ForestQC: quality control on genetic variants from next-generation sequencing data using random forest

Jiajin Li[1], Brandon Jew[2], Lingyu Zhan[3], Sungoo Hwang[4], Giovanni Coppola[4], Nelson B. Freimer[1,4], Jae Hoon Sul[4,*]

1. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

2. Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, CA 90095, USA

3. Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

4. Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA 90095, USA

*. Correspondence: jaehoonsul@mednet.ucla.edu

ABSTRACT

Next-generation sequencing technology (NGS) enables discovery of nearly all genetic variants present in a genome. A subset of these variants, however, may have poor sequencing quality due to limitations in sequencing technology or in variant calling algorithms. In genetic studies that analyze a large number of sequenced individuals, it is critical to detect and remove those variants with poor quality as they may cause spurious findings. In this paper, we present a statistical approach for performing quality control on variants identified from NGS data by combining a traditional filtering approach and a machine learning approach. Our method uses information on sequencing quality such as sequencing depth, genotyping quality, and GC contents to predict whether a certain variant is likely to contain errors. To evaluate our method, we applied it to two whole-genome sequencing datasets where one dataset consists of related individuals from families while the other consists of unrelated individuals. Results indicate that our method outperforms widely used methods for performing quality control on variants such as VQSR of GATK by considerably improving the quality of variants to be included in the analysis. Our approach is also very efficient, and hence can be applied to large sequencing datasets. We conclude that combining a machine learning algorithm trained with sequencing quality information and the filtering approach is an effective approach to perform quality control on genetic variants from sequencing data.

Keywords: machine learning, genetic variant, quality control, next-generation sequencing, random forest, filtering

Author Summary

Genetic disorders can be caused by many types of genetic mutations, including common and rare single nucleotide variants, structural variants, insertions and deletions. Nowadays, next generation sequencing (NGS) technology allows us to identify various genetic variants that are associated with diseases. However, variants detected by NGS might have poor sequencing quality due to biases and errors in

sequencing technologies and analysis tools. Therefore, it is critical to remove variants with low quality, which could cause spurious findings in follow-up analyses. Previously, people applied either hard filters or machine learning models for variant quality control (QC), which failed to filter out those variants accurately. Here, we developed a statistical tool, ForestQC, for variant QC by combining a filtering approach and a machine learning approach. We applied ForestQC to one family-based whole genome sequencing (WGS) dataset and one general case-control WGS dataset, to evaluate our method. Results show that ForestQC outperforms widely used methods for variant QC by considerably improving the quality of variants. Also, ForestQC is very efficient and scalable to large-scale sequencing datasets. Our study indicates that combining filtering approaches and machine learning approaches enables effective variant QC.

Introduction

Over the past few years, genome-wide association studies (GWAS) have been playing an important role in identifying genetic variations associated with common diseases or complex traits(1,2). GWAS have found many associations between common variants and human diseases, such as schizophrenia(3), type 2 diabetes(4,5) and Parkinson's Disease(6). However, these common variants typically explain only a small fraction of heritability for the complex traits(7,8). Rare variants are another type of genetic variants that have been considered as an important risk factor for complex traits and common diseases(9–12). With the next generation sequencing (NGS) technology, geneticists may now gain insights into the roles of novel or rare variants. For instance, deep targeted sequencing was applied to discover rare variants associated with inflammatory bowel disease(13). Whole genome sequencing (WGS) has been used to identify rare variants associated with prostate cancer(14), and with whole exome sequencing, studies have also detected rare variants associated with LDL cholesterol(15) and autism(16).

70 NGS data are not, however, perfect, and the quality of variants detected by sequencing may be
71 adversely influenced by several factors. First, genome sequencing is known to have errors or biases(17–
72 21), which might cause inaccuracy in detecting variants. Second, sequence mappability of different
73 regions may not be uniform, but correlated with sequence-specific biological features, leading to
74 alignment biases. For instance, it is shown that introns have significantly lower mappability levels than
75 exons(22). Third, variant calling algorithms may be sources of errors as no algorithm is 100% accurate.
76 For example, GATK HaplotypeCaller and GATKUnifiedGenotyper(23), which are the widely used
77 variant callers, have sensitivity of about 96% and precision of about 98%(24). Additionally, different
78 variant callers may generate discordant calls on some variants(25), which indicates inaccuracy of those
79 calls, and in certain cases, different versions of even the same software may generate inconsistent calls.
80 All these factors may generate false positive sites or incorrect genotypes, which may then lead to false
81 positive associations in the follow-up association test. For example, Alzheimer's Disease Sequencing
82 Project reports that they found spurious associations in the case-control analysis where one of the causes
83 for the problem could be inconsistent variant calling processes for sequenced samples(26).
84 It is extremely important to perform quality control (QC) on genetic variants identified from
85 sequencing to remove variants that may contain sequencing errors and hence are likely to be false
86 positive calls. Traditionally, genetic studies have utilized two types of QC approaches; we call them,
87 "filtering" and "classification" approaches. In filtering approaches, several filters are applied to remove
88 problematic variants such as variants with high genotype missing rate (e.g. > 5%), low Hardy-Weinberg
89 Equilibrium (HWE) p-value (e.g. < 1E-4), or very high or low allele balance of heterozygous calls
90 (ABHet) (e.g. > 0.75 or < 0.25). One main problem with this type of approaches is that these thresholds
91 are arbitrarily determined without strong statistical justification. We may also remove variants whose
92 metrics are very close to the thresholds (e.g. variants with missing rate of 5.1%). Another type of QC is
93 a classification approach that attempts to learn variants with low quality using machine learning

94 approaches. One example is VQSR of GATK(24,27) that uses a Gaussian mixture model to learn the
95 multidimensional annotation profile of variants with high and low quality. However, one of issues with
96 VQSR is that one needs training datasets acquired from existing databases on variants such as 1000
97 Genomes Project(28) and HapMap(29), which may be biased to keep known variants and filter out novel
98 variants. Another issue is that those known databases of genetic variants may not be always accurate,
99 which would lead to inaccurate classification of variants, and they may not even be available for some
100 species. It may also be a challenge to apply VQSR to a variant call set generated by variant callers other
101 than GATK as VQSR needs metrics of variants that are not often calculated by non-GATK variant
102 callers.

103 In this article, we present ForestQC for performing QC on genetic variants discovered through
104 sequencing. Our method aims to identify whether a specific variant is of high sequencing quality
105 ("good" variants) or of low quality ("bad" variants) by combining the filtering and classification
106 approaches. We first apply a filtering approach to detect obviously good and bad variants from data. We
107 use stringent filters such that those variants are truly good or bad while the rest of variants that are
108 neither good nor bad are considered to have ambiguous quality ("gray" variants). Given this set of good
109 and bad variants, we train a machine learning model whose goal is to classify whether gray variants are
110 good or bad. With an insight that good variants would have higher genotype quality and sequencing
111 depth than do bad variants, we use information of several sequencing quality measures of variants for
112 model training. ForestQC then uses sequencing quality measures of gray variants to predict whether
113 each gray variant has high or low sequencing quality. Our approach is different from the filtering
114 strategy in that it only uses filters to identify truly good or bad variants and does not attempt to classify
115 gray variants with filters. Our method is also different from VQSR as our training strategy allows us to
116 train our model without known datasets for variants and solves several issues with VQSR mentioned

above. Another advantage of our software is that it can be applied to standard Variant Call Format (VCF) files from any variant callers and is very efficient.

To demonstrate accuracy of ForestQC, we apply it to two high-coverage WGS datasets; 1) large extended pedigrees ascertained for bipolar disorder (BP) from Costa Rica and Colombia(30), and 2) a sequencing study for Progressive Supranuclear Palsy (PSP). The first dataset includes 449 related individuals from families while the latter dataset consists of 495 unrelated individuals. We show that ForestQC outperforms VQSR and a filtering approach based on ABHet as good variants detected from ForestQC have higher sequencing quality than those from VQSR and the filtering approach in both datasets. This suggests that our tool identifies high-quality variants more accurately than other approaches in both family and unrelated datasets. ForestQC is publicly available at https://github.com/avallonking/ForestQC

Results

Overview of ForestQC

ForestQC takes a raw VCF file as input and determines whether each variant has “good” sequencing quality or “bad” quality. Our method combines a filtering approach that determines good and bad variants by a set of pre-defined filters and a classification approach that uses machine learning to classify whether a variant is good or bad. As illustrated in Figure 1, our method first calculates statistics of each variant for several filters that are commonly used in performing QC in GWAS. These statistics consist of ABHet, HWE p-value, genotype missing rate, Mendelian error rate for family data, and any user-defined statistics (details described in Method session). ForestQC then identifies three sets of variants using these statistics for filters: 1) a set of good variants that pass all filters, 2) a set of bad

140 variants that fail any filter(s), and 3) a set of gray variants that are neither good nor bad variants. We use
141 stringent thresholds for filters (Table S2, S3), and hence we are highly confident that good variants are
142 of high quality while bad variants are truly false positives or have unequivocally poor sequencing
143 quality. The next step in ForestQC is to train a random forest machine learning model using the good
144 and bad variants we detect from the filtering step. In ForestQC, seven sequencing quality metrics of
145 good and bad variants are used as features to train the random forest model, including three related to
146 sequencing depth, three related to genotype quality, and one related to the GC content. Finally, the fitted
147 model predicts whether each gray variant is good or bad. We combine the predicted good variants from
148 the random forest model and the good variants from the filtering step, and they are all good variants
149 determined by ForestQC. The same procedure is applied to identify bad variants.

150 One major challenge in classifying gray variants is to identify a set of sequencing quality metrics that
151 are used as features to train the random forest model. We choose three sets of features based on quality
152 metrics that variant callers provide and prior knowledge in genome sequencing. The first set of features
153 is genotype quality (GQ) where we have three metrics: mean, standard deviation (SD), and outlier ratio.
154 The outlier ratio is the proportion of samples whose GQ scores are lower than a particular threshold, and
155 it measures a fraction of individuals who are poorly sequenced at a mutation site. A good variant is
156 likely to have high mean, low SD, and low outlier ratio of GQ values. The second set of features is
157 sequencing depth (DP) as low depth often introduces sequencing biases and reduces variant calling
158 sensitivity(31). We also use the same three sets of metrics for DP as those for GQ: mean, SD, and outlier
159 ratio. The last set of features is related to genomic characteristics instead of sequencing quality, which is
160 GC content. High or low GC content may decrease the coverage of certain regions(32,33) and thus may
161 lower the quality of variant calling. Hence, the GC content of the DNA region containing a good variant
162 would not be too high or too low. Given these three sets of features, ForestQC learns how those features
163 determine good and bad variants and classifies gray variants according to rules that it learns.

164 ## Comparison of different machine learning algorithms

165 As there are many different machine learning algorithms available, we first seek to find the most
166 accurate and efficient algorithm for performing QC on NGS variant data. To ensure the quality of
167 training and prediction, we choose supervised learning algorithms rather than unsupervised algorithms.
168 Several major types of supervised algorithms are selected for comparison: random forest, logistic
169 regression, k nearest neighbors (KNN), Naive Bayes, quadratic discriminant analysis (QDA), AdaBoost,
170 artificial neural network (ANN), and single support vector machine (SVM). We use the BP WGS
171 dataset, which consists of large pedigrees from Costa Rica and Colombia, to compare the performance
172 of different algorithms. We use the aforementioned three sets of features related to sequencing quality
173 for all algorithms we test. We apply the filtering approach (Table S2, S3) to the BP data to identify
174 good, bad, and gray variants, and we choose 100,000 good and 100,000 bad variants randomly for
175 model training. We then choose another 100,000 good and 100,000 bad variants randomly from the rest
176 of variants for model testing. Each learning algorithm will be trained with the same training set and
177 tested with the same test set. We use 10-fold cross validation, area under the receiver operating
178 characteristic curve (AUC), and F1-score to estimate classification accuracy during model testing. F1-
179 score is the harmonic average of precision (positive predictive value) and recall (sensitivity). The closer
180 F1-score is to 1, the better the performance is. To assess the efficiency of each algorithm, we measure its
181 time cost during training and predicting. We use eight threads for algorithms that support parallelization.

182 **Table 1: Performance of eight different machine learning algorithms**

Machine learning algorithm	Time cost (sec)	F1-score for indel classification	F1-score for SNV classification
Random Forest	9.85	0.9428	0.9740
ANN	75.34	0.9400	0.9707
SVM	1253.48	0.9381	0.9704
AdaBoost	25.27	0.9270	0.9672
Logistic Regression	2.49	0.9074	0.9668
KNN	24.71	0.9200	0.9486
QDA	0.30	0.9006	0.9241

Naïve Bayes	0.18	0.8716	0.9012

Performance metrics, including F1-scores, total time cost of model fitting and prediction, are ranked by F1-score for SNV classification. Random forest, ANN, logistic regression and KNN are set to run with eight threads. "ANN": artificial neural network. "SVM": single support vector machine. "KNN": K-nearest neighbors classifier. "QDA": quadratic discriminant analysis.

