trait | Locus | Lead variant | Position | Alleles (Ref/Alt) | $AF_{Chinese}$ (Alt) | $AF_{Malay}$ (Alt) | Beta (Alt) | SE | P | Known | Target or off-target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HbA1c | *GPATCH8-SLC4A1* | rs369762319 | 17:42477360 | T/C | 0.000 | 0.014 | -0.409 | 0.058 | $2.56 \times 10^{-12}$ | NO | Target |
| HbA1c | *ROR2* | rs1201042 | 9:94453525 | A/T | 0.776 | 0.721 | -0.059 | 0.011 | $3.24 \times 10^{-8}$ | NO | Off-target |
| FBG | *G6PC2* | rs2232326 | 2:169764491 | T/C | 0.046 | 0.121 | -0.188 | 0.025 | $1.07 \times 10^{-13}$ | YES | Both |
| TC | *APOB* | rs13306194 | 2:21252534 | G/A | 0.140 | 0.105 | -0.251 | 0.041 | $1.38 \times 10^{-9}$ | YES | Both |
| TC | *SUGP1* | rs10401969 | 19:19407718 | T/C | 0.096 | 0.113 | -0.247 | 0.044 | $2.42 \times 10^{-8}$ | YES | Both |
| TC | *APOE* | rs7412 | 19:45412079 | C/T | 0.081 | 0.087 | -0.424 | 0.047 | $6.11 \times 10^{-19}$ | YES | Both |
| HDL | *AC012181.1* | rs12149545 | 16:56993161 | G/A | 0.158 | 0.130 | 0.010 | 0.013 | $2.94 \times 10^{-14}$ | YES | Both |
| LDL | *APOE* | rs7412 | 19:45412079 | C/T | 0.081 | 0.087 | -0.595 | 0.042 | $1.15 \times 10^{-44}$ | YES | Both |
| LDL | *APOB* | rs13306194 | 2:21252534 | G/A | 0.140 | 0.105 | -0.215 | 0.038 | $1.35 \times 10^{-8}$ | YES | Both |
| TG | APOA5 | rs662799 | 11:116663707 | G/A | 0.723 | 0.714 | -0.166 | 0.018 | $2.17 \times 10^{-20}$ | YES | Both |

Each locus is labelled by the nearest gene, and the lead variant is the one with the smallest *P*-value. For each lead variant, we reported genomic position (in the format of chromosome:position), reference allele (Ref) and alternative allele (Alt), frequencies of the alternative allele (AF) in Chinese and Malay, and the effect size (beta) and its standard error (SE) of the alternative allele. We consider a locus as known if its lead variant is within 500 kb of reported genes in the GWAS Catalog for the corresponding trait. The last column indicates where the variants with $P < 5 \times 10^{-8}$ in each locus come from. HbA1c, glycated haemoglobin; FBG, fasting blood glucose; TC, total cholesterol; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TG, triglycerides.

*GPATCH8-SLC4A1* [MIM: 614396, 109270] (rs369762319; $P = 2.56 \times 10^{-12}$) and *ROR2* [MIM: 602337] (rs1201042; $P = 3.24 \times 10^{-8}$) (Figure 5). All three significant associations in the *GPATCH8-SLC4A1* locus were from target variants, including a 27-bp deletion in the *SLC4A1* gene (rs769664228, $P = 1.19 \times 10^{-11}$). In contrast, the *ROR2* locus reached genome-wide significance only at variants from off-target regions. In addition, we identified 157 suggestive associations in 28 loci across 10 traits ($5 \times 10^{-8} < P < 10^{-6}$, Supplementary Data, Table S5). These suggestive loci included four known ones reported by the GWAS Catalog, including *MTNR1B* [MIM: 600804] associated with FBG [46], *HMGCR* [MIM: 142910] associated with TC [47], *LIPC* [MIM: 151670] associated with HDL [48], and *SUGP1* [MIM: 607992] associated with LDL [49]. Except for *SUGP1*, which was identified by both target and off-target variants, the other three known loci together with 20 novel suggestive loci were identified by off-target SNPs rather than target variants.

