Zhongsheng Chen ORCID iD: 0000-0002-0828-2044

Combining sequence data from multiple studies: impact of analysis strategies on rare variant calling and association results

Zhongsheng Chen[1], Michael Boehnke [1,4], and Christian Fuchsberger [1,2,3,4]+

[1] Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI

[2] Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy

[3] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria

[4] These authors jointly supervised this work

+Correspondence: Christian.Fuchsberger@eurac.edu

Grant number: HG009976 (MB)

Running title: Combining sequence data

Abstract

Individual sequencing studies often have limited sample sizes and so limited power to detect trait associations with rare variants. A common strategy is to aggregate data from multiple studies. For studying rare variants, jointly calling all samples together is the gold standard strategy but can be difficult to implement due to privacy restrictions and computational burden. Here, we compare joint calling to the alternative of single-study calling in terms of variant detection sensitivity and genotype accuracy as a function of sequencing coverage and assess their impact on downstream association analysis. To do so, we analyze deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data on 2,250 individuals from the GoT2D study jointly and separately within five geographic cohorts.

For rare SNVs: (1) ≥97% of discovered SNVs are found by both calling strategies; (2) non-reference concordance with a set of highly accurate genotypes is ≥99% for both calling strategies; (3) meta-analysis has similar power to joint analysis in deep-coverage sequence data but can be less powerful in low-coverage sequence data. Given similar data processing and quality control steps, we recommend single-study calling as a viable alternative to joint calling for analyzing SNVs of all MAF in deep-coverage data.

Keywords: Sequencing studies, rare variants, joint analysis, meta-analysis

Introduction

Genome-wide association studies (GWAS) based on genotype arrays have identified thousands of common (minor allele frequency [MAF]>5%) genetic variants associated with a wide range of human diseases and traits (Hindorff et al., 2012). However, these common variants comprise only 10% of the ~84 million variant sites discovered in the human genome by the 1000 Genomes Project (2015) with the rest being low-frequency (MAF 0.5-5%; ~14%) and rare (MAF<0.5%; ~76%) variants that are less well captured by genotype arrays and subsequent genotype imputation (Zuk et al., 2014). With the advance of genome sequencing technology, we can now directly study the role of variants across the full allele-frequency spectrum. Although sequencing studies to date have reaffirmed and expanded on the common variant associations of array-based GWAS, the modest sample sizes of most sequencing studies to date have limited the discovery of rare and low-frequency variant associations (Fuchsberger et al., 2016; Auer et al., 2016; Luo et al., 2017).

To increase sample size, researchers often aggregate sequence data across multiple studies. To combine sequence data across studies, the gold standard strategy is to jointly call all samples together (Auer et al., 2016). This joint calling strategy increases the quality of variant calls and minimizes batch effects such as those due to different sequencing centers or platforms (Auer, et al., 2016). However, joint calling for sequence data can be difficult to implement due to restrictions on data sharing (Paltoo et al., 2014; Jiang et al., 2014) and the potentially heavy computation burden (Lek et al., 2016). An alternative strategy that adheres to privacy rules and mitigates computing load is single-study calling (Okada et al., 2018) in which variants are identified and genotypes called separately within each study and then combined through meta-analysis of study-level association statistics or joint analysis of pooled individual-level data (i.e. mega-analysis). Although single-study calling is easier to implement than the gold standard joint calling, there is a need to quantify the difference in calling results between these two strategies and assess how it affects downstream association analysis.

Past research has shown that meta-analysis of study-level association results is as statistically efficient as joint analysis of individual-level data for combining common-variant GWAS (Lin & Zeng, 2010). More recent research has extended methods for meta-analysis to sequencing studies for rare variants (Tang & Lin, 2015). However, this research only analyzes the relative power of joint and meta-analysis under a single-study calling strategy and does not consider the impact of joint calling on association results. In addition, sequencing studies often differ in sequencing coverage depending on project needs and goals. For example, deep-coverage sequencing results in improved genotyping accuracy, particularly for rare variants (Lee et al., 2014; Xu et al., 2017), while low-coverage sequencing results in more sequenced samples at the same cost (Li et al., 2011). Thus, there is also a need to compare rare variant association tests for joint and single-study calling under different sequencing coverage.