Results show that random forest is the most accurate model in both SNV classification and indel classification with the highest F1-scores, accuracy and the largest AUC (Table 1, Table S1, Figure S1). Its time cost is only 9.85 seconds in model training and prediction (Table 1), which ranks as the fourth fastest algorithm. As random forest randomly divides the entire dataset into several subsets of the same size and constructs decision trees independently in each subset, it is highly scalable, and it has low error rates and high robustness with respect to noise(34). As for other machine learning algorithms, both SVM and ANN are highly accurate (both with F1-score of 0.97 and AUC > 0.985 in SNV classification) but they are not as efficient as random forest. ANN is the second slowest algorithm that is about 8x slower than random forest because it has to estimate many parameters. Especially, SVM is the slowest algorithm because of its inability to parallelize, which costs about 125x as much time as random forest (Table 1). This suggests that it may be computationally very expensive to use SVM in large-scale WGS datasets that have tens of millions of variants. Normally, a real dataset is at least 10 times larger than the dataset used here. For example, in the BP dataset, the training set has 2.20 million (M) SNVs and there are 2.73M gray SNVs for prediction. We find that random forest only spends 80.51 seconds for training and predicting, while ANN needs 489.63 seconds and SVM needs 14.74 hours. Therefore, random forest is much faster than ANN and SVM, although all three algorithms have similar performance in terms of AUC (Figure S1). In addition, there are even a larger number of variants in large-scale WGS projects such as NHLBI Trans-Omics for Precision Medicine (TOPMed) program that includes about 463M variants. Hence, it is more practical to use random forest when processing this very large datasets. Logistic regression, Naive Bayes and QDA are more efficient than random forest, but their predictions

207 are not as accurate as those of random forest. For example, Naive Bayes needs only 0.18 seconds for
208 training and prediction while its F1-score is the lowest among all algorithms (0.90 and 0.87 in SNV and
209 indel classification, respectively) (Table 1). This result demonstrates that random forest is both accurate
210 and efficient, and hence we use it as the machine learning algorithm in our approach. To further improve
211 the random forest algorithm, we test a different number of trees in the algorithm and we find that
212 random forest with 50 trees balances efficiency and accuracy (Figure S2). To identify good variants
213 from gray variants, we use the probability of each gray variant being a good variant calculated from
214 random forest, and we consider gray variants with the probability of being good variants > 50% as good
215 variants as this probability threshold achieves the highest F1-score (Figure S3).

216 Measuring performance of QC methods on WGS data

217 To evaluate the accuracy of ForestQC and other methods on WGS data, we apply them to two WGS
218 datasets and calculate several statistics. For a family-based dataset, we calculate Mendelian error rate
219 (ME) of each variant, which measures inconsistency in genotypes between parents and offspring.
220 Another statistic we measure is genotype discordance rate between microarray and sequencing if
221 individuals who are sequenced are also genotyped. In both WGS datasets we analyze, microarray data
222 are available. These two statistics are important indicators of quality of variants because good variants
223 would follow Mendelian inheritance patterns and their genotypes would be consistent between
224 microarray and sequencing. In addition to these statistics, we measure several other statistics that are
225 reported in sequencing studies such as the number of variants (SNVs and indels),
226 transitions/transversions (Ti/Tv) ratio, the number of multi-allelic variants, genotype missing rate. We
227 compute these QC-related statistics separately for SNVs and indels. We use these statistics to compare
228 the performance of ForestQC with that of three approaches. The first is one without performing any QC
229 (no QC). The second method is VQSR which is a classification approach that requires known truth sets
230 for model training, such as HapMap or 1000 genomes. We use recommended resources and parameter

231 settings to run VQSR as of 2018-04-04(35), but we also look at different settings. The third method is an
232 ABHet approach, which is a filtering approach that retains variants according to allele balance of
233 variants (see Methods).

234 Performance of ForestQC on family WGS data

235 We apply ForestQC to the BP WGS dataset that consists of 449 subjects with the average coverage of
236 36. There are 25.08M SNVs and 3.98M indels(30). The variant calling is performed with GATK-
237 HaplotypeCaller v3.5. This is an ideal dataset for assessing the performance of different QC methods
238 because this dataset contains individuals from families who are both sequenced and genotyped. This
239 study design allows us to calculate both ME rate and genotype discordance rate of variants between
240 WGS and microarray. For this dataset, we test ForestQC with two different filter settings, one using ME
241 rate as a filter and the other not using ME as a filter. The results of the former approach would filter out
242 bad variants based on ME rate, and hence ME rate of good variants would be very low. However, we
243 observe that both approaches have similar performance in terms of ME rate and other statistics (Table
244 S4, Figure S4, Figure S5), and hence we show results of only ForestQC using ME rate as a filter.

245 **Table 2: Variant-level quality metrics of good variants in the BP dataset processed by different**
246 **methods**

Metric	No QC	ABHet	VQSR	ForestQC
Total SNVs	25081636	22415368	24239357	22227503
Known SNVs	21165051	19665276	20675746	19361635
Known SNVs (%)	84.38%	87.73%	85.30%	87.11%
Total indels	3976710	2670647	3212886	2789037
Known indels	3094271	2188996	2758783	2237002
Known indels (%)	77.81%	81.97%	85.87%	80.21%
Multi-allelic SNVs	153836	26549	128894	77693
Multi-allelic SNVs (%)	0.61%	0.12%	0.53%	0.35%

247 Four methods are compared, including no QC applied, ABHet approach, VQSR and ForestQC. "Known" stands for variants
248 found in dbSNP. The version of dbSNP is 150.

249 Results show that ForestQC outperforms ABHet and VQSR in terms of the quality of good SNVs
250 while it detects fewer good SNVs than the other approaches (detailed variant-level metrics in Table S5).
251 ForestQC identifies 22.23M (88%) good SNVs, which is fewer than 22.42M (89%) and 24.24M (97%)
252 good SNVs from ABHet and VQSR, respectively (Table 2). However, ABHet has 3.57x and VQSR has
253 9.99x higher ME rate on good SNVs than ForestQC (Figure 2a), and ABHet has 1.50x (p-value < 2.2e-
254 16) and VQSR has 1.26x higher genotype discordance rate (p-value < 2.2e-16) on good SNVs than
255 ForestQC (Figure 2b). In addition, ABHet and VQSR have 81.48x and 97.72x higher genotype missing
256 rate on good SNVs than ForestQC, respectively (Figure 2c), but it is important to note that genotype
257 missing rate is used as a filter in ForestQC, which means SNVs with high genotype missing rate are
258 filtered out. We observe that VQSR and ABHet have 319 thousand (K) (1.32%) and 235K (1.05%) good
259 SNVs with very high genotype missing rate (>10%), respectively, and there are also 118K (0.49%,
260 VQSR) and 53K (0.24%, ABHet) good SNVs with very high ME rate (>15%) while ForestQC has none
261 of them due to its filtering approach. The better quality of good SNVs from ForestQC means that bad
262 SNVs detected from ForestQC would have lower quality, and results show that bad SNVs detected by
263 our method have higher genotype missing rate, higher ME rates and higher genotype discordance rate
264 than those of ABHet, and higher genotype missing rate than those of VQSR (Figure S6a, b, c). The no
265 QC method keeps the greatest number of good SNVs (25.08M), but they have the highest ME rate,
266 genotype missing rate, and genotype discordance rate as expected.

267 Next, we obtain several statistics of good SNVs commonly used in sequencing studies to evaluate the
268 performance of ForestQC. One such statistic is Ti/Tv ratio, which is expected to be around 2.0 over the
269 whole genome(36). If this ratio is smaller than 2.0, it means that there may be false positive variants in
270 the dataset. We compute Ti/Tv ratio for each individual across all good SNVs and look at the
271 distribution of those ratios across all individuals (sample-level statistics). We find that the mean Ti/Tv

272 ratio of good known SNVs (present in dbSNP) is around 2.0 for all four methods, which suggests that
273 they have similar accuracy on known SNVs in terms of Ti/Tv ratio (Figure S7a). However, results show
274 that the mean Ti/Tv ratio of good novel SNVs (not in dbSNP) from ForestQC is better than that of those
275 SNVs from other methods; the mean Ti/Tv ratio is 1.68 for ForestQC, which is closest to 2.0 among
276 other methods (1.41 for VQSR, 1.53 for ABHet, and 1.29 for No QC) (Figure 3a). Paired t-tests for the
277 difference in the mean Ti/Tv ratio between ForestQC and other methods are all significant (p-value <
278 2.2e-16 versus all other methods). This result suggests that novel SNVs predicted to be good by
279 ForestQC are more likely to be true positives than those SNVs from other QC methods. Another statistic
280 commonly used in sequencing studies is the percentage of multi-allelic SNVs, which are variants with
281 more than one alternative allele. Given this sample size (449), many of them are likely to be false
282 positives, and ForestQC has 33.96% and 42.62% smaller fraction of multi-allelic SNVs among good
283 SNVs than do VQSR and no QC methods while the ABHet approach has the smallest fraction of such
284 SNVs (Table 2). Note that ABHet values can only calculated for biallelic mutation sites, so ABHet does
285 not work properly for multi-allelic variants. It might mistakenly filter out many high quality multi-allelic
286 SNVs, so it has the fewest multi-allelic SNVs.

287 In addition to SNVs, we apply the four QC methods to indels. Similar to results of SNVs, ForestQC
288 identifies fewer good indels than does VQSR, but the quality of those indels from ForestQC is better
289 than that of good indels from ABHet and VQSR. Out of total 3.98M indels, ForestQC predicts 2.79M
290 indels (70%) to have good sequencing quality while VQSR and ABHet find 3.21M (81%) and 2.67M
291 (67%) good indels, respectively (Table 2). Good indels from VQSR and ABHet, however, have 8.54x
292 and 3.18x higher ME rate, and 22.25x and 25.28x higher genotype missing rate, than those from
293 ForestQC, respectively (Figure 2d, e). Bad indels identified by ForestQC have 2.25x and 1.32x higher
294 ME rate, and 1.48x and 2.36x higher genotype missing rate than those from VQSR and ABHet,
295 respectively (Figure S6d, e). Besides, we observe that there are 95K (2.97%, VQSR) and 86K (3.23%,

296 ABHet) good indels with very high genotype missing rate (>10%) and also 167K (5.21%, VQSR) and
297 44K (1.66%, ABHet) good indels with very high ME rate (>15%) while there are no such indels in
298 ForestQC. This result suggests that many good indels detected by ABHet or VQSR may be false
299 positives or indels with poor sequencing quality. One of the reasons why VQSR does not perform well
300 on indels could be the database it uses for training its machine learning model as VQSR considers all
301 indels found in the database (Mills gold standard call set(37) and 1000G Project(38)) to be true variants.
302 This leads VQSR to have a significantly higher proportion of known indels among good indels (86%),
303 compared with 80% from ForestQC and 82% from ABHet (Table 2). The poor performance of VQSR
304 on indels may be because not all indels in the database are true variants, or because even if they are true
305 indels, those indels would not necessarily have high sequencing quality in the sequencing dataset of
306 interest. Hence, this result demonstrates one of the limitations of using known databases for finding
307 good variants. It is also important to note that in general, indels have much higher ME rate (0.41% for
308 no QC) than that of SNVs (0.08% for no QC), which is expected given the greater difficulty of calling
309 indels.

310 Another major difference between ForestQC and the other approaches is the allele frequency of
311 variants after QC as ForestQC keeps a greater number of rare variants in its good variant set. Our
312 method has 1.77% and 1.64% higher proportion of rare SNVs, and 5.30% and 15.37% higher proportion
313 of rare indels than ABHet and VQSR do, respectively (Table S6). We also observe this phenomenon in
314 the variant-level and sample-level statistics for the number of SNVs. The variant-level statistics show
315 that the number of good SNVs detected by ForestQC is similar to those from ABHet (Table 2).
316 However, the sample-level statistics show that each individual on average carries fewer alternative
317 alleles of good SNVs from ForestQC (3.58M total SNVs) than those from VQSR and ABHet (3.99M
318 and 3.77M total SNVs, respectively) (Figure 3b, c, Figure S7b). We observe a similar phenomenon for
319 indels between ABHet and ForestQC (Table 2, Figure 3d, Figure S7c, d). This phenomenon could be

320 explained by the higher fraction of rare variants among good variants from ForestQC, as individuals
321 would carry fewer variants if there are a greater fraction of rare variants. One main reason why
322 ForestQC has the higher proportion of rare variants is that common variants have higher ME rate,
323 genotype discordance rate and genotype missing rate than do rare variants (Figure S8); because common
324 variants are more heterozygous, it is more difficult to accurately call them. This suggests that while a
325 majority of common variants may be true variants, some of them may not necessarily have high
326 sequencing quality, and hence their calls may not be accurate enough for downstream analyses.