### Off-target data improve polygenic risk prediction

Finally, we compared the prediction accuracy of PRS before and after incorporating off-target SNPs for nine traits, of which the summary statistics are available from UKB [30] and BBJ [31]. Despite the population difference between the reference samples (British in UKB and Japanese in BBJ) and the target samples (Chinese and Malays), which might compromise the prediction accuracy, PRS derived from either target SNPs or both target and off-target SNPs had adjusted $R^2$ significantly >0 for all traits except for SBP and DBP (Table 2). PRS for lipid traits had the highest prediction accuracy, consistent with a strong role of genetic regulation in lipid metabolism (Figure 4) [50]. Only a small proportion, ranging from 1.8 to 3.7% for different traits, of the GWAS SNPs with $P < 10^{-6}$ were in the target regions (Supplementary Data, Table S6). By incorporating off-target SNPs, the proportion of GWAS SNPs included in the PRS calculation were increased by an order of magnitude to 42.3–58.5% (Supplementary Data, Table S6). Consequently, the adjusted $R^2$ values of PRS were increased for all nine traits, six of which were significant with $P < 0.01$ while the other three were insignificant with $P > 0.05$ (HDL, SBP and DBP). For example, the adjusted $R^2$ increased from 6.06 to 8.10% for LDL ($P = 0.0046$), from 5.19 to 8.21% for TC ($P = 0.0004$), and from 0.32 to 1.67% for HbA1c ($P = 0.0015$). The PRS prediction accuracy for BMI was similar to those for HbA1c and FBG ($R^2 \approx 1.7\%$ using both target and off-target SNPs) even though we did not identify any significant association for BMI in our dataset, reflecting the polygenic genetic architecture of BMI.

## Discussion

WES has been widely used by researchers and clinicians all over the world. More than 60 000 exomes from over 20 studies have been contributed to the Exome Aggregation Consortium Browser [51]. Most studies, however, discarded the off-target data, even though it has been shown that extremely low-coverage data can be used to increase the power of genetic association tests [17]. Such an analysis strategy is highly inefficient considering that off-target data typically account for a substantial amount of the total sequencing data [17–20]. The low utilization of off-target data is partly due to the computational complexity and the lack of a user-friendly pipeline customized for analysing off-target data from WES. In this study, we filled in this gap by developing the WEScall pipeline as a general tool to perform LD-based genotype calling using both target and off-target data from WES with additional information from an external reference panel. Furthermore, our pipeline not only considers known variants that are in the reference panel, such as the 1KGP dataset, but also considers variants newly discovered in the target regions, which might be particularly important for clinical diagnosis. We do not consider novel variants in the off-target regions due to high error rates in variant discovery from shallow sequencing data.

We performed extensive evaluation of different analysis strategies using the Living Biobank WES dataset. We showed that with appropriate QC filtering, we could obtain genotypes at 0.993 concordance rate from 5.2 million SNPs in the off-target region, and thus enable genome-wide downstream analyses. With more WGS data from diverse populations become available as the external reference panel, we expect our WEScall pipeline to identify more variants with higher genotyping accuracy. In contrast to previous studies [17, 25], we also demonstrated that off-target data can be used to improve genotyping accuracy for SNPs in the high-coverage target regions. The improvement is noticeable only when we incorporate both off-target data and an external reference panel to obtain reliable LD information (Figure 2; Supplementary Data, Table S2). Even though the improvement gradually diminishes as the sequencing depth increases, this analysis strategy remains important to reduce the variation in genotyping accuracy across SNPs because the