In this paper, we aim to quantify the difference between the gold standard joint calling and the alternative single-study calling strategies and assess their impact on association testing of rare single nucleotide variants (SNVs) in deep and low-coverage sequence data. Specifically, we compare variant detection and genotyping accuracy for joint and single-study callsets on deep-coverage whole exome sequence (WES) and low-coverage whole genome sequence (WGS) dataset from the Genetics of Type 2 Diabetes (GoT2D) study (Fuchsberger et al., 2016) using the GotCloud variant calling pipelines (Jun et al., 2015) at default settings. Then for each data type, we compare single-variant and gene-based association test results for rare SNVs between three types of joint and single-study strategies: 1) joint calling with joint analysis, 2) single-study calling with meta-analysis, and 3) single-study calling with mega-analysis.

Methods

Data description

We analyzed data on 2,250 individuals from the GoT2D study (Fuchsberger et al., 2016) for whom deep-coverage whole exome sequence (mean depth 82X), low-coverage whole genome sequence (mean depth 5X), and Illumina HumanOmni 2.5M array data were all available. Study participants came from five geographical regions: (1) Augsburg, Germany (n=193; KORA study), (2) the Botnia region of western Finland (n=303; DGI study), (3) Sweden (n=391; DGI study), (4) the United Kingdom (n=473; UKT2D study), and (5) Finland (n=890; FUSION study). For clarity, we will refer to the sample of 2,250 individuals as the “joint” cohort and the five subsets as the “single-study” cohorts (Figure 1).

DNA sample preparation and sequencing

DNA samples were processed at the Broad Institute (FUSION and DGI), Wellcome Trust Centre for Human Genetics (UKT2D), and Helmholtz Zentrum München (KORA). DNA samples were genome and exome sequenced using the Illumina GAII or HiSeq. 2000 sequencers. Sequence data were aligned to human reference genome version 19 (hg19) using Picard (DePristo et al., 2011) and BWA (Li & Durbin, 2009). Further details on data generation, processing, and quality control can be found in Fuchsberger et al. (2016).

Processed and filtered sequence reads for the joint and single-study cohorts were analyzed by the GotCloud and GATK (McKenna et al., 2010; Van der Auwera et al., 2013) variant calling pipelines according to the best practice workflows recommended by their developers at default settings. We restricted our analyses to chromosome 2 (~8% of the human genome) to reduce computational burden.

Whole-genome and exome sequence data processing: GotCloud and GATK pipeline

We called SNVs with GotCloud at default settings using processed BAM files (Figure 1). We used SAMtools pileup and glfFlex to generate genotype likelihoods for all samples in 5 Mb chromosomal segments. We then used a support vector machine classifier to filter out likely false-positive variant sites (Jun et al., 2015).

Adhering to the recommended GATK workflow, we “hard called” every variable site in each sample for the number of non-reference alleles (0, 1, or 2) using HaplotypeCaller in GVCF mode. To parallelize this step, we divided chromosome 2 into 5 Mb segments with 100 bp overlap and simultaneously carried out hard-calls within each segment. We merged intermediate genomic VCF (gVCF) files from each sample into batches of 100 samples with CombineGVCFs and then jointly genotyped them with GenotypeGVCFs. We used the GATK CatVariants tool to concatenate variant sets from all genomic regions to form a combined callset. We identified a set of high-quality variant calls from the raw variant callset using the Variant Quality Score Recalibration (VQSR) method which applies machine learning algorithms to score each variant call and filter them at a desired level of sensitivity. We used GATK VariantRecalibrator and ApplyRecalibration to filter the raw variant callset at the recommended tranche threshold of 99.9% which provides high sensitivity while maintaining a reasonable level of specificity. Finally, we removed indels from the filtered variant callset in keeping with our settings for the GotCloud pipeline and to focus on SNVs in subsequent analyses.

We used haplotype-based refinement to improve genotype and haplotype quality for whole genome genotype calls from both pipelines (Figure 1). Specifically, we used Beagle (Browning & Yu, 2009) to phase the genotype data in chunks of 10,000 SNVs with 1,000 SNVs overlaps and refined the phased sequences using Thunder (Jun et al., 2015) with 300 states.

We ran whole exome sequence reads through the GotCloud and GATK discovery pipelines under the same settings as the whole-genome data. We did not apply any refinement steps to the exome calls, consistent with standard practice for both pipelines for deep-coverage sequence data.