327 ForestQC uses several filters to remove variants whose sequencing quality is poor while other two
328 approaches (VQSR and ABHet) do not use these filters, which might have artificially improved the
329 performance of ForestQC. Hence, to compare the performance of ForestQC with other approaches
330 without this potential bias due to the filtering step, we measure the performance metrics on only gray
331 variants as their sequencing quality is not determined by the filtering approach. From 2.73M gray SNVs
332 and 1.09M gray indels, ForestQC identifies 979K (35.83%) good SNVs and 532K (48.58%) good
333 indels, while ABHet approach detects 620K (22.70%) SNVs and 195K (17.80%) indels, and VQSR
334 selects 2.16M (79.18%) SNVs and 643K (58.76%) indels as good variants, respectively (Table S7). For
335 good SNVs from gray variants, ABHet and VQST have 2.75x and 22.67x higher ME rate than
336 ForestQC, respectively (Figure S9a), and ABHet and VQSR have 5.15x (p-value = 1.367e-14) and
337 3.86x (p-value = 1.926e-14) higher genotype discordance rate than ForestQC (Figure S9b). In addition,
338 ABHet and VQSR have 15.50x and 7.05x higher genotype missing rate on good SNVs than ForestQC,
339 respectively (Figure S9c). We observed similar results for indels (Figure S9d and S8e). Sample-level
340 metrics also show that ForestQC has better Ti/Tv ratio on known SNVs (mean Ti/Tv: 1.64, 1.85, 1.72,
341 1.88 for No QC, ABHet, VQSR, ForestQC, respectively), and novel SNVs (mean Ti/Tv: 1.14, 1.04,
342 1.21, 1.22 for No QC, ABHet, VQSR, ForestQC, respectively) than other methods (Figure S10d and
343 S9e). Paired t-tests for the difference in the mean Ti/Tv ratio of novel SNVs and known SNVs between

344 ForestQC and other methods are all significant (p-value < 0.05 versus all other methods). These results
345 show that even on those variants for whom we do not use the filtering approach, ForestQC has better
346 performance than ABHet and VQSR. These results further imply that if we use the same filtering
347 approach to all three approaches, our method will still outperform other approaches.

348 Performance of ForestQC on WGS data with unrelated individuals

349 To evaluate the performance of ForestQC on WGS datasets that contain only unrelated individuals, we
350 apply it to the PSP dataset that has 495 individuals who are whole-genome sequenced at average
351 coverage of 29, generating 33.27M SNVs and 5.09M indels. Among the 495 individuals who are
352 sequenced, 381 individuals (77%) of them are also genotyped with microarray, which enables us to
353 check the genotype discordance rate between WGS and microarray data. Because the PSP dataset
354 contains only unrelated individuals, we do not report ME rate. Similar to BP WGS data, we apply four
355 methods (ForestQC, VQSR, ABHet, and No QC) to the PSP dataset, although the parameter setting of
356 VQSR has slightly changed. As the PSP dataset is called with GATK v3.2, the StrandOddsRatio (SOR)
357 information from the VCF file is missing, which is recommended to use in VQSR, and hence this
358 annotation is excluded from VQSR. However, we find that SOR information has little impact on the
359 results of VQSR as we test VQSR without SOR information using the BP dataset and obtain similar
360 results with one using SOR information (Figure S11).

361 **Table 3: Variant-level quality metrics of good variants in the PSP dataset processed by four**
362 **different methods**

Metric	No QC	ABHet	VQSR	ForestQC
Total SNVs	33273111	29771182	31281620	29352329
Known SNVs	25960464	24142744	24910728	23514257
Known SNVs (%)	78.02%	81.09%	79.63%	80.11%
Total indels	5093443	3311136	3682319	3418242
Known indels	3679990	2532899	3012662	2567879

Known indels (%)	72.25%	76.50%	81.81%	75.12%
Multi-allelic SNVs	250418	6685	188180	146247
Multi-allelic SNVs (%)	0.75%	0.02%	0.60%	0.50%

363 Four methods are compared, including no QC applied, ABHet approach, VQSR and ForestQC. "Known" stands for variants
364 found in dbSNP. The version of dbSNP is 150.

365 Similar to the results of the BP dataset, ForestQC identifies good variants with higher quality
366 although it detects fewer good variants than other approaches (detailed variant-level metrics in Table
367 S8). ForestQC identifies 29.25M (88%) good SNVs, which is slightly fewer than 29.77M (89%) good
368 SNVs from ABHet but about 2 million fewer than 31.28M (94%) good SNVs from VQSR (Table 3).
369 However, good SNVs from ABHet and VQSR have 53.76x and 42.55x higher genotype missing rate
370 than those from ForestQC, respectively (Figure 4a), but it is important to note that missing rate is
371 included as a filter in ForestQC. In addition, there are 311K (0.99%, VQSR) and 331K (1.13%, ABHet)
372 good SNVs with very high genotype missing rate (>10%), while ForestQC removes all these SNVs. We
373 also observe that bad SNVs from ForestQC have 2.4x higher genotype missing rate than those from
374 ABHet, although bad SNVs from GATK have slightly higher missing rate than those from ForestQC
375 (Figure S12a). Good SNVs from ABHet and VQSR have 1.28x (p-value < 2.2e-16) and 1.29x higher
376 genotype discordance rate (p-value < 2.2e-16) than those from ForestQC, respectively (Figure 4b). As
377 for the genotype discordance rate of bad SNVs, both ABHet and VQSR have higher genotype
378 discordance rate than does ForestQC (Figure S12b), but this may be inaccurate because of the small
379 number of bad SNVs genotyped with microarray (10,130, 4,121, and 553 such SNVs for ForestQC,
380 ABHet, and VQSR, respectively). The variant-level and sample-level statistics also demonstrate the
381 better quality of good SNVs from ForestQC. Although all methods have mean Ti/Tv ratio of good
382 known SNVs above 2.0, the mean Ti/Tv ratio of good novel SNVs among all sequenced individuals is
383 1.65 for ForestQC, which is closer to 2.0 than other methods (1.27, 1.54, and 1.24 for VQSR, ABHet, no
384 QC, respectively). (Figure S13a, Figure 5a). Paired t-tests for the difference in the mean Ti/Tv ratio

385 between ForestQC and other methods are all significant (p-value < 2.2e-16 versus all other methods).
386 ForestQC has 16.67% and 33.33% smaller fraction of multi-allelic SNVs among good SNVs than do
387 VQSR and no QC methods, respectively, while the ABHet approach has the smallest proportion of such
388 SNVs (Table 3). ABHet has the smallest number of multi-allelic SNVs because it can only work
389 properly for biallelic SNVs where all subjects are either heterozygous or homozygous and therefore it
390 might remove many multi-allelic SNVs by mistakes. Lastly, consistent with the results of the BP dataset,
391 the sample-level statistics show that each individual on average carries fewer alternative alleles of good
392 SNVs from ForestQC than those from VQSR and ABHet (Figure 5b, c, Figure S13b). Rare SNVs in
393 good SNVs from ForestQC account for 1.70% and 1.32% higher proportion, compared with those from
394 ABHet and VQSR (Supplemental Table 5). This may be because rare SNVs have lower genotype
395 missing rate and genotype discordance rate than do common variants (Figure S14a, b).

396 For indels, our method predicts 3.42M indels (67% of total 5.09M indels) to be good variants, which
397 is slightly more than 3.31M (65%) good indels from ABHet and fewer than 3.68M (72%) good indels
398 from VQSR (Table 3). Because the PSP dataset lacks ME rate as it contains only unrelated individuals
399 and indels are not called in microarray, it is difficult to compare the performance of the QC methods on
400 indels. We find that good indels from ABHet and VQSR have 27.02x and 18.77x higher genotype
401 missing rate than those from our method, respectively (Figure 4c). Additionally, VQSR and ABHet have
402 107K (2.91%) and 131K (4.08%) good indels with high genotype missing rate (>10%), respectively
403 while ForestQC filters out all of these indels. Also, bad indels from ForestQC have 2.05x and 1.21x
404 higher genotype missing rate than those from ABHet and VQSR, respectively (Figure S12c). This,
405 however, may be biased comparison as ForestQC removes indels with high genotype missing rate in its
406 filtering step. Consistent with the results of SNVs, the sample-level statistics indicate that each
407 individual has fewer good indels from ForestQC than those from VQSR and ABHet (Figure 5d, Figure
408 S13c, d). Among good indels, ForestQC has 6% and 1% more novel indels than VQSR and ABHet,

409 respectively (Table 3). In terms of allele frequency, rare indels detected by ForestQC accounts for
410 12.35% and 3.49% larger proportions than those by VQSR and ABHet, respectively (Table S9). Similar
411 to the results of the BP dataset, we also observe that the missing rate of rare indels is lower than that of
412 common indels (Figure S14c).

413 Similar with the analysis of the BP dataset, we also compare the performance of ForestQC, ABHet
414 approach and VQSR only on gray variants in PSP dataset. From 3.95M gray SNVs and 1.60M gray
415 indels, ForestQC identifies 1.71M (43.33%) good SNVs and 719K (45.01%) good indels, while ABHet
416 approach detects 780K (19.74%) SNVs and 248K (15.51%) indels, and VQSR selects 2.75M (69.52%)
417 SNVs and 820K (51.34%) indels as good variants, respectively (Table S10). For good SNVs from gray
418 variants, ABHet and VQSR have 14.84x and 5.38x higher genotype missing rate than ForestQC,
419 respectively (Figure S15a). In addition, ABHet has 2.09x (p-value = 2.183e-11) and VQSR has 2.13x
420 higher genotype discordance rate (p-value = 1.584e-10) on than ForestQC (Figure S15b). For indels,
421 ABHet and VQSR have 9.39x and 3.61x higher genotype missing rate on good indels than ForestQC,
422 respectively (Figure S15c). Sample-level metrics also show that ForestQC has better Ti/Tv ratio on
423 known SNVs (mean Ti/Tv: 1.75, 1.87, 1.82, 1.96 for No QC, ABHet, VQSR and ForestQC,
424 respectively) and novel SNVs (mean Ti/Tv: 1.17, 1.03, 1.20, 1.39 for No QC, ABHet, VQSR and
425 ForestQC, respectively) than other methods (Figures S15d and S15e). Paired t-tests for the difference in
426 the mean Ti/Tv ratio of novel SNVs and known SNVs between ForestQC and other methods are all
427 significant (p-value < 2.2e-16 versus all other methods). Similar to results of the BP dataset, ForestQC
428 has higher accuracy in identifying good variants from gray variants, compared with ABHet approach
429 and VQSR.

430 Feature importance in random forest classifier

431 ForestQC uses several sequencing features in the random forest classifier to predict whether a variant
432 with undermined quality is good or bad. To understand which sequencing features are more important

433 indicators for quality of variants than other features, we analyze weight or importance of each feature
434 that the random forest classifier learns during its model training. We first find that GC-content has the
435 lowest importance in both BP and PSP datasets and also for both SNVs and indels (Figure S17). This
436 means that GC-content may not be as a strong indicator of quality of variants as other features related to
437 sequencing quality such as depth (DP) and genotype quality (GQ). Second, the results show that
438 classification results are not determined by one or two most important features as there is no feature with
439 much higher importance than other features except GC-content. This suggests that all sequencing
440 features except GC-content are important indicators for quality of variants and need to be included in
441 our model. We also check correlation among features and find that while certain pairs of features are
442 highly correlated, like outlier GQ and mean GQ, SD DP and mean DP, some features have low
443 correlation to other features, such as GC, suggesting that they may capture different information on
444 quality of genetic variants (Figure S19). Third, we observe that the same features have different
445 importance between the BP dataset and the PSP dataset. For example, for SNVs, an outlier ratio of GQ
446 feature has the highest importance for the PSP dataset while it has the third lowest importance for the BP
447 dataset (Figure S17a). Also, the importance of features varies between SNVs and indels. One example is
448 a SD of DP feature that has the highest importance for SNVs in the BP dataset, but it has the third lowest
449 importance for indels (Figure S17a, b). Therefore, these results suggest that each feature may have a
450 different contribution to classification results depending on sequencing data and types of genetic
451 variants.