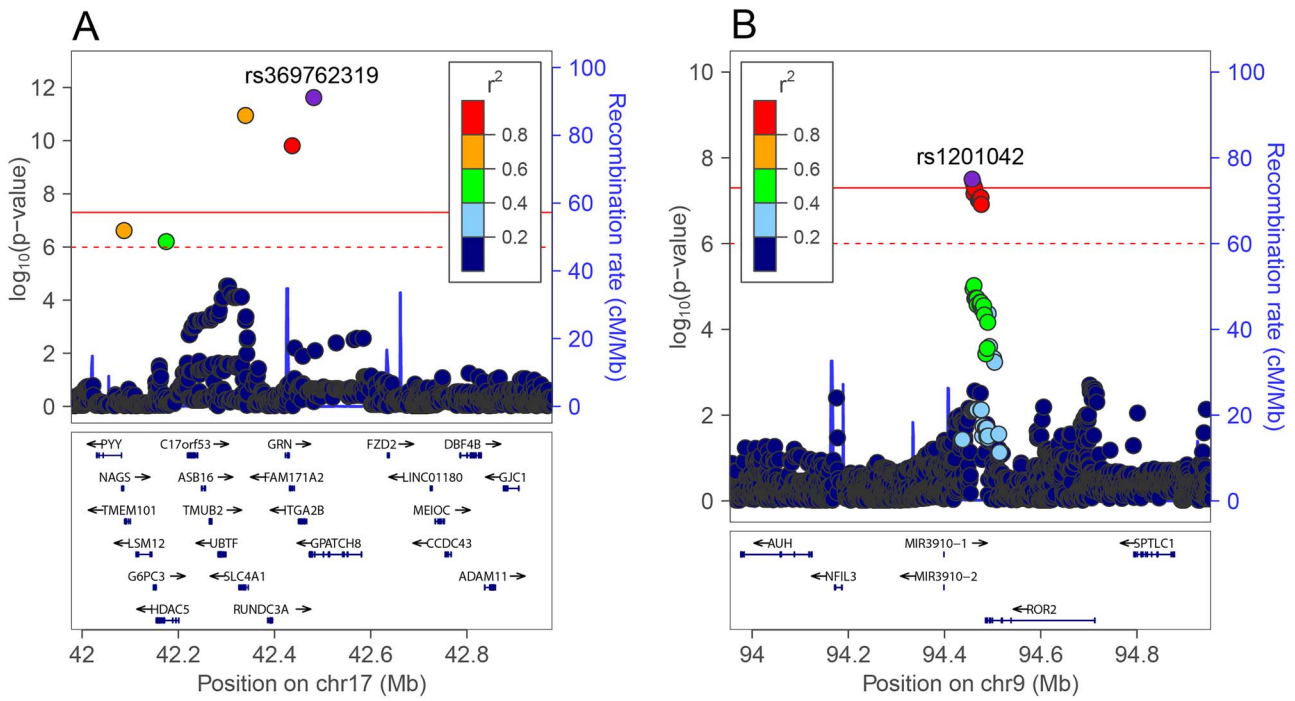


**Figure 5.** LocusZoom plots for two significant loci in association with HbA1c. (**A**) *GPATCH8-SLC4A1*; (**B**) *ROR2*. In each panel, the top variant is represented by a purple dot, while other variants are coloured according to LD with the top variant. Red solid line: $P = 5 \times 10^{-8}$; red dotted line: $P = 10^{-6}$.

**Table 2.** Comparison of polygenic risk prediction for metabolic traits with and without off-target data

| Trait | UKB sample size | BBJ sample size | (a) Target only | | | (b) Target and off-target | | | $P\ (H_0 : R_a^2 = R_b^2)$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | SNPs in PRS | $R_a^2$ | $P\ (H_0 : R_a^2 = 0)$ | SNPs in PRS | $R_b^2$ | $P\ (H_0 : R_b^2 = 0)$ | |
| HbA1c | 344 182 | 42 790 | 215 | 0.32% | $7.54 \times 10^{-3}$ | 556 | 1.67% | $6.29 \times 10^{-10}$ | 0.0015 |
| FBG | 314 916 | — | 74 | 0.35% | $6.24 \times 10^{-3}$ | 162 | 1.64% | $2.12 \times 10^{-9}$ | 0.0056 |
| TC | 344 278 | 128 305 | 437 | 5.19% | $1.42 \times 10^{-30}$ | 934 | 8.21% | $2.89 \times 10^{-48}$ | 0.0004 |
| HDL | 315 133 | 70 657 | 539 | 3.32% | $2.26 \times 10^{-21}$ | 1276 | 3.77% | $3.80 \times 10^{-24}$ | 0.1303 |
| LDL | 343 621 | 72 866 | 348 | 6.06% | $1.37 \times 10^{-35}$ | 792 | 8.10% | $1.59 \times 10^{-47}$ | 0.0046 |
| TG | 343 992 | 105 597 | 403 | 3.06% | $1.01 \times 10^{-15}$ | 965 | 4.71% | $1.49 \times 10^{-23}$ | 0.0010 |
| SBP | 340 162 | 136 597 | 96 | 0.01% | 0.605 | 378 | 0.10% | 0.093 | 0.1615 |
| DBP | 340 159 | 136 615 | 109 | 0.01% | 0.544 | 357 | 0.06% | 0.202 | 0.2371 |
| BMI | 359 983 | 158 284 | 285 | 0.51% | $3.57 \times 10^{-4}$ | 1042 | 1.74% | $3.63 \times 10^{-11}$ | 0.0081 |