The final dataset for each of the four combinations of sequencing coverage (genome and exome) and pipeline (GotCloud and GATK) consists of a joint callset for all 2,250 samples, five separate single-study callsets for the geographically subdivided cohorts, and a union callset which merges the five single-study callsets. Since comparing the joint callset to five single-study callsets individually is difficult because detection of rare SNVs is heavily dependent on sample size and the results would be potentially skewed by the considerable sample size differences between cohorts, we use the union callset as an overall representation of single-study calling to provide a more apt comparison with the joint callset. For the union callset, we set genotype calls for SNVs not found in one or more of the single-study callset(s) as missing.

Non-reference genotype accuracy

For both pipelines, we assessed the accuracy of whole genome calls by comparing the Thunder-refined non-reference genotypes against a set of 192,322 variants of highly accurate ("high-confidence") genotypes determined through joint statistical analysis of deep-coverage (~82X) exome sequence and Illumina HumanOmni 2.5 array data in the GoT2D whole genome sequencing study (Fuchsberger et al., 2016). We assessed the accuracy of exome calls by comparing unrefined non-reference genotypes against the set of high-confidence genotypes from Illumina HumanOmni 2.5 array data.

Single-variant association analysis

We evaluated the impact of joint and single-study calling on single-variant association tests by comparing -$\log_{10}$p-values from joint analysis of the joint callset against those from meta-analysis of single-study summary statistics and joint analysis of the union callset (i.e. mega-analysis). In each single-study callset, we used the logistic score test to test for T2D association under an additive genetic model with the top two principal components as covariates (Figure 1). For meta-analysis, we combined summary-level results from the single-study callsets with fixed-effects sample-size weighted meta-analysis using METAL (Willer et al., 2010) and with trans-ethnic meta-analysis using MR-MEGA software (Mägi et al., 2017).

Gene-based association analysis

We used SKAT-O to test for association with multiple rare and low-frequency SNVs within coding regions of the genome. We prepared four lists of SNVs ("masks") based on MAF and functional annotation. For the creation of the masks, we considered a SNV to have MAF<1% if its MAF in every one of the single-study callsets is <1%. Mask 1 contained SNVs predicted to be protein-truncating, Mask 2 included all SNVs from Mask 1 together with missense SNVs with MAF<1%, Mask 3 included all SNVs from Mask 1 and those predicted to be deleterious by all five algorithms applied (Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster, and SIFT), and Mask 4 included all SNVs from Mask 1 and those predicted to be deleterious by at least one algorithm with MAF<1%.

We performed SKAT-O (Lee et al., 2012) analysis on the four masks separately within each single-study callset (Figure 1). We combined SKAT-O results from each single-study callset using Meta-SKAT-O test in the MetaSKAT R package (Lee et al., 2013) once assuming homogeneous genetic effects across single-study cohorts and again assuming heterogeneous genetic effects.

Results

Overview

We evaluated the utility of single-study calling as an alternative to the gold standard joint calling by comparing these methods in terms of variant detection, genotype accuracy, and impact on power of association tests for different sequencing coverage. For our analysis (restricted to chromosome 2 due to computational burden), we focus on the gold standard *joint callset*, which are calls from analyzing all 2,250 samples together (the "joint" cohort), the five *single-study callsets*, which are calls from the five geographically subdivided cohorts (the "single-study" cohorts: Germany, Botnia, Sweden, UK, Finland), and the *union callset*, which pools calls from the five single-study callsets. There are 25,689 deep-coverage WES SNVs and 2,101,401 (15,344 when restricted to coding regions) low-coverage WGS SNVs in the joint callset and 26,364 deep-coverage WES SNVs and 2,249,181 (16,457) low-coverage WGS SNVs in the union callset. We present only GotCloud results as we found choice of software pipelines (GotCloud or GATK) to have no meaningful impact on variant calling and association results.

Calling results

Union callset

The union callset pools calling results from the five single-study cohorts by merging their SNV calls. For SNV sites found in only a subset of the studies, we assign missing genotypes for studies in which the SNV site was not called. Using the union callset, we examine the overlap in variant detection between single-study cohorts. For deep-coverage data, 78% of all rare SNVs detected by single-study calling (i.e. those in the union callset) are "study specific" (Table 1), meaning they were found in only one of the single-study callsets and missing in all others, compared with 1.2% of low-frequency SNVs and 0.05% of common SNVs (Table 1). Conversely, only 2.3% of rare SNVs in the union callset are found in all five studies (Table 1) compared with 80% of low-frequency and 99% of common SNVs (Table 1). Similar numbers are seen for low-coverage data (restricted to coding regions) (Table 1). Overall, there are three possible reasons for a missing SNV site in a study: 1) the SNV was monomorphic in the study sample; 2) the variant caller did not have confidence to declare the SNV site; or 3) the SNV site was identified but removed by quality control as likely false-positive. However, for single-study calling, we are unable to differentiate between the three types of missingness because of privacy restrictions for individual-level data such as BAM files and calling results.