452 Performance of VQSR with different settings

453 For SNVs, GATK recommends three SNV call sets for training its VQSR model; 1) SNVs found in
454 HapMap ("HapMap"), 2) SNVs in the omni genotyping array ("Omni"), and 3) SNVs in the 1000
455 Genomes Project ("1000G"). According to the VQSR parameter recommendation, SNVs in HapMap
456 and Omni call sets are considered to contain only true variants while SNVs in 1000G contain both true

457 and false positive variants(35). We call this recommended parameter setting "original VQSR." We,
458 however, find that considering SNVs in Omni to contain both true and false positive variants
459 considerably improves the quality of good SNVs from VQSR for the BP dataset. We call this modified
460 parameter setting "Omni_Modified VQSR". Results show that the mean Ti/Tv on good novel SNVs
461 from Omni_Modified VQSR is 1.76, which is much higher than that from original VQSR (1.41) and
462 slightly higher than that from ForestQC (1.68) (Figure S19a). We also find that the mean number of total
463 SNVs from Omni_Modified VQSR is 3.68M which is much smaller than that from original VQSR
464 (3.99M) but higher than that from ForestQC (3.58M) (Figure S19b). In terms of other statistics, good
465 SNVs from original VQSR has 3.66x higher ME rate, 7.40x higher genotype missing rate, and 1.16x
466 higher genotype discordance rate (p-value = 0.0001118) than those SNVs from Omni_Modified VQSR
467 (Figure S19c-e). Interestingly, we do not observe the improved performance of Omni_Modified VQSR
468 for the PSP dataset as the mean novel Ti/Tv on good novel SNVs of Omni_Modified VQSR is 1.23,
469 which is slightly smaller than that of original VQSR (1.27) (Figure S19a), although individuals have
470 fewer good SNVs from Omni_Modified VQSR (3.53M) than that from original VQSR (3.75M) (Figure
471 S19b). These results suggest that the performance of VQSR may change significantly depending on
472 whether to consider a certain SNV call set to contain only true variants or both true and false positive
473 variants, and it appears that the difference in performance is more noticeable in certain sequencing
474 datasets than others.

475 Although Omni_Modified VQSR has slightly better Ti/Tv on good novel SNVs and identifies more
476 good SNVs than does ForestQC, good SNVs from Omni_Modified VQSR have 2.76x higher ME rate,
477 13.20x higher genotype missing rate, and 1.35x higher genotype discordance rate (p-value < 2.2e-16)
478 than good SNVs from ForestQC (Figure S19c-e). Hence, the results show that good SNVs from
479 ForestQC have higher quality than those from VQSR even with the modification in the parameter
480 setting.

481 # Discussion

482 We developed an accurate and efficient method called ForestQC to identify a set of variants with high
483 sequencing quality from NGS data. ForestQC combines the traditional filtering approach for performing
484 QC in GWAS and the classification approach that uses a machine learning algorithm to classify whether
485 a variant has good quality. Our method first uses stringent filters to identify good and bad variants that
486 unequivocally have high and low sequencing quality, respectively. ForestQC then trains a random forest
487 classifier using the good and bad variants obtained from the filtering step, and predicts whether a variant
488 with ambiguous quality (a gray variant) is good or bad in an unbiased manner. To evaluate ForestQC,
489 we applied our method to two WGS datasets where one dataset consists of related individuals from
490 families and the other dataset has unrelated individuals. We demonstrated that good variants identified
491 from ForestQC in both datasets had higher sequencing quality than those from other approaches such as
492 VQSR and a filtering approach based on ABHet.

493 To measure the performance of methods for variant quality control, one typically plans to apply these
494 methods to benchmarking datasets where the true variants with high sequencing quality are verified. A
495 few high-quality benchmarking variant sets have been proposed, including Genome In A Bottle (GIAB)
496 (39), Platinum Genome (PlatGen) (40) and Syndip (41). GIAB has seven samples, PlatGen sequenced
497 17 individuals, and Syndip includes only two cell lines, CHM1 and CHM13. The sample sizes of these
498 datasets are very small while we usually need to perform variant QC on an entire large dataset
499 containing tens of millions of variants from hundreds of subjects or more. Thus, these datasets cannot be
500 used as benchmarking datasets for variant QC. Apart, it is not expected to have a new benchmarking
501 dataset with large sample size in the near future because it is expensive to construct such a dataset.
502 Hence, in this study, we used real WGS datasets to evaluate different approaches for variant QC. Their
503 large sample sizes allow more accurate calculation of various quality metrics and statistics used by the
504 approaches for variant QC, and therefore enable more reliable performance evaluation.

505 To measure the quality of variants, we used 21 sample-level metrics and 20 variant-level metrics, plus
506 genotype missing rate, ME rate and genotype discordance rate, resulting in a comprehensive evaluation
507 of the performance of different methods. ME rate is found to be nearly linearly correlated with genotype
508 errors(42–44), so it is a good quality metric for variants with pedigree information. Low genotype
509 missing rate has been considered as an indicator of high-quality variant call set as a variant with high
510 genotype missing rate indicates poor genotyping or sequencing quality(45). Also, high-quality variants
511 would have the same genotypes generated by different genotyping technologies, such as sequencing and
512 microarray. Thus, variant sequencing quality may be measured with genotype discordance rate between
513 microarray and sequencing. One challenge with this approach is that genotypes generated by microarray
514 are usually available for a small proportion of variants in the whole genome, especially for common and
515 known variants, so it might not be able to show the sequencing quality of the entire variant call set.
516 Another frequently used variant quality metric is Ti/Tv ratio (46–49). It is supposed to be around 2.0 for
517 whole genome sequencing data(36). That is because transitions have higher frequency according to
518 molecular mechanisms although the number of transversions is twice as many as transitions. Previous
519 studies found that mitochondrial DNA and some non-human DNA sequences might be biased towards
520 transitions or transversions(50,51). In this study, we only computed Ti/Tv ratio for each QC method
521 using the same human variant call set excluding mitochondria, in order to achieve an unbiased
522 evaluation of all methods.

523 A main advantage of our approach over the traditional filtering approach is that our method does not
524 attempt to classify gray variants using filters. It is difficult to determine the quality of those gray variants
525 using filters if their QC metrics (e.g. genotype missing rate) are close to the thresholds of filters. Hence,
526 ForestQC avoids a limitation of the traditional filtering approaches that determine the quality of every
527 variant using filters, which may exclude some of good variants from the downstream analysis. We did
528 not compare our approach with the traditional filtering approach used in GWAS that removes variants

529 according to HWE p-values, ME rates and genotype missing rates. One main reason is that the
530 performance of this approach changes dramatically depending on filters and thresholds for each filter,
531 and there are numerous different thresholds of filters as well as many combinations of filters that could
532 be tested. Another reason is that its performance could be arbitrarily determined depending on the filters
533 we use. For example, if one filter is to remove any variants having more than zero Mendel errors, the
534 ME rate of good variants would be zero, but we may be removing many other good variants. We
535 checked the accuracy of a filtering approach based on ABHet as ABHet is often used in performing QC
536 of NGS data and is a good indicator for variant quality(26,52,53). Also, as this approach is not based on
537 standard QC metrics such as genotype missing rate, its performance is independent of those metrics
538 unlike the standard filtering approaches. We showed that our approach outperformed the ABHet
539 approach as the quality of good variants from ForestQC was better than that from ABHet, regardless of
540 similar total number of good variants, as demonstrated by ME rate, missing rate, genotype discordance
541 rate and Ti/Tv ratio in the BP and PSP dataset.

542 Although our approach is similar to VQSR as both approaches train machine learning classifiers to
543 predict quality of variants, they have a few distinct differences. First, our approach trains the model
544 using good and bad variants detected from sequencing data on which quality control is performed, while
545 VQSR uses variants in existing databases, such as HapMap and 1000 genomes, as its training set. As
546 VQSR uses previously known variants for model training, good variants from VQSR are likely to
547 contain more known (and likely to be common) variants than novel (and rare) variants. We showed in
548 both WGS datasets that it did identify more common and known SNVs and indels as good variants than
549 ForestQC. This may not be a desirable outcome for some sequencing studies if one of their main goals is
550 to identify rare and novel variants not captured in chips. Another difference between ForestQC and
551 VQSR is the set of features used in the classifiers. While both methods use features related to
552 sequencing depth and genotyping quality, VQSR uses some features that are specifically calculated by

553 GATK software while our method uses quality information reported in the standard VCF file. This
554 suggests that our method is more generalizable than VQSR as it can be applied to VCF files generated
555 from variant callers other than GATK. The last difference is the machine learning algorithms that
556 ForestQC and VQSR use. Our method trains a random forest model while VQSR trains a Gaussian
557 Mixture model. Using the BP and PSP dataset, we found that random forest model was much faster than
558 Gaussian Mixture model (Table S11).

559 In addition to SNVs, we applied our method to indels in both WGS datasets and found that indels had
560 much lower sequencing quality than do SNVs as the fraction of good indels detected by ForestQC was
561 considerably smaller than that of SNVs. This is somewhat expected because indel or structural variant
562 calling is much more difficult than SNV calling from sequencing data, and some of them are likely to be
563 false positives(54,55). It is, however, important to note that VQSR classifies many more indels as good
564 variants than does ForestQC or ABHet, but those good indels from VQSR may not have high
565 sequencing quality. We showed that good indels from VQSR had similar Mendelian error rate to that
566 without performing QC, indicating the poor performance of VQSR on indels. VQSR considers indels
567 from Mills gold standard call set(37) as true variants, and while those indels might represent true variant
568 sites, it does not necessarily mean that genotyping on those sites is accurate. Therefore, genetic studies
569 need to perform stringent QC on indels to remove those erroneous calls and not to have false positive
570 findings in their downstream analysis.

571 We found that the performance of VQSR was improved dramatically for the BP dataset when we
572 considered SNVs in Omni genotyping array to have both true and false positive sites, compared with
573 when they were assumed to have all true sites. We, however, did not observe this performance
574 enhancement for the PSP dataset. This suggests that users may need to try different parameter settings to
575 obtain optimal results from VQSR for specific sequencing datasets they analyze. Another issue with
576 VQSR and also with ABHet is that some of good SNVs or indels have high genotype missing rate and

577 ME rate, which may not be suitable for the downstream analysis such as association analysis. Thus,
578 those variants need to be filtered out separately, which means users may need to perform an additional
579 filtering step in addition to applying VQSR and ABHet to the dataset. As the filtering step is
580 incorporated in ForestQC, our method does not have this issue.

581 Our approach is an extension of a previous approach that uses a logistic regression model to predict
582 the quality of variants in the BP dataset(30). While our approach is similar to the previous approach in
583 that they both combine filtering and classification approaches, ForestQC uses a random forest classifier
584 that has higher accuracy than a logistic regression model, according to our simulation results. It includes
585 more bad variants for model training, leading to predictions with fewer biases. ForestQC also includes
586 more features than the previous approach as well as more filters to improve the quality of good variants.
587 Additionally, compared with the previous approach, ForestQC is more user-friendly and generalizable
588 because users can choose or define different features and filters and tune the parameters according to
589 their research goals.

590 ForestQC is efficient, modularized and flexible with following features. First, users are allowed to
591 change thresholds for filters as needed. This is important because filters that are stringent for one dataset
592 may not be stringent for another dataset. For example, variants from sequence data with very small
593 sample size (e.g. < 100) may not have statistical power to have significant HWE p-values, and hence
594 higher p-value thresholds may need to be used, compared with studies with larger sample size. If filters
595 are not stringent enough, there may be many bad variants, and ForestQC would train a very stringent
596 classifier, leading to the possible removal of good variants. On the contrary, if the filters are too
597 stringent, there would be too few good variants or bad variants, which would lower the accuracy of our
598 random forest classifier. In this study, after the filtering step, 4.39% of SNVs and 15.72% of indels in
599 the BP dataset, and 5.06% of SNVs and 15.66% of indels in PSP dataset, were determined as bad
600 variants. Empirically, we suggest filters for ForestQC such that after the filtering step, a fraction of bad

601 variants is about 4-16%. Normally, the default parameter settings are recommended, which are the same
602 sets of filters and features described in this paper. The selection of threshold values for these filters are
603 based on our previous study for WGS data of extended pedigrees for bipolar disorder(30). Second, users
604 are allowed to use their own filters and features provided that they specify values for those new filters
605 and features at each variant site, and our software also allows users to remove existing filters and
606 features. As there may be filters and features that capture sequencing quality of variants more accurately
607 than current set of filters and features, this option allows users to improve ForestQC further. For
608 example, users can employ mappability, strand bias and micro-repeats as features, instead of sequencing
609 depth and genotyping quality used in this study, because DP and GQ might penalize disease-causing
610 variants with low coverage. Also, if users want to obtain more variants after QC, they may lower the
611 standard for good variants, that is, increase the threshold values of ME or missing rate for determining
612 good variants. Third, ForestQC generates the probability of each gray variant being a good variant. This
613 probability needs to be greater than a certain threshold for a gray variant to be predicted to be good, and
614 it can also be used to analyze sequencing quality of certain variants. If studies find that a certain gray
615 variant is associated with a phenotype, they may consider checking whether its probability of being a
616 good variant is high enough. Lastly, ForestQC allows users to change the probability threshold for
617 determining whether each gray variant is good or bad. Users may lower this threshold if they are
618 interested in obtaining more good variants at the cost of including more bad variants.