We calculated the adjusted $R^2$ of PRS for each metabolic trait, adjusting for PC1, PC2, age, age$^2$ and sex. The PRS was calculated using SNPs within target regions (a) or from both target and off-target regions (b). Summary statistics from the meta-analysis of UKB and BBJ and a LD reference panel based on 1KGP Europeans were used for the PRS calculation. The adjusted $R^2$ and the corresponding P-value for the null hypothesis of $R^2 = 0$ were reported for both (a) and (b). We performed 10 000 bootstraps to obtain the P-value for the comparison of $R^2$ from (a) and (b).

sequencing depth fluctuates dramatically across the target regions in WES due to the imperfect capture technology [52]. While genotyping quality for SNPs from the off-target regions is lower than those from the target regions, we note that the 0.70% genotyping error rate after QC for the off-target SNPs in our Living Biobank dataset is even lower than the 1KGP-based imputation error rates for GWAS data from most populations, including Europeans [29], suggesting a good quality for downstream analyses based on off-target SNPs. These results also demonstrate that imputation algorithms such as BEAGLE [23] are useful to improve genotyping quality even at extremely shallow sequencing setting with only ~1× sequencing depth.

We illustrated the applicability of off-target genotypes produced by WEScall in inferring population structure and cryptic relatedness, testing for genetic associations, and predicting the PRS of complex traits. In our previous work, we have developed LASER and SEEKIN to estimate population structure and cryptic relatedness directly from low-coverage data without calling genotypes [18, 21, 22]. Here, we directly used the genotypes produced by WEScall without modelling the genotype uncertainty. Our estimates of population structure (top two PCs) and the kinship coefficients were almost identical to those obtained from high-quality array genotypes, further confirming the genotyping quality at off-target SNPs.

In the association analysis with 10 metabolic traits, eight out of the 10 significant loci were well-known association loci for the corresponding traits, such as the Apolipoproteins A5, B and E for lipid traits. We also identified two novel loci associated with HbA1c that reached genome-wide significance level in our samples but have not been reported in the GWAS Catalog. The first one is the *GPATCH8-SLC4A1* locus, composing of three significant variants locating from 42.3 to 42.5 Mb on chromosome 17. In particular, the variant rs769664228 ($P = 1.19 \times 10^{-11}$) is a 27-bp deletion in the *SLC4A1* gene, which is in high LD with the index SNP (rs369762319, a synonymous variant in *GPATCH8*) and is known to cause Southeast Asian Ovalocytosis [MIM: 166900], a

rare hereditary red cell membrane defect prevalent in Southeast Asia [53]. In our samples, this 27-bp deletion is mostly found in Malays (MAF: 1.44% in Malays versus 0.1% in Chinese), among whom the association signal with HbA1c was highly significant ($P = 4.36 \times 10^{-15}$ in Malays). In fact, we have replicated the association between this 27-bp deletion and HbA1c in an independent cohort of Singapore Malays (unpublished data). This locus has also been reported to associate with various red cell and platelet traits [54]. We thus hypothesize that mutations at this locus might affect the level of HbA1c by altering erythrocyte physiology [55]. The second locus composed of three significant SNPs near *ROR2*, which is responsible for neurological disorders such as Alzheimer's disease [MIM: 104300] and schizophrenia [MIM: 181500] [56–58], and has been observed to associate with diabetes [59–61]. In addition, *ROR2* has been reported to weakly associate with HbA1c in Lebanese ($P = 4.05 \times 10^{-6}$) [62]. Our finding based on Chinese and Malays corroborates this association signal, making *ROR2* a good candidate for functional studies for HbA1c. While we also reported 20 novel suggestive loci in off-target regions, we should interpret the results with caution due to lower statistical stringency and small sample size. Nevertheless, some suggestive loci are biologically interesting. For example, *C6orf203* [MIM: 618583] ($P = 4.73 \times 10^{-7}$ with HbA1c) encodes an RNA-binding protein involved in mitochondrial protein synthesis [63]. Knock out of *C6orf203* in mouse model would lead to oxidative phosphorylation (OXPHOS) deficiency [MIM: 609060] [64, 65]. *CES1* [MIM: 114835] ($P = 5.91 \times 10^{-7}$ with WHR) encodes carboxylesterase 1, which is a key enzyme for endogenous esters hydrolysis and lipid homeostasis [66, 67]. Replication using independent high-quality data or validation by functional experiments will be required to confirm these findings. Here, our main goal was to illustrate that our pipeline could be useful for screening candidate association loci in non-coding regions by mining existing WES data without additional cost.