Variant detection: callset size

We evaluated variant detection for joint and single-study strategies by comparing the joint and union callsets across a range of MAFs. For low-frequency and common SNVs in both deep-coverage exome and low-coverage genome (restricted to coding regions) sequence data, there is almost complete overlap between the joint and union callsets (Figure 2C-F). However, for rare SNVs, there are noticeable discrepancies between the two callsets as described below.

The overwhelming majority of rare SNVs detected in deep-coverage data are found in both the joint and union callsets (97% of all rare SNVs) with the remaining SNVs found exclusively in the joint (0.1%) and union (2.9%) callsets (Figure 2A). Contrary to expectations, the union callset is larger than the joint callset, mainly due to inconsistencies in variant filtering. Of the 631 rare SNVs exclusive to the union callset, 540 of them were filtered out during joint calling and excluded from the final joint callset. SNVs in joint calling go through variant filters once whereas SNVs in single-study calling have one chance per study to pass filters and be included in the union callset. In this scenario, a lack of consistent variant filtering between joint and single-study calling can lead to the differences seen here.

For rare SNVs in low-coverage data (Figure 2B), we observed a similar pattern of variant detection as for deep-coverage data. However, inconsistencies in variant filtering only accounts for a small fraction of differences between the joint and union callsets. Only 128 of the 1,107 rare SNVs exclusive to the union callset were filtered out during joint calling.

Variant detection: genotype calls

In addition to comparing the number of SNVs detected by joint and single-study calling, we also compared the genotype calls made by the two strategies at different sequencing coverage. We show in Tables 2 and 3 the comparison of genotype calls between joint and the single-study calling for 9,096 rare SNVs found in the joint and union callsets from deep-coverage exome as well as from low-coverage genome (restricted to coding regions) sequence data. Genotype comparisons for 2,127 low-frequency and 2,027 common SNVs are shown in Supplementary Tables 1-4. Excluding missing calls, overall genotype discordance between joint and single-study calling is lower in deep-coverage data than in low-coverage data. Furthermore, for rare SNVs, 64% of all genotype calls from single-study calling in deep-coverage data (Table 2) are missing compared with 70% for low-coverage data (Table 3). Breaking down rare SNVs further by minor allele count (MAC), we observe this missingness to be a function of MAC in both types of sequencing data with the rarest categories most affected (Supplementary Tables 5-12). In deep-coverage data, we can attribute almost all missing calls for rare SNVs to monomorphic SNVs in the single-study cohort(s) since 13,093,060 of the 13,093,128 missing single-study calls were called as homozygous reference by joint calling (Table 2). Using the GATK pipeline, it is possible to identify monomorphic SNVs in gVCFs and assign homozygous reference genotypes to the 13,093,060 missing calls. However, we were unable to do this for the GotCloud pipeline since it does not support gVCFs. In low-coverage data, 6,365 of 14,246,613 missing single-study calls were called as non-reference by joint calling (Table 3) compared with 68 non-reference calls for deep-coverage data (Table 2). Since rare SNVs naturally have low allele counts to begin with, any small change to their overall allele counts will have a noticeable impact on association testing and other downstream analyses. Finally, the missingness appears to be mostly localized to rare SNVs as we observe only a slight number of missing genotype calls in low-frequency SNVs (4.3% in deep-coverage data, 9.2% in low-coverage data; Supplementary Tables 1 and 2) and a negligible number in common SNVs (0.21% and 0.78%; Supplementary Tables 3 and 4).

Genotype concordance

We assessed non-reference genotype accuracy (hereafter referred to as “genotype concordance”) of joint and single-study calling in deep-coverage exome sequence data by comparing non-reference calls for SNVs found in both the joint and union callsets against a “truth” set of high confidence genotypes from Illumina HumanOmni 2.5 array data (Fuchsberger et al., 2016). The joint and union callsets have nearly identical genotype concordance with the truth set for SNVs of all MAFs and negligible differences in raw counts (Table 4).