619 Materials and Methods

620 ForestQC

621 ForestQC consists of two approaches: a filtering approach and a machine learning approach based on a
622 random forest algorithm.

623 **Filtering** Given a variant call set from next generation sequencing data, ForestQC first applies several

624 stringent filters to identify good, bad, and gray variants. Good variants are ones that pass all filters while
625 bad variants fail any of them (Table S2, S3). Gray variants are variants that neither pass filters for good
626 variants nor fail filters for bad variants. We use following filters in the filtering step.

627 • Mendelian error (ME) rate. The Mendelian error occurs when a child's genotype is inconsistent
628 with genotypes from parents. ME rate is calculated as the number of ME among all trios divided
629 by the number of trios for a given variant. Note that this statistic is only available for family-based
630 data.

631 • Genotype missing rate. This is the proportion of missing alleles in each variant.

632 • Hardy-Weinberg equilibrium (HWE) p-value. This is a p-value for hypothesis testing whether a
633 variant is in Hardy-Weinberg equilibrium. Its null hypothesis is that the variant is in Hardy-
634 Weinberg equilibrium. We use the algorithm used in an open-source software, VCFtools(56) for
635 the calculation of Hardy-Weinberg equilibrium p-value.

636 • ABHet. This is allele balance for heterozygous calls. ABHet is calculated as the number of
637 reference reads from individuals with heterozygous genotypes divided by the total number of
638 reads from such individuals, which is supposed to be 0.50 for good variants. For variants in
639 chromosome X, we only calculate ABHet for females.

640 **Random forest classifier** Random forest algorithm is a machine learning algorithm that runs efficiently
641 on large datasets with high accuracy(34). Briefly, random forest builds several randomized decision
642 trees, each of which is trained to classify the input objects. For classification of a new object, the fitted
643 random forest model passes the input vector down to each of the decision trees in the forest. Each
644 decision tree has its classification result, then the forest would output the classification that the majority
645 of the decision trees make. Balancing efficiency and accuracy, we train a random forest classifier using
646 50 decision trees (Figure S2) and 50% as probability threshold (Figure S3).

647 To train random forest, we use good and bad variants identified from the previous filtering step as a
648 training dataset, after balancing their sample size by random sampling. Normally, good variants are
649 much more numerous than bad variants, so we randomly sample from good variants with the sample
650 size of bad variants. Hence, the sample size of the balanced training set would be twice as large as the
651 sample size of bad variants. We also need features in training random forest, which characterize
652 datasets, and we use following features.

653 • Mean and standard deviation of depth (DP) and genotyping quality (GQ). Depth and genotyping
654 quality values are extracted from DP and GQ fields of each sample in VCF files, respectively, and
655 mean and standard deviation are calculated over all samples for each variant.

656 • Outlier depth and outlier genotype quality. These are the proportions of samples whose DP or GQ
657 is lower than a particular threshold. We choose this threshold as the first quartile value of all DP
658 or GQ values of variants on chromosome 1. We use DP and GQ of variants on only chromosome
659 1 to reduce the computational costs.

660 • GC content: We first split a reference genome into window size of 1,000 bp and calculate GC
661 content for each window as (# of G or C alleles) / (# of A, G, C or T alleles). Then, each variant is
662 assigned a GC content value according to its position in the reference genome.

663 After training random forest with the training dataset using above features, we next use the fitted
664 model to make predictions on gray variants on being good variants. Gray variants with the predicted
665 probability of being good larger than 50% are labeled as predicted good variants. Then the predicted
666 good variants and good variants from the previous filtering step are combined to form the final set of
667 good variants. We apply the same procedure to identify bad variants.

668 Comparison of different machine learning algorithms

669 We compare eight different machine learning algorithms, in order to identify the best algorithm used for
670 ForestQC. They are 1) k-nearest neighbors for supervised 2-class classification (8 threads); 2) logistic

671 regression (8 threads); 3) single support vector machine with Gaussian kernel function and penalty
672 parameter C of 1 (1 thread); 4) random forest with 50 trees (8 threads); 5) naïve Bayes without any prior
673 probabilities of the classes (1 thread); 6) artificial neural network with sigmoid function as activation
674 function (8 threads). It has 1 hidden layer with 10 units; 7) AdaBoost with 50 estimators and learning
675 rate of 1, which uses SAMME.R real boosting algorithm (1 thread); 8) and quadratic discriminant
676 analysis without any prior on classes. Its regularization is 0 and its threshold for rank estimation is 1e-4
677 (1 thread). Other parameters of these machine learning algorithm are default, as described in the
678 documentations of Python scikit-learn package(57). All learning algorithms use the seven
679 aforementioned features: mean and standard deviation of sequencing depth, mean and standard deviation
680 of genotype quality, outlier depth, outlier quality and GC content.

681 To test these eight machine learning algorithms, we obtain training and test datasets from the BP
682 dataset, using filters described in Table S2 and S3. There are 21,248,103 good SNVs and 2,257,506
683 good indels while there are 1,100,325 bad SNVs and 624,965 bad indels. We sample 100,000 variants
684 randomly from good variants and 100,000 variants from bad variants to generate a training set.
685 Similarly, 100,000 good variants and 100,000 bad variants are randomly chosen from the rest of variants
686 to form a test set. Each machine learning model shares the same training and test sets. We train the
687 machine learning models and measure training time at a training stage, and then test their accuracy and
688 measure prediction time at a testing stage. We measure the time cost of each algorithm, which is the
689 elapsed clock time between the start and end of each algorithm. To assess the performance of each
690 algorithm, we compute F1-score for the test set. F1-score is the harmonic average of precision and
691 recall, which is calculated as $2 \cdot precision \cdot \frac{recall}{(precision+recall)}$. The closer F1-score is to 1, the higher
692 classification accuracy is. Recall is the fraction of true positive results over all samples that should be
693 given positive prediction. Precision is the number of true positive results divided by the number of

694 positive results predicted by the classifier. We also measure the model accuracy using 10-fold cross
695 validation, as well as the area under the receiver operating characteristic curve.

696 ABHet approach and VQSR

697 We compare ForestQC with two other approaches for performing QC on genetic variants. One is a
698 filtering approach based on ABHet and the other is a classification approach called VQSR from GATK
699 software. For the ABHet approach, we consider variants with ABHet > 0.7 or < 0.3 as bad variants, and
700 the rest as good variants. We chose this threshold setting of ABHet (> 0.3 and < 0.7) because the ADSP
701 project could not reliably confirm heterozygous calls with ABHet > 0.7 with Sanger sequencing(26). We
702 also exclude variants with small ABHet values (< 0.3) to ensure high quality. For GATK, we use
703 recommended arguments as of 2018-04-04(35). For SNVs, VQSR takes SNVs in HapMap 3 release 3,
704 1000 Genome Project and Omni genotyping array as training resources, and dbSNP135 as known sites
705 resource. HapMap and Omni sites are considered as true sites, meaning that SNVs in these datasets are
706 all true variants, while 1000 Genome Project sites are regarded as false sites, meaning that there could
707 be both true and false-positive variants. The desired level of sensitivity of true sites is set to be 99.5%. In
708 the BP dataset, we run VQSR version 3.5-0-g36282e4 with following annotations; quality by depth
709 (QD), RMS mapping quality (MQ), mapping quality rank sum test (MQRankSum), read position rank
710 sum test (ReadPosRankSum), fisher strand (FS), coverage (DP) and strand odds ratio (SOR) to evaluate
711 the likelihood of true positive calls. In the PSP dataset, we use VQSR version 3.2-2-gec30cee that uses
712 all previous annotations and additional inbreeding coefficient (InbreedingCoeff) except SOR because
713 variants in PSP dataset do not have the SOR annotation. For indels, VQSR takes indels in Mills gold
714 standard call set(37) as true training resource, and dbSNP135 as known sites resource. The desired level
715 of sensitivity of true sites is set to be 99.0%. We use VQSR version 3.5-0-g36282e4 with QD, DP, FS,
716 SOR, ReadPosRankSum and MQRankSum annotations to evaluate the likelihood of true positive calls

717 in the BP dataset, while we run VQSR version 3.2-2-gec30cee with the same annotations and additional
718 InbreedingCoeff except SOR for the PSP dataset.

719 BP and PSP WGS datasets

720 The BP WGS dataset is for studying bipolar disorder whose average coverage is 36. This study recruited
721 individuals from 11 Colombia (CO) and 15 Costa Rica (CR) extended pedigrees in total. 454 subjects
722 from 10 CO and 12 CR families are both whole genome sequenced and genotyped with microarray.
723 There are 144 individuals diagnosed with BP1 and 310 control samples that are unaffected or have non-
724 BP traits. We use highly scalable Churchill pipeline(58) to do variant calling for the BP data set, where
725 GATK-HaplotypeCaller 3.5-0-g36282e4 is used as the variant caller according to the GATK best
726 practices(23) and the reference genome is HG19. After initial QC on individuals, five individuals are
727 removed because of poor sequencing quality and possible sample mix-ups. Finally, 449 individuals are
728 included in an analysis, resulting in 25,081,636 SNVs and 3,976,710 indels. 1,814,326 SNVs in the
729 WGS dataset are also genotyped with microarray, which are used to calculate genotype discordance rate.
730 In this study, we use the BP dataset before any QC performed on genetic variants. In a previous
731 study(30), genetic variants in the BP WGS dataset are first processed with VQSR and then filtered with
732 a trained logistic regression model to remove variants with low quality.

733 The PSP WGS dataset is for studying progressive supranuclear palsy with average coverage of 29.
734 544 unrelated individuals are whole genome sequenced, 518 of whom are also genotyped with
735 microarray. Among them, 119 individuals have 547,644 SNPs and 399 individuals have 1,682,489 SNPs
736 genotyped with microarray, respectively. That 119 individuals would be excluded when calculating
737 genotype discordance rate in case of biases caused by fewer SNPs. There are 356 individuals diagnosed
738 with PSP and 188 individuals as controls. Variant calling for the PSP dataset is performed using
739 Churchill pipeline, where GATK-HaplotypeCaller 3.2-2-gec30cee is used as the variant caller according
740 to the GATK best practices and the reference genome is HG19. 49 samples are found to have high

741 missing rate or high relatedness with other samples, or are diagnosed with diseases other than PSP, so
742 they are removed. Next, we extract variant data with only 495 individuals with VCFtools. Monomorphic
743 variants are then removed. After preprocessing, the PSP WGS dataset has 33,273,111 SNVs and
744 5,093,443 indels. There are 1,682,489 SNVs from 381 samples genotyped by both microarray and WGS,
745 which are used for calculating genotype discordance rate.

746 **Performance metrics**

747 21 sample-level metrics and 20 variant-level metrics are defined to measure the sequencing quality of
748 the variant call set after performing quality control (Table S12). Note that we do not show all sample-
749 level metrics and variant-level metrics in the main text. Other metrics are available in supplemental
750 materials. Variant-level metrics provide us with a summarized assessment report of the sequencing
751 quality of a variant call set, such as total SNVs of the whole dataset. They are calculated based on the
752 information of all variants in a variant call set. For example, the number and the proportion of multi-
753 allelic SNVs are counted for the entire dataset, each of which is identified according to its reference and
754 alternate alleles. On the other hand, sample-level metrics enable the inspection of the sequencing quality
755 for sequenced individuals in a variant call set. For instance, we check the distribution of novel Ti/Tv or
756 other quality metrics among all individuals in the study. Sample-level metrics are calculated for each
757 sample, using its genotype information on all variants in the dataset, and a distribution of those metrics
758 across all individuals is shown as a box plot. For example, the number of SNV singletons on a sample
759 level shows the distribution of the number of SNV singletons across all sequenced individuals. In this
760 study, both sample-level and variant-level metrics are used to evaluate the sequencing quality of WGS
761 variant datasets.