Finally, we demonstrated that for WES samples, using off-target data could significantly improve prediction accuracy of PRS for several complex traits, supporting the important role played by non-coding variants in regulating many biological processes. We used summary statistics derived from meta-analysis of GWASs based on UKB and BBJ, and calculated PRS for Chinese and Malays using a standard P + T method. This analysis strategy, however, is not meant to be optimal because trans-ethnic PRS prediction is currently under active research and should be interpreted with caution [68]. We expect the prediction accuracy to be further improved when summary statistics from large GWAS of diverse ancestry background are available. Because PRS are often used as the instrumental variable in MR analyses, the improved PRS accuracy can be translated into higher statistical power to infer causal relationship between risk factors (such as lipid and glycaemic traits) and complex diseases (such as coronary artery disease and cancer) [13–15]. In addition to quantitative traits, we also expect to obtain better PRS for complex diseases when using both target and off-target data from WES, which might be useful for clinical care [12]. Remarkably, a recent study has reported that polygenic risk can powerfully modify the risk conferred by monogenic risk variants in coronary artery disease, breast cancer [MIM: 114480], and colorectal cancer [MIM: 114500] [16]. Our method can therefore facilitate better risk prediction and stratification in a cost-effective WES experiment by assessing both monogenic risk variants, which are mostly protein coding within target regions, and polygenic risk contributed by genome-wide variants.

Several limitations of the proposed method should be noted. First, the performance of WEScall depends on the amount of off-target sequencing data. While several studies have shown substantial amount of off-target data from WES experiments [17–20], the off-target coverage varies across different experimental protocols and might decrease with the improvement of the capture technology. An interesting experimental design would be to combine exome enrichment with low-coverage WGS to balance between experimental cost and genomic coverage. Second, because LD-based calling relies on shared haplotypes, WEScall is expected to have less improvement for genotyping rare variants and will not be applicable to discover novel variants in the off-target regions. Because of the above two limitations, we do not expect to call genotypes for all variants in the off-target regions. Finally, further evaluation will be needed to assess the performance of WEScall in diverse populations, whose LD structure and allele frequency spectrum can be substantially different from Chinese and Malays.

## Key points

- We develop a bioinformatics pipeline for accurate genotype calling in both target and off-target regions of WES.
- We identify two novel loci in significant association with HbA1c, including a Malay-specific 27-bp deletion in SLC4A1.
- We demonstrate that off-target data from whole sequencing can significantly improve polygenic risk prediction.

## Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

## Funding

The Natural Science Foundation of China (NSFC 81973148), and the Biomedical Research Council (BMRC 03/1/27/18/216), National Medical Research Council (0838/2004), National Research Foundation (through BMRC 05/1/21/19/425 and 11/1/21/19/678) and the Ministry of Health, Singapore. Genotyping and WES for Living Biobank were jointly funded by the Agency for Science, Technology and Research, Singapore (https://www.a-star.edu.sg/) and Merck Sharp & Dohme Corp., Whitehouse Station, NJ, USA (http://www.merck.com). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Conflict of Interest

The authors declare no conflict of interests.

## Author contributions

C.W. and S.C. conceived and supervised the study. J.D. and D.W. implemented the software pipeline. J.D., D.W., L.D., K.W., M.J. and X.C. tested the pipeline and analysed data. D.F.R., E.S.T., J.L., X.S. and C.W. contributed data. J.D., X.S., S.C. and C.W. interpreted results and wrote the manuscript. All authors contributed to the revision of the manuscript and approved the final version.

## Web resources

WEScall, https://github.com/dwuab/WEScall

GotCloud (TopMed freeze 3 version), https://github.com/statgen/topmed_freeze3_calling