Next, we assessed genotype concordance for SNVs in low-coverage genome sequence data (not restricted to coding regions to preserve a meaningful number of comparisons) by comparing against high confidence genotypes from Illumina HumanOmni 2.5 array data and/or from deep (~82X) exome sequence in the GoT2D integrated panel (Fuchsberger et al., 2016). The joint callset correctly calls 0.4% more genotypes than the union callset for rare SNVs, 0.5% more for low-frequency SNVs, and 0.2% more for common SNVs (Table 4). Compared with deep-coverage data, here we observe a larger difference in genotype concordance with the truth set between the joint and union callsets. For example, the joint callset calls 13,322 more genotypes correctly (out of 3,575,402 total comparisons) than the union callset for rare SNVs in low-coverage data while it only calls 1 more genotype correctly (out of 91,756) for rare SNVs in deep-coverage data. As expected, the improvements to calling accuracy offered by larger sample sizes in the joint strategy are more pronounced when the average read coverage is low.

Effect of GC bias on genotype concordance

It is a well-known that sequencing read coverage tends to be lower in high GC-content regions. To investigate the effect of this GC bias on joint and single-study calling, we compared genotype concordance between the joint and union callset in regions of low GC-content (<60% of base pairs are GC) and in regions of high GC-content (≥60%) in chromosome 2. In low GC-content regions, we observe similar genotype concordance between the joint and union callset in both deep- and low-coverage sequence data (Supplementary Table 13). In high GC-content regions, we observe similar genotype concordance between the two callsets in deep-coverage data but notice larger differences in low-coverage data where the joint callset correctly calls 0.7% more genotypes than the union callset for rare and low-frequency SNVs (Supplementary Table 14). The performance of the two calling strategies in high GC-content regions are nearly equal in deep-coverage data but single-study calling can be slightly less accurate than joint calling in low-coverage data.

Association analysis

Overall, we observe similar p-values between joint analysis of the joint callset, fixed-effects meta-analysis of single-study summary statistics, and joint analysis of the union callset (mega-analysis) for rare SNVs in deep-coverage data (Figure 3A-C). This is due to almost perfect concordance in genotype calls between joint and single-study calling and the fact that missing variant calls for rare SNVs from single-study calling were almost all called as homozygous reference in the joint callset. However, for low-coverage data, we observe large discrepancies in p-values between joint and meta-analysis (Figure 3D) as well as between joint and mega-analysis for rare SNVs (Figure 3E). These differences in association results is caused by a combination of lower concordance in genotype calls between the two calling strategies for low-coverage data and an increase in the number of missing single-study calls being called as non-reference in the joint callset. Since both meta-analysis and mega-analysis use single-study calling, their association results are more similar (Figure 3F).

We evaluated association power between joint and single-study calling for gene-based tests by comparing -log$_{10}$p-values from SKAT-O test of the joint callset versus those from meta-analysis of single-study SKAT-O test results assuming homogeneous genetic effects. For all masks, SKAT-O based joint analysis and Meta-SKAT-O based meta-analysis produce similar p-values (Supplementary Figure 3).

Heterogeneity between single-study cohorts

To address possible heterogeneity in genetic effects between our single-study cohorts, we combined single-study summary statistics using a trans-ethnic meta-analysis implemented in MR-MEGA and combined single-study SKAT-O test results using Meta-

SKAT-O assuming heterogeneous genetic effects. For single-variant tests, we observe that trans-ethnic meta-analysis had slightly greater power to detect variants whose heterogeneity in genetic effects were correlated with ancestry compared with fixed-effects meta-analysis (Supplementary Figure 4). However, none of these variants are close to reaching genome-wide significance (p-value<5x10^{-8}) while those that are have more significant p-values under a fixed-effects meta-analysis. For gene-based tests, we observe slight variations in p-values between homogeneous and heterogeneous effect meta-analyses for Masks 1 and 3 but much greater p-value variability for Masks 2 and 4 (Supplementary Figure 5).

Discussion

Although jointly calling all samples together is the gold standard strategy for analyzing rare SNVs in sequencing studies, single-study calling is more appealing due to fewer privacy restrictions and smaller computation burden. In this study, we compared joint and single-study calling in terms of variant detection, non-reference genotype concordance, and their impact on association power as a function of sequencing coverage.