762 In addition, we also use genotype missing rate, ME rate and genotype discordance rate as variant
763 quality metrics, which are computed using the entire variant call set. The definitions of genotype
764 missing rate and ME rate have been described above. Note that ME rate is only available for family-

765 based datasets, such as the BP dataset, so we do not calculate ME rate for the PSP dataset that only
766 includes unrelated individuals. Genotype discordance rate is the proportion of individuals whose
767 genotypes are inconsistent between next-generation sequencing and microarray. This metric can only be
768 calculated with a subset of variants due to the limited number of variants genotyped by both sequencing
769 and microarray. Note that microarray might also have biases in genotyping, leading to some limitations
770 of genotype discordance rate. For example, microarray usually genotype selected variants, especially
771 common and known variants, so genotype discordance rate is only available for these selected variants
772 and it cannot provide quality evaluation for all variants, especially rare variants. Genotype missing rate,
773 ME rate and genotype discordance rate provide us with accurate evaluation of variant quality, because
774 true positive variants with high quality are very likely to have low values of these three metrics.

775 Competing Interests

776 The authors declare no competing interests.

777 Acknowledgements

778 We thank Dr. Susan K. Service from Department of Psychiatry and Biobehavioral Sciences, UCLA for
779 the precious comments and suggestions to our project and this manuscript. We thank all study participants
780 in the BP and PSP datasets.

781 Funding

782 This study was supported by the National Institute of Environmental Health Sciences (NIEHS) grant
783 K01 ES028064 and the National Science Foundation grant #1705197.

784 # Availability and Implementation

785 ForestQC is available at https://github.com/avallonking/ForestQC under the open source MIT license.
786 There are detailed installation instructions and user guide. It is implemented with Python3 and is
787 compatible with Linux, Mac OSX and Windows 64-bit operating systems.

788 # Authors' contributions

789 JL and JHS designed the method and conceived this study. JL developed the method and did the analysis
790 of the BP and PSP datasets. JHS did preprocessing and variant calling for the BP dataset. BJ and SH did
791 preprocessing and variant calling for the PSP dataset. LZ tested the software. NF provided the BP
792 dataset and GC contributed the PSP dataset. JL and JHS wrote the manuscript. All authors read and
793 approved the final manuscript.

794 # References

795 1. Pray L. Genome-wide association studies and human disease networks. Nat Educ. nature.com;
796 2008;1(1):220.

797 2. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex
798 traits. Nat Rev Genet. 2005 Feb;6(2):95–108.

799 3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from
800 108 schizophrenia-associated genetic loci. Nature. 2014 Jul;511(7510):421–7.

801 4. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study
802 identifies novel risk loci for type 2 diabetes. Nature. search.proquest.com; 2007
803 Feb;445(7130):881–5.

804 5. Ng MCY, Shriner D, Chen BH, Li J, Chen W-M, Guo X, et al. Meta-analysis of genome-wide

805 association studies in African Americans provides insights into the genetic architecture of type 2
806 diabetes. PLoS Genet. journals.plos.org; 2014 Aug;10(8):e1004517.
807 6. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, et al. Large-scale meta-analysis
808 of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet.
809 nature.com; 2014 Sep;46(9):989–93.
810 7. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum
811 Genet. 2012 Jan;90(1):7–24.
812 8. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009
813 Apr;360(17):1696–8.
814 9. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through
815 whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415–25.
816 10. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common
817 diseases. Nat Genet. 2008 Jun;40(6):695–701.
818 11. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex
819 diseases. Curr Opin Genet Dev. 2009 Jun;19(3):212–9.
820 12. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum
821 Genet. 2001 Jul;69(1):124–37.
822 13. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of
823 GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat
824 Genet. 2011 Oct;43(11):1066–73.
825 14. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, et
826 al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with
827 prostate cancer. Nat Genet. 2012 Dec;44(12):1326–9.

828 15. Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang Z-Z, et al. Whole-exome sequencing
829 identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum
830 Genet. 2014 Feb;94(2):233–45.

831 16. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, et al. Using
832 whole-exome sequencing to identify inherited causes of autism. Neuron. 2013 Jan;77(2):259–73.

833 17. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets
834 from high-throughput DNA sequencing. Nucleic Acids Res. 2008 Sep;36(16):e105.

835 18. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific
836 error profile of Illumina sequencers. Nucleic Acids Res. 2011 Jul;39(13):e90.

837 19. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and
838 measuring bias in sequence data. Genome Biol. 2013 May;14(5):R51.

839 20. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale
840 variation in metagenomic sequencing data. BMC Bioinformatics. 2016 Mar;17:125.

841 21. Manley LJ, Ma D, Levine SS. Monitoring Error Rates In Illumina Sequencing. J Biomol Tech.
842 2016 Dec;27(4):125–8.

843 22. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov M V, Oparina NY, et
844 al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. 2014 Mar;4:4532.

845 23. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for
846 variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011
847 May;43(5):491–8.

848 24. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, et al. An analytical framework
849 for optimizing variant discovery from personal genomes. Nat Commun. 2015 Feb;6:6275.

850 25. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-

851 calling pipelines: practical implications for exome and genome sequencing. Genome Med.
852 genomemedicine.biomedcentral. …; 2013 Mar;5(3):28.

853 26. ADSP. Review and Proposed Actions for False-Positive Association Results in ADSP Case-
854 Control Data | ADSP [Internet]. https://www.niagads.org/adsp/content/review-and-proposed-
855 actions-false-positive-association-results-adsp-case-control-data. 2016. Available from:
856 https://www.niagads.org/adsp/content/review-and-proposed-actions-false-positive-association-
857 results-adsp-case-control-data

858 27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
859 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
860 Genome Res. 2010 Sep;20(9):1297–303.

861 28. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et
862 al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

863 29. International HapMap Consortium. The International HapMap Project. Nature. 2003
864 Dec;426(6968):789–96.

865 30. Sul JH, Susan K Service, Huang AY, Ramensky V, Hwang S-G, Teshiba TM, et al. Contribution
866 of common and rare variants to bipolar disorder susceptibility in extended pedigrees from
867 population isolates. bioRxiv. 2018 Jul. doi: 10.1101/363267

868 31. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance
869 comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011 Sep;29(10):908–14.

870 32. Wang W, Wei Z, Lam T-W, Wang J. Next generation sequencing has lower sequence coverage
871 and poorer SNP-detection capability in the regulatory regions. Sci Rep. 2011 Aug;1:55.

872 33. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing
873 PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011 Feb;12(2):R18.

874 34. Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32.

875 35. GATK Dev Team. Which training sets / arguments should I use for running VQSR?
876 https://software.broadinstitute.org/gatk/documentation/article.php?id=1259. 2017 Sep;

877 36. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, et al. Targeted
878 enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant
879 densities. Genome Biol. 2011 Jul;12(7):R68.

880 37. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic
881 variation caused by small insertions and deletions in the human genome. Genome Res. 2011
882 Jun;21(6):830–9.

883 38. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated
884 map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75–81.

885 39. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human
886 sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat
887 Biotechnol [Internet]. 2014 Mar 16 [cited 2019 Feb 13];32(3):246–51. Available from:
888 http://www.ncbi.nlm.nih.gov/pubmed/24531798

889 40. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data
890 set of 5.4 million phased human variants validated by genetic inheritance from sequencing a
891 three-generation 17-member pedigree. Genome Res [Internet]. 2017 Jan [cited 2019 Feb
892 13];27(1):157–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27903644

893 41. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid
894 benchmark for accurate variant-calling evaluation. Nat Methods [Internet]. NIH Public Access;
895 2018 Aug [cited 2019 Feb 13];15(8):595–7. Available from:
896 http://www.ncbi.nlm.nih.gov/pubmed/30013044

897 42. Saunders IW, Brohede J, Hannan GN. Estimating genotyping error rates from Mendelian errors in

898 SNP array genotypes and their impact on inference. Genomics [Internet]. Academic Press; 2007
899 Sep 1 [cited 2018 Dec 14];90(3):291–6. Available from:
900 https://www.sciencedirect.com/science/article/pii/S088875430700136X

901 43. Sobel E, Papp JC, Lange K. Detection and Integration of Genotyping Errors in Statistical
902 Genetics. Am J Hum Genet [Internet]. Cell Press; 2002 Feb 1 [cited 2018 Dec 14];70(2):496–508.
903 Available from: https://www.sciencedirect.com/science/article/pii/S0002929707639627

904 44. Hao K, Li C, Rosenow C, Hung Wong W. Estimation of genotype error rate using samples with
905 pedigree information—an application on the GeneChip Mapping 10K array. Genomics [Internet].
906 Academic Press; 2004 Oct 1 [cited 2018 Dec 14];84(4):623–30. Available from:
907 https://www.sciencedirect.com/science/article/pii/S0888754304001193

908 45. Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation
909 distortion in molecular marker data on the construction of linkage maps. Heredity (Edinb)
910 [Internet]. Nature Publishing Group; 2003 Jan 9 [cited 2018 Dec 14];90(1):33–8. Available from:
911 http://www.nature.com/articles/6800173

912 46. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of
913 human genome variation from population-scale sequencing. Nature [Internet]. Nature Publishing
914 Group; 2010 Oct 28 [cited 2018 Dec 14];467(7319):1061–73. Available from:
915 http://www.nature.com/doifinder/10.1038/nature09534

916 47. Guo Y, Zhao S, Sheng Q, Ye F, Li J, Lehmann B, et al. Multi-perspective quality control of
917 Illumina exome sequencing data using QC3. Genomics [Internet]. Academic Press; 2014 May 1
918 [cited 2019 Jan 12];103(5–6):323–8. Available from:
919 https://www.sciencedirect.com/science/article/pii/S0888754314000354

920 48. Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read
921 sequencing data. BMC Genomics [Internet]. 2012 [cited 2018 Dec 14];13(1):666. Available from:

922 http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-666

923 49. Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, et al. Exome sequencing generates high quality
924 data in non-target regions. BMC Genomics [Internet]. 2012 [cited 2018 Dec 14];13(1):194.
925 Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-194

926 50. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of
927 the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A [Internet]. National
928 Academy of Sciences; 2008 Jul 8 [cited 2018 Dec 14];105(27):9272–7. Available from:
929 http://www.ncbi.nlm.nih.gov/pubmed/18583475

930 51. Lanave C, Tommasi S, Preparata G, Saccone C. Transition and transversion rate in the evolution
931 of animal mitochondrial DNA. Biosystems [Internet]. Elsevier; 1986 Jan 1 [cited 2018 Dec
932 14];19(4):273–83. Available from:
933 https://www.sciencedirect.com/science/article/pii/0303264786900043

934 52. Aylward A, Cai Y, Lee A, Blue E, Rabinowitz D, Haddad Jr J, et al. Using Whole Exome
935 Sequencing to Identify Candidate Genes With Rare Variants In Nonsyndromic Cleft Lip and
936 Palate. Genet Epidemiol. 2016 Jul;40(5):432–41.

937 53. Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, et al.
938 Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in
939 1779 cases and 1273 controls. Neurobiol Aging. 2017 Nov;59:220.e1--220.e9.

940 54. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation
941 Sequencing Data. Front Bioeng Biotechnol [Internet]. 2015 Jun;3:92. Available from:
942 http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract

943 55. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read
944 data. Hum Genomics. 2015 Aug;9:20.

945 56. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call

946 format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156–8.

947 57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
948 Machine Learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.

949 58. Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, et al. Churchill: an ultra-fast,
950 deterministic, highly scalable and balanced parallelization strategy for the discovery of human
951 genetic variation in clinical and population-scale genomics. Genome Biol. 2015 Jan;16:6.

952

953 # Supporting Information

954 Supporting Information includes 19 figures and 11 tables. Captions listed below.

955 Figure S1: Receiver operating charateristic (ROC) curves and area under the curve of eight machine
956 learning models.

957 Figure S2: Relationship between the number of trees in random forest model and the performance of
958 ForestQC. Relationship between the number of trees and (a) CPU time and (b) F1-score.

959 Figure S3: Relationship between the probability threshold for predicting a variant to be good and the
960 precision of ForestQC. If the probability of a variant predicted to be good is larger than the probability
961 threshold, this variant would be labeled as a good variant. Classification precision changes along with
962 the probability threshold in SNV classification (a) and indel classification (b). The precision of
963 ForestQC is measured in F1-score.

964 Figure S4: Overall quality of good and bad variants in the BP dataset identified by ForestQC using ME
965 rate as a filter or not. The average Mendelian error rate and genotype missing rate for SNVs and indels,
966 and genotype discordance rate to microarray data for SNVs are shown. Data are represented as the mean
967 ± SEM.

968 Figure S5: Sample-level quality metrics of good variants in the BP dataset identified by ForestQC using
969 ME rate as a filter or not. (a) Total number of SNVs. (b) The number of SNVs found in dbSNP. (c) the
970 number of SNVs not found in dbSNP. (d) Ti/Tv ratio of SNVs found in dbSNP. (e) Ti/Tv ratio of SNVs
971 not found in dbSNP. (f) Total number of indels. (g) the number of indels found in dbSNP. (h) the
972 number of indels not found in dbSNP. The version of dbSNP is 150.