SEEKIN, https://github.com/chaolongwang/SEEKIN/

LASER, http://csg.sph.umich.edu/chaolong/LASER/

BEAGLE, https://faculty.washington.edu/browning/beagle/beagle.html

LocusZoom, http://locuszoom.org/

PRSice-2, http://www.prsice.info/

METAL, http://csg.sph.umich.edu/abecasis/Metal/

1000 Genomes Project, http://www.internationalgenome.org/

Singapore Genome Variation Project, http://phg.nus.edu.sg/#sgvp

GWAS Catalog, https://www.ebi.ac.uk/gwas/

GWAS results of UK Biobank, http://www.nealelab.is/uk-biobank

GWAS results of Biobank Japan, http://jenger.riken.jp/en/result

OMIM, http://www.omim.org/

## References


1. Visscher PM, Wray NR, Zhang Q, *et al*. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 2017;**101**:5–22.
2. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;**299**:1335–44.
3. Ng SB, Turner EH, Robertson PD, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;**461**:272–6.
4. Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.
5. Do R, Stitziel NO, Won HH, *et al*. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2015;**518**:102–6.
6. Lange LA, Hu Y, Zhang H, *et al*. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet* 2014;**94**:233–45.
7. Bernstein BE, Birney E, Dunham I, *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
8. The GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**:648–60.
9. Kundaje A, Meuleman W, Ernst J, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
10. Farh KK, Marson A, Zhu J, *et al*. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015;**518**:337–43.
11. Yang J, Benyamin B, McEvoy BP, *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;**42**:565–9.
12. Khera AV, Chaffin M, Aragam KG, *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–24.
13. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med* 2016;**35**:1880–906.
14. Guo Y, Warren Andersen S, Shu XO, *et al*. Genetically predicted body mass index and breast cancer risk: Mendelian randomization analyses of data from 145,000 women of European descent. *PLoS Med* 2016;**13**:e1002105.
15. Holmes MV, Asselbergs FW, Palmer TM, *et al*. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J* 2015;**36**:539–50.
16. Fahed AC, Wang M, Homburger JR, *et al*. Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer. *medRxiv* 2019. https://doi.org/10.1101/19013086.
17. Pasaniuc B, Rohland N, McLaren PJ, *et al*. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 2012;**44**:631–5.
18. Wang C, Zhan X, Bragg-Gresham J, *et al*. Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* 2014;**46**:409–15.
19. Zhan X, Larson DE, Wang C, *et al*. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat Genet* 2013;**45**:1375–9.
20. Taliun D, Chothani SP, Schönherr S, *et al*. LASER server: ancestry tracing with genotypes or sequence reads. *Bioinformatics* 2017;**33**:2056–8.
21. Wang C, Zhan X, Liang L, *et al*. Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *Am J Hum Genet* 2015;**96**:926–37.
22. Dou J, Sun B, Sim X, *et al*. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genet* 2017;**13**:e1007021.
23. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* 2009;**85**:847–61.
24. Li Y, Sidore C, Kang HM, *et al*. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* 2011;**21**:940–51.
25. Jun G, Wing MK, Abecasis GR, *et al*. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* 2015;**25**:918–25.
26. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
27. Browning BL, Browning SR. Genotype imputation with millions of reference samples. *Am J Hum Genet* 2016;**98**:116–26.
28. McCarthy S, Das S, Kretzschmar W, *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**:1279–83.
29. Wu D, Dou J, Chai X, *et al*. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 2019;**179**:736–49.
30. Sudlow C, Gallacher J, Allen N, *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
31. Nagai A, Hirata M, Kamatani Y, *et al*. Overview of the BioBank Japan project: study design and profile. *J Epidemiol* 2017;**27**:S2–s8.
32. Flickinger M, Jun G, Abecasis GR, *et al*. Correcting for sample contamination in genotype calling of DNA sequence data. *Am J Hum Genet* 2015;**97**:284–90.
33. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics* 2015;**31**:2202–4.