For single-study calling, we found that low overlap in variant detection among single-study cohorts for rare SNVs results in an abundance of “missing” genotype calls where we lose information for variant sites in cohorts where they were not detected. We show that for deep-coverage data, the impact of missing genotype calls on association testing of rare SNVs from single-study calling is minimal because almost all of this missingness is due to monomorphic SNVs, as evident by corresponding homozygous reference calls in the joint callset. However, for low-coverage data, average read depth is low and thus, a portion of the missing genotype calls may be due to lack of coverage at the variant sites (Xu et al., 2017). Indeed, we show that a fraction amount of missing single-study calls for rare SNVs in low-coverage data have corresponding non-reference calls in the joint callset, resulting in lower than expected allele counts and reduced power for association testing of these SNVs. In addition, these missing calls can have a negative impact on gene-based aggregation tests, which will be underpowered if too many variant sites within a gene have missing genotype calls, and genotype-based callbacks, since the majority of loss-of-function SNVs are rare. A possible, but resource-intensive solution is to generate a list of SNV sites based on the union callset and then go back and genotype these sites within each single-study cohort. With parallel computation for each sample and every 5 Mb chromosomal segment, this process takes on average one hour CPU-time per sample per cohort with a maximum memory usage of approximately 0.5 GB to re-call 1 to 1.2 million variants in chromosome 2.

Although the low overlap in variant detection among single-study cohorts for rare SNVs can arise naturally due to sample population differences between cohorts, another contributing factor is the inconsistency of variant calling filters (i.e. false-positive screening). In our analysis, rare SNVs that were filtered out during joint calling may pass filters during calling in some single-study cohorts while being filtered out in others. This increases the possibility of introducing false-positive SNVs to downstream analyses since they only need to pass filters in one of the single-study cohorts to be included in association tests.

Recommendations

For deep-coverage data, single-study calling and either meta-analysis or mega-analysis can be recommended as a viable alternative to joint calling and analysis for rare SNVs based on almost perfect concordance of genotype calls between the two calling strategies, comparable non-reference genotype concordance with an external truth set, and comparable association results. Furthermore, missing genotype calls in single-study calling for deep-coverage data can be assumed to be homozygous reference and attributed to monomorphic variant due to a matching homozygous reference call for their counterparts in the joint callset. When combining many smaller single studies, meta-analysis can be more conservative and less powerful than mega-analysis (Ma et al., 2013).

For low-coverage data or low-coverage regions in deep data, single-study calling cannot be recommended as a viable alternative to joint calling for rare SNVs. Discordance in genotype calls between the two calling strategies is approximately 150 times higher than that in deep-coverage data (0.09% versus 0.0006%) and combined with a sizable number of genotype calls in single-study calling being missing due to lack of coverage at variant sites, we observe large discrepancies in association results between the two calling strategies.

In general, for studying low-frequency and common SNVs, single-study calling can be used as an alternative to joint calling in both deep-coverage and low-coverage data (Supplementary Figures 1 and 2). The only exception is for studying low-frequency SNVs in low-coverage data (Supplementary Figure 1D-F) where there remain noticeable discrepancies in association results between joint and meta/mega-analysis, although less than that seen for rare SNVs in low-coverage data.

Comparison with GATK pipeline

In addition to the GotCloud pipeline, we ran our analyses with the widely used GATK pipeline at default settings. Choice of software pipeline had a limited impact on variant detection (Supplementary Figure 6) and genotype accuracy (Supplementary Table 15) with little to no impact on association results (Supplementary Figures 7-10). There is more overlap in detected SNVs between joint and single-study calling for the GotCloud pipeline in deep-coverage data and vice versa for the GATK pipeline in low-coverage data. The GotCloud pipeline was slightly more accurate in calling common and low-frequency SNVs; however, on average this difference amounts to less than 1.5% more correctly called non-reference genotypes.

Study limitations

Due to computation time and burden, we limited our study to SNVs in chromosome 2 and we were unable to compare joint and single-study calling strategies for indels. Additional SNVs from analyzing more chromosomes would be helpful in comparing association power between joint and single-study strategies for genome-wide significant (p-value<$5x10^{-8}$) rare SNVs. Currently, our single-variant and gene-based analysis of rare SNVs are centered on those with p-values≥$5x10^{-5}$ with limited information on rare SNVs near the genome-wide significance threshold.

Due to similar performances between the GotCloud and GATK pipelines, we only presented results for the GotCloud pipeline. One difference between the two pipelines is that GATK, in contrast to GotCloud, also supports the gVCF file format which eliminates almost all of the missing homomorphic reference genotype calls in the union callset by including calls on monomorphic SNVs in the single-study callset.