973 Figure S6: Overall quality of bad variants in the BP dataset detected by four different methods, including
974 no QC applied, ABHet approach, VQSR and ForestQC. The average Mendelian error rate and genotype
975 missing rate for SNVs and indels, and genotype discordance rate to microarray data for SNVs are
976 shown. Data are represented as the mean ± SEM.

977 Figure S7: Sample-level quality metrics of good variants in the BP dataset identified by four different
978 methods, including no QC applied, ABHet approach, VQSR and ForestQC. (a) Ti/Tv ratio of SNVs
979 found in dbSNP. (b) The number of SNVs found in dbSNP. (c) The number of indels found in dbSNP.
980 (d) The number of indels not found in dbSNP. The version of dbSNP is 150.

981 Figure S8: Overall quality of rare variants (MAF < 0.03) and common variants (MAF 0.03) in the BP
982 dataset. The average Mendelian error rate and genotype missing rate for SNVs and indels, and genotype
983 discordance rate to microarray data for SNVs are shown. Data are represented as the mean ± SEM.

984 Figure S9: Overall quality of good variants identified from gray variants in the BP dataset processed by
985 four different methods, including no QC applied, ABHet approach, VQSR and ForestQC. The average
986 Mendelian error rate and genotype missing rate for SNVs and indels, and genotype discordance rate to
987 microarray data for SNVs are shown. Data are represented as the mean ± SEM.

988 Figure S10: Sample-level quality metrics of good variants identified from gray variants in the BP dataset
989 processed by four different methods, including no QC applied, ABHet approach, VQSR and ForestQC.
990 (a) Total number of SNVs. (b) The number of SNVs found in dbSNP. (c) the number of SNVs not found
991 in dbSNP. (d) Ti/Tv ratio of SNVs found in dbSNP. (e) Ti/Tv ratio of SNVs not found in dbSNP. (f)

992 Total number of indels. (g) the number of indels found in dbSNP. (h) the number of indels not found in
993 dbSNP. The version of dbSNP is 150.

994 Figure S11: Selected sample-level quality metrics of good variants in BP dataset identified by VQSR
995 using "SOR" or not. (a) Ti/Tv ratio of SNVs not found in dbSNP, (b) the number of total SNVs and (c)
996 the number of total indels in the BP dataset processed with VQSR using "SOR" or not. SOR stands for
997 StrandOddsRatio, which is a metric for strand bias measured by the Symmetric Odds Ratio test. The
998 version of dbSNP is 150.

999 Figure S12: Overall quality of bad variants in the PSP dataset detected by four different methods,
1000 including no QC applied, ABHet approach, VQSR and ForestQC. The average genotype missing rate for
1001 both SNVs and indels, and genotype discordance rate to microarray data for SNVs are shown. Data are
1002 represented as the mean ± SEM.

1003 Figure S13: Sample-level quality metrics of good variants in PSP dataset identified by four different
1004 methods, including no QC applied, ABHet approach, VQSR and ForestQC. (a) Ti/Tv ratio of SNVs
1005 found in dbSNP. (b) The number of SNVs found in dbSNP. (c) The number of indels found in dbSNP.
1006 (d) The number of indels not found in dbSNP. The version of dbSNP is 150.

1007 Figure S14: Overall quality of rare variants (MAF < 0.03) and common variants (MAF 0.03) in the PSP
1008 dataset. The average genotype missing rate for SNVs and indels, and genotype discordance rate to
1009 microarray data for SNVs are shown. Data are represented as the mean ± SEM.

1010 Figure S15: Overall quality of good variants identified from gray variants in the PSP dataset processed
1011 by four different methods, including no QC applied, ABHet approach, VQSR and ForestQC. The
1012 average genotype missing rate for both SNVs and indels, and genotype discordance rate to microarray
1013 data for SNVs are shown. Data are represented as the mean ± SEM.

1014 Figure S16: Sample-level quality metrics of good variants identified from gray variants in the PSP
1015 dataset processed by four different methods, including no QC applied, ABHet approach, VQSR and

1016 ForestQC. (a) Total number of SNVs. (b) The number of SNVs found in dbSNP. (c) the number of
1017 SNVs not found in dbSNP. (d) Ti/Tv ratio of SNVs found in dbSNP. (e) Ti/Tv ratio of SNVs not found
1018 in dbSNP. (f) Total number of indels. (g) the number of indels found in dbSNP. (h) the number of indels
1019 not found in dbSNP. The version of dbSNP is 150.

1020 Figure S17: Feature importance of each feature in the random forest model of ForestQC applied to the
1021 BP and PSP datasets. DP stands for sequencing depth. GQ stands for genotyping quality. SD means
1022 standard deviation. Outlier DP or GQ means the proportion of samples having genotyping quality or
1023 sequencing depth lower than the first quartile of depth or genotyping quality in chromosome 1. GC
1024 stands for the GC content of a 1000-bp window where the variant is located. (a) Feature importance in
1025 SNV classification. (b) Feature importance in indel classification.

1026 Figure S18: Pearson's correlation coefficients between each pair of features in the BP and PSP dataset.

1027 Figure S19: Quality of good SNVs identified by VQSR with two different settings of training resources
1028 and ForestQC. (a) Ti/Tv ratio of SNVs not found in dbSNP v150 and (b) total number of SNVs in the
1029 BP and PSP dataset. (c)-(e) Average Mendelian error rate, average genotype missing rate, and average
1030 genotype discordance rate of good SNVs in the BP dataset. Data are represented as the mean ± SEM.
1031 "Omni_Modified VQSR": SNVs in Omni chip array call set are considered to contain both true and
1032 false positive sites. "original VQSR": SNVs in Omni chip array call set are considered to contain only
1033 true sites.

1034

1035 Table S1: Accuracy of eight different machine learning algorithms

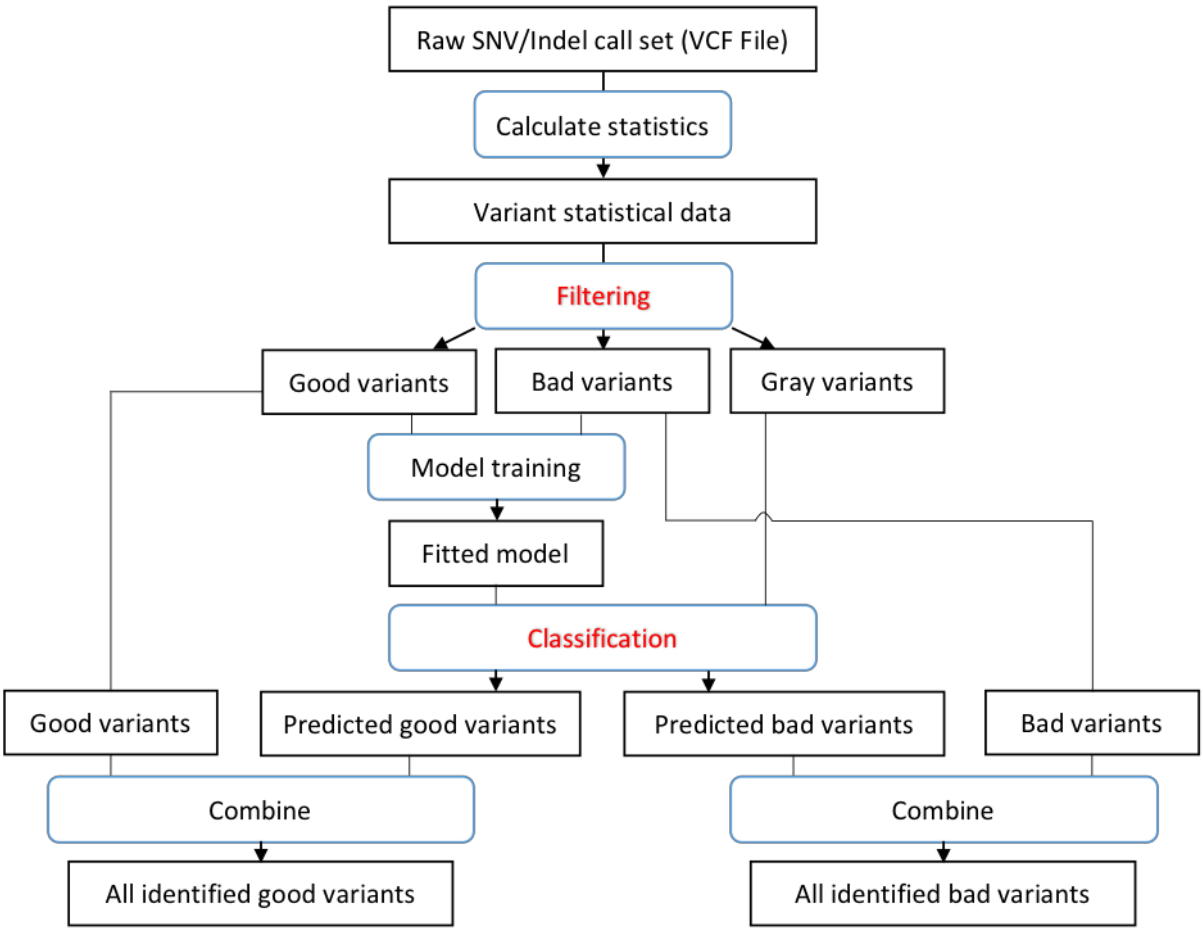
1036 Table S2: Thresholds of four filters for the selection of good variants from the original dataset

1037 Table S3: Thresholds of four filters for the selection of bad variants from the original dataset

1038 Table S4: Variant-level quality metrics of variants in the BP dataset processed by ForestQC with
1039 different settings

1040 Table S5: Variant-level quality metrics of good variants in the BP dataset processed by different
1041 methods
1042 Table S6: Rare variants and common variants in the BP dataset processed by different methods
1043 Table S7: Variant-level quality metrics of good variants identified from gray variants in the BP dataset
1044 Table S8: Variant-level quality metrics of good variants in the PSP dataset processed by four different
1045 methods
1046 Table S9: Rare variants and common variants in the PSP dataset processed by different methods
1047 Table S10: Variant-level quality metrics of good variants identified from gray variants in the PSP dataset
1048 Table S11: Running time of ForestQC and VQSR in two datasets, measured in real time
1049 Table S12: Definitions of 23 metrics for sequencing quality control calculated for sample-level and
1050 variant-level
1051
1052
1053

1054 # Figures



1055

1056 Figure 1: Workflow of ForestQC. ForestQC takes a raw variant call set in the VCF format as input. Then
1057 it calculates the statistics of each variants, including MAF, mean depth, mean genotyping quality, etc..
1058 In the filtering step, it separates the variant call set into good, bad, and gray variants by applying various
1059 hard filters, such as Mendelian error rate and genotype missing rate. In classification step, good and bad
1060 variants are used to train a random forest model, which is then applied to assign labels to gray variants.
1061 Variants predicted to be good among gray variants are combined with good variants from the
1062 classification step for the final set of good variants. The same procedure applies to find the final set of
1063 bad variants.