34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv* 2013;13033997. https://arxiv.org/abs/1303.3997v2.
35. Tarasov A, Vilella AJ, Cuppen E, *et al*. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;**31**:2032–4.
36. DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491.
37. Jun G, Flickinger M, Hetrick KN, *et al*. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 2012;**91**:839–48.
38. Teo YY, Sim X, Ong RT, *et al*. Singapore genome variation project: a haplotype map of three southeast Asian populations. *Genome Res* 2009;**19**:2154–62.
39. Manichaikul A, Mychaleckyj JC, Rich SS. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;**26**:2867–73.
40. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;**44**:821–4.
41. MacArthur J, Bowler E, Cerezo M, *et al*. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;**45**:D896–901.
42. Pruim RJ, Welch RP, Sanna S, *et al*. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;**26**:2336–7.
43. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;**26**:2190–1.
44. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. *GigaScience* 2019;**8**. doi: 10.1093/gigascience/giz082.
45. Bulik-Sullivan B, Finucane HK, Anttila V, *et al*. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015;**47**:1236.
46. Prokopenko I, Langenberg C, Florez JC, *et al*. Variants in MTNR1B influence fasting glucose levels. *Nat Genet* 2009;**41**:77–81.
47. Jiang SY, Li H, Tang JJ, *et al*. Discovery of a potent HMG-CoA reductase degrader that eliminates statin-induced reductase accumulation and lowers cholesterol. *Nat Commun* 2018;**9**:5138.
48. Guerra R, Wang J, Grundy SM. A hepatic lipase (LIPC) allele associated with high plasma concentrations of high density lipoprotein cholesterol. *Proc Natl Acad Sci U S A* 1997;**94**:4532–7.
49. Kim MJ, Yu CY, Theusch E, *et al*. SUGP1 is a novel regulator of cholesterol metabolism. *Hum Mol Genet* 2016;**25**:3106–16.
50. Willer CJ, Schmidt EM, Sengupta S, *et al*. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–83.
51. Lek M, Karczewski KJ, Minikel EV, *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285.
52. Wang Q, Shashikant CS, Jensen M, *et al*. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep* 2017;**7**:885.
53. Rosanas-Urgell A, Lin E, Manning L, *et al*. Reduced risk of plasmodium vivax malaria in Papua New Guinean children with southeast Asian ovalocytosis in two cohorts and a case-control study. *PLoS Med* 2012;**9**:e1001305.
54. Astle WJ, Elding H, Jiang T, *et al*. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 2016;**167**:1415–29.
55. Chen P, Ong RT-H, Tay W-T, *et al*. A study assessing the association of glycated hemoglobin A1C (HbA1C) associated variants with HbA1C, chronic kidney disease and diabetic retinopathy in populations of Asian ancestry. *PLoS One* 2013;**8**:e79767.
56. Cerpa W, Latorre-Esteves E, Barria A. RoR2 functions as a noncanonical Wnt receptor that regulates NMDAR-mediated synaptic transmission. *Proc Natl Acad Sci U S A* 2015;**112**:4797–802.
57. Green JL, Kuntz SG, Sternberg PW. ROR receptor tyrosine kinases: orphans no more. *Trends Cell Biol* 2008;**18**:536–44.
58. Oishi I, Suzuki H, Onishi N, *et al*. The receptor tyrosine kinase Ror2 is involved in non-canonical Wnt5a/JNK signalling pathway. *Genes Cells* 2003;**8**:645–54.
59. Calkin CV, Gardner DM, Ransom T, *et al*. The relationship between bipolar disorder and type 2 diabetes: more than just co-morbid disorders. *Ann Med* 2013;**45**:171–81.
60. Charles EF, Lambert CG, Kerner B. Bipolar disorder and diabetes mellitus: evidence for disease-modifying effects and treatment implications. *Int J Bipolar Disord* 2016;**4**:13.
61. Lustman PJ, Anderson RJ, Freedland KE, *et al*. Depression and poor glycemic control: a meta-analytic review of the literature. *Diabetes Care* 2000;**23**:934–42.
62. Ghassibe-Sabbagh M, Haber M, Salloum AK, *et al*. T2DM GWAS in the Lebanese population confirms the role of TCF7L2 and CDKAL1 in disease susceptibility. *Sci Rep* 2014;**4**:7351.
63. Gopalakrishna S, Pearce SF, Dinan AM, *et al*. C6orf203 is an RNA-binding protein involved in mitochondrial protein synthesis. *Nucleic Acids Res* 2019;**47**:9386–99.
64. Ketterer C, Müssig K, Heni M, *et al*. Genetic variation within the TRPM5 locus associates with prediabetic phenotypes in subjects at increased risk for type 2 diabetes. *Metabolism* 2011;**60**:1325–33.
65. Palacios-Zambrano S, Vázquez-Fonseca L, González-Páramos C, *et al*. C6orf203 controls OXPHOS function through modulation of mitochondrial protein biosynthesis. *bioRxiv* 2019;704403. doi: 10.1101/704403.
66. Lian J, Nelson R, Lehner R. Carboxylesterases in lipid metabolism: from mouse to human. *Protein Cell* 2018;**9**:178–95.
67. Wang D, Zou L, Jin Q, *et al*. Human carboxylesterases: a comprehensive review. *Acta Pharm Sin B* 2018;**8**:699–712.
68. Martin AR, Kanai M, Kamatani Y, *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**:584–91.