Summary

We show single-study calling to be a viable alternative to joint calling for deep-coverage sequence data but show them to have noticeable discrepancies in rare variant calling and association results for low-coverage sequence data.

Acknowledgements

We thank Laura Scott for helpful discussions, Hyun Min Kang and Yun Li for technical help with GotCloud, and the GoT2D consortium for allowing us to use their sequence data. This research was supported by the National Institutes of Health grants HG000376 and HG009976.

Conflict of interest

The authors declare that there is no conflict of interest.

Data availability

Genotypes and phenotypes from the WGS and WES panels are available at the European Genome-phenome Archive (EGA) and the database of Genotypes and Phenotypes (dbGAP). Details can be found in Flannick et al., 2017.

References

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68. *doi:*10.1038/nature15393

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs000840.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001093.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001095.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001096.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001097.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001098.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001099v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001100.v1.p1. *dbGAP*

Altshuler, D., Boehnke, M., McCarthy, M. & Florez, J. (2016). phs001102.v1.p1. *dbGAP*

Auer, P. L., Reiner, A. P., Wang, G., Kang, H. M., Abecasis, G. R., Altshuler, D., ... & Leal, S. M. (2016). Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. *The American Journal of Human Genetics*, 99(4), 791-801. *doi:*10.1016/j.ajhg.2016.08.012

Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, 85(6), 847-861. *doi:*10.1016/j.ajhg.2009.11.004

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), 491. *doi:*10.1038/ng.806

Duggirala, R. et al. (2016). phs000849.v1.p1. *dbGAP*

Flannick, J., Fuchsberger, C., Mahajan, A., Teslovich, T. M., Agarwala, V., Gaulton, K. J., ... & McCarthy, D. J. (2017). Sequence data and association statistics from 12,940 type 2 diabetes cases and controls. *Scientific data*, *4*, 170179. *doi*: 10.1038/sdata.2017.179

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., ... & Rivas, M. A. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614), 41. *doi:*10.1038/nature18642

Hindorff LA, MacArthur J, Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA. 2012. A catalog of published genome-wide association studies. NHGRI. Available at: www.ebi.ac.uk/gwas/diagram.

Jiang, W., Chen, S. Y., Wang, H., Li, D. Z., & Wiens, J. J. (2014). Should genes with missing data be excluded from phylogenetic analyses?. *Molecular Phylogenetics and Evolution*, 80, 308-318. *doi:*10.1016/j.ympev.2014.08.006

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome research*, gr-176552. *doi:*10.1101/gr.176552.114

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., ... & NHLBI GO Exome Sequencing Project. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2), 224-237. *doi:*10.1016/j.ajhg.2012.06.007

Lee, S., Teslovich, T. M., Boehnke, M., & Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics*, 93(1), 42-53. *doi:*10.1016/j.ajhg.2013.05.010

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1), 5-23. *doi:*10.1016/j.ajhg.2014.06.009

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285. *doi:*10.1038/nature19057

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14), 1754-1760. *doi:*10.1093/bioinformatics/btp324

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome research*. *doi:*10.1101/gr.117259.110

Lin, D. Y., & Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology*, 34(1), 60-66. *doi:*10.1002/gepi.20435

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & GoT2D Investigators. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*, 37(6), 539-550. *doi:*10.1002/gepi.21742

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., ... & Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human molecular genetics*, *26*(18), 3639-3650. *doi*:10.1093/hmg/ddx280

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. *doi:*10.1101/gr.107524.110

Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., ... & Suematsu, M. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature communications*, 9(1), 1631. *doi:*10.1038/s41467-018-03274-0

Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J. L., ... & Caulder, M. (2014). Data use under the NIH GWAS data sharing policy and future directions. *Nature genetics*, 46(9), 934. *doi:*10.1038/ng.3062

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R., Steinberg, J., Walter, K., ... & McCarthy, S. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *The American Journal of Human Genetics*, 100(6), 865-884 *doi:*10.1016/j.ajhg.2017.04.014

Tang, Z. Z., & Lin, D. Y. (2015). Meta-analysis for discovering rare-variant associations: statistical methods and software programs. *The American Journal of Human Genetics*, 97(1), 35-53. *doi:*10.1016/j.ajhg.2015.05.001

The European Genome-phenome Archive EGAS00001001459 (2016).

The European Genome-phenome Archive EGAS00001001460 (2016).