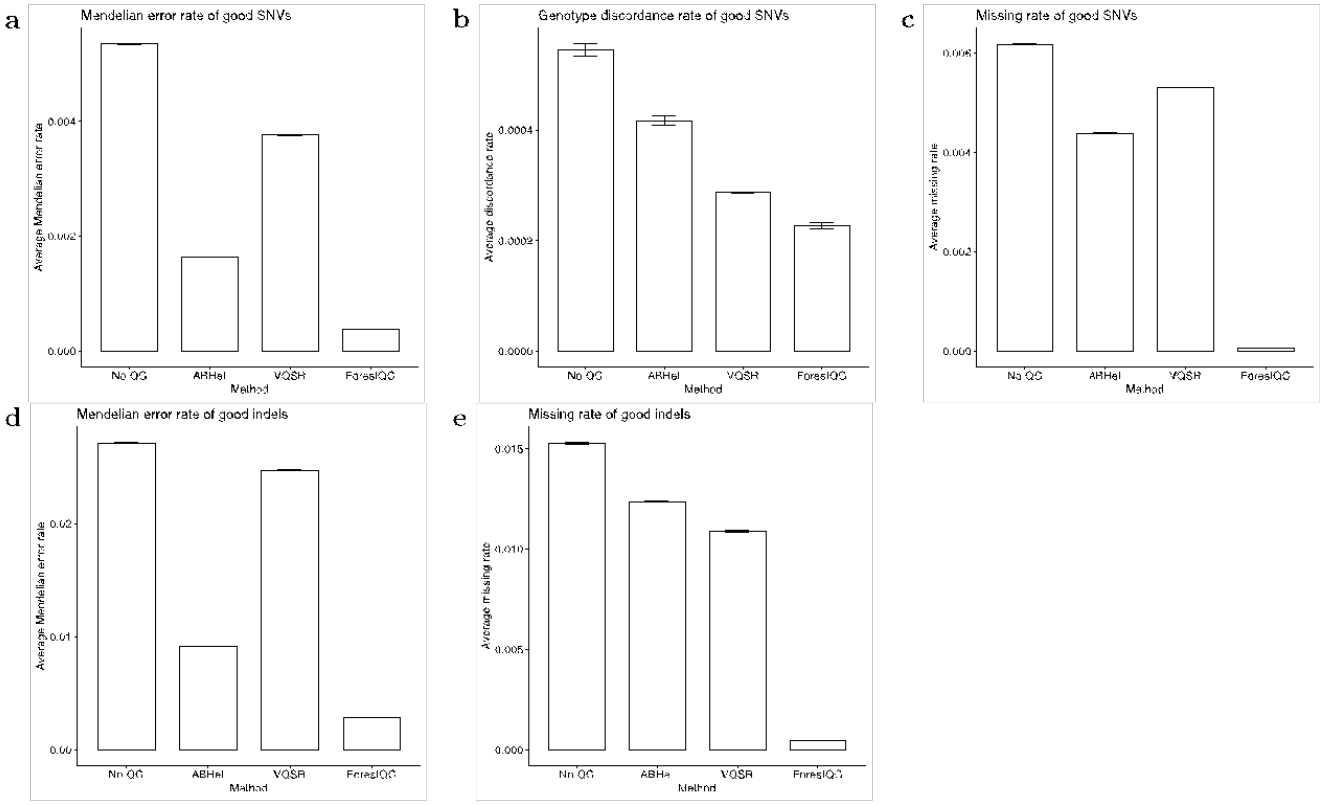


Figure 2: Overall quality of good variants in the BP dataset detected by four different methods, including no QC applied, ABHet approach, VQSR and ForestQC. The average Mendelian error rate and genotype missing rate for SNVs and indels, and genotype discordance rate to microarray data for SNVs are shown. Data are represented as the mean ± SEM.

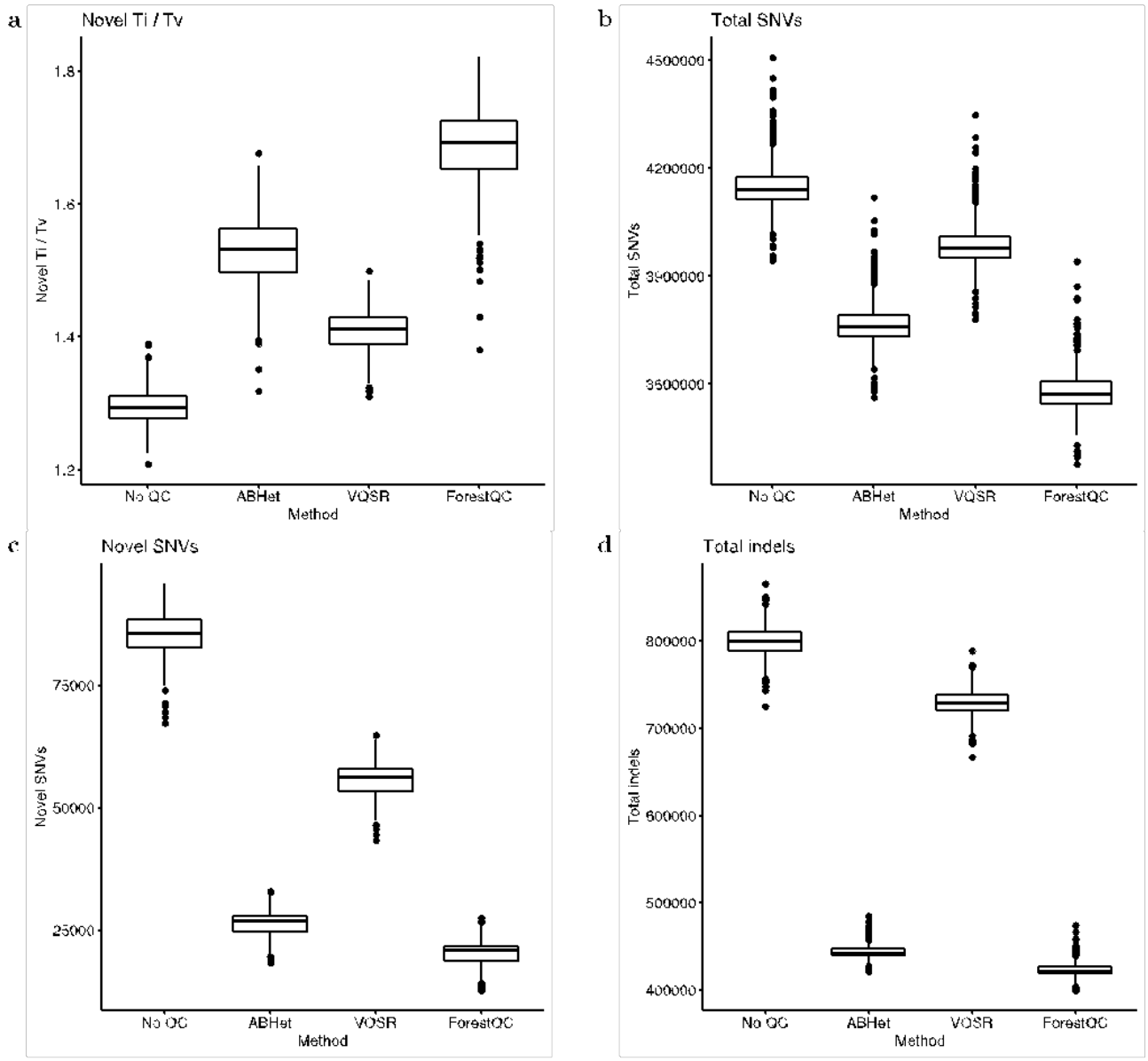
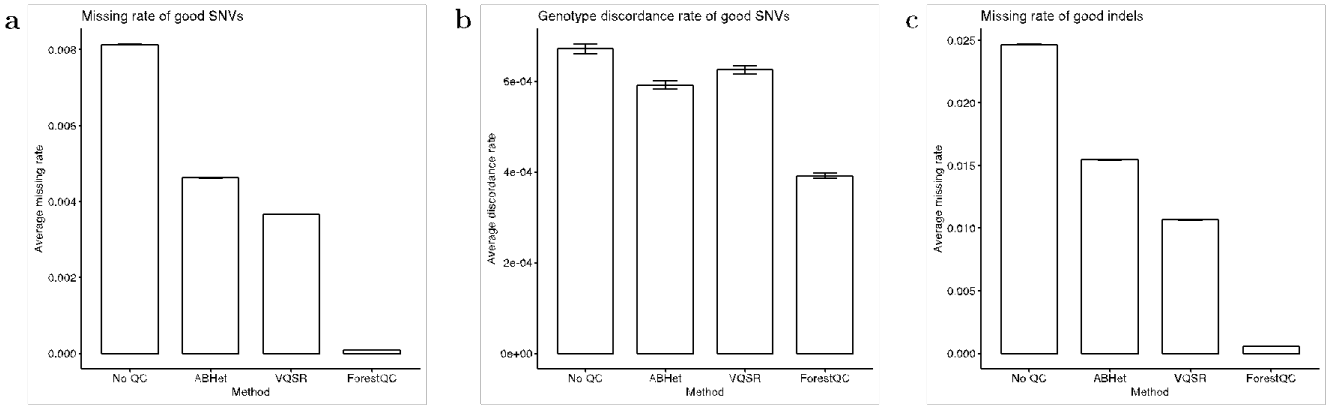


Figure 3: Sample-level quality metrics of good variants in the BP dataset identified by four different methods, including no QC applied, ABHet approach, VQSR and ForestQC. (a) Ti/Tv ratio of SNVs not found in dbSNP. (b) Total number of SNVs. (c) The number of SNVs not found in dbSNP. (d) Total number of indels. The version of dbSNP is 150.

1077 

1078 Figure 4: Overall quality of good variants in the PSP dataset detected by four different methods,
1079 including no QC applied, ABHet approach, VQSR and ForestQC. The average genotype missing rate for
1080 both SNVs and indels, and genotype discordance rate to microarray data for SNVs are shown. Data are
1081 represented as the mean ± SEM.

1082

1083

1084

1085

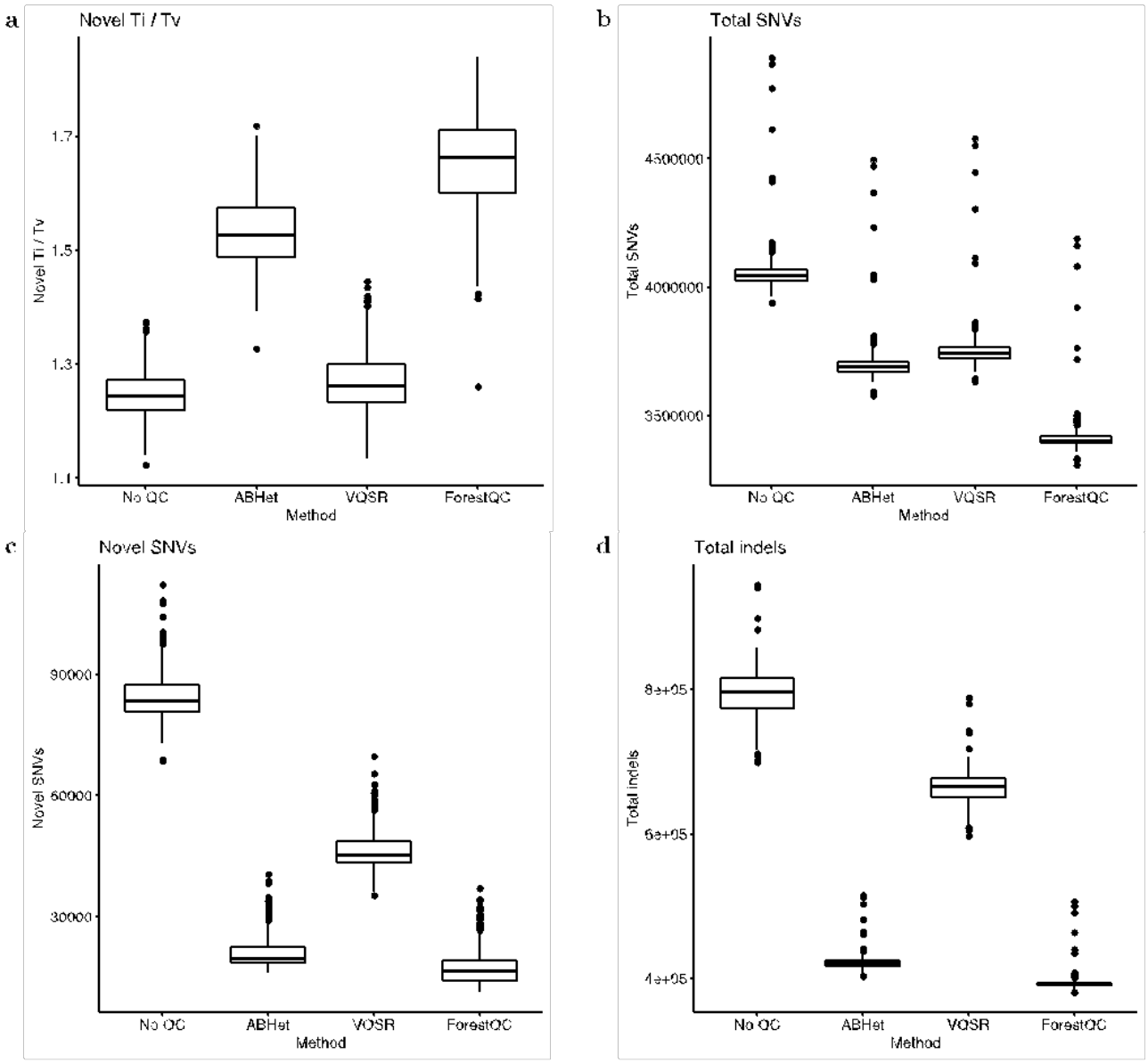


Figure 5: Sample-level quality metrics of good variants in the PSP dataset identified by four different methods, including no QC applied, ABHet approach, VQSR and ForestQC. (a) Ti/Tv ratio of SNVs not found in dbSNP. (b) Total number of SNVs. (c) The number of SNVs not found in dbSNP. (d) Total number of indels. The version of dbSNP is 150.