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & Banks, E. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1), 11-10. *doi:*10.1002/0471250953.bi1110s43

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190-2191. *doi:*10.1093/bioinformatics/btq340

Xu, C., Wu, K., Zhang, J. G., Shen, H., & Deng, H. W. (2017). Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. *Genetic epidemiology*, 41(3), 187-197. *doi:*10.1002/gepi.22015

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., ... & Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4), E455-E464.

*doi:*10.1073/pnas.1322563111

Figure 1: Workflow for variant calling and association analysis. Sequencing and alignment procedures are described in Fuchsberger et al., 2016. Haplotype-based refinement was only applied to low-coverage whole genome sequence data.

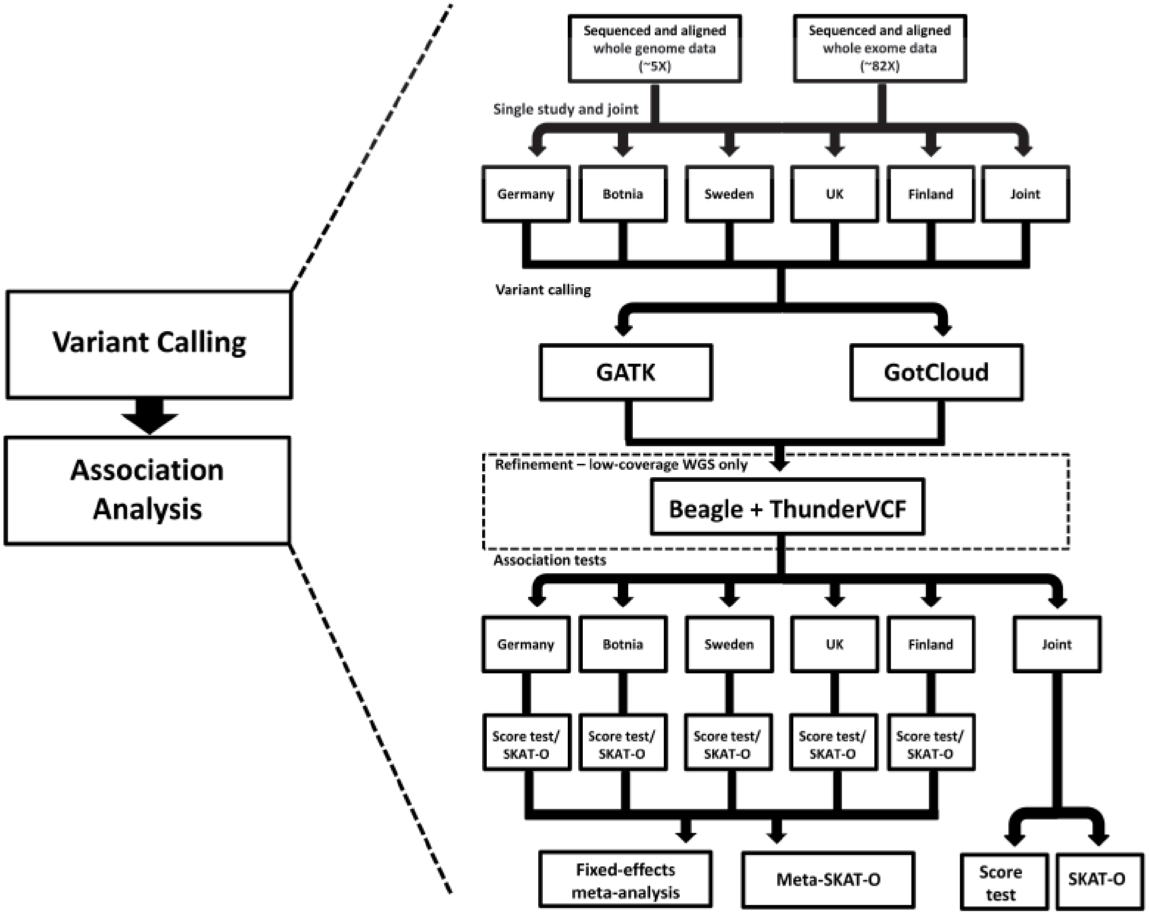

Figure 2: Comparison of variant detection between joint and single study calling strategies for rare (MAF<0.5%), low-frequency (MAF 0.5-5%), and common (MAF>5%) SNVs in deep-coverage (~82X) exome sequence data and low-coverage (~5X) genome sequence data restricted to coding regions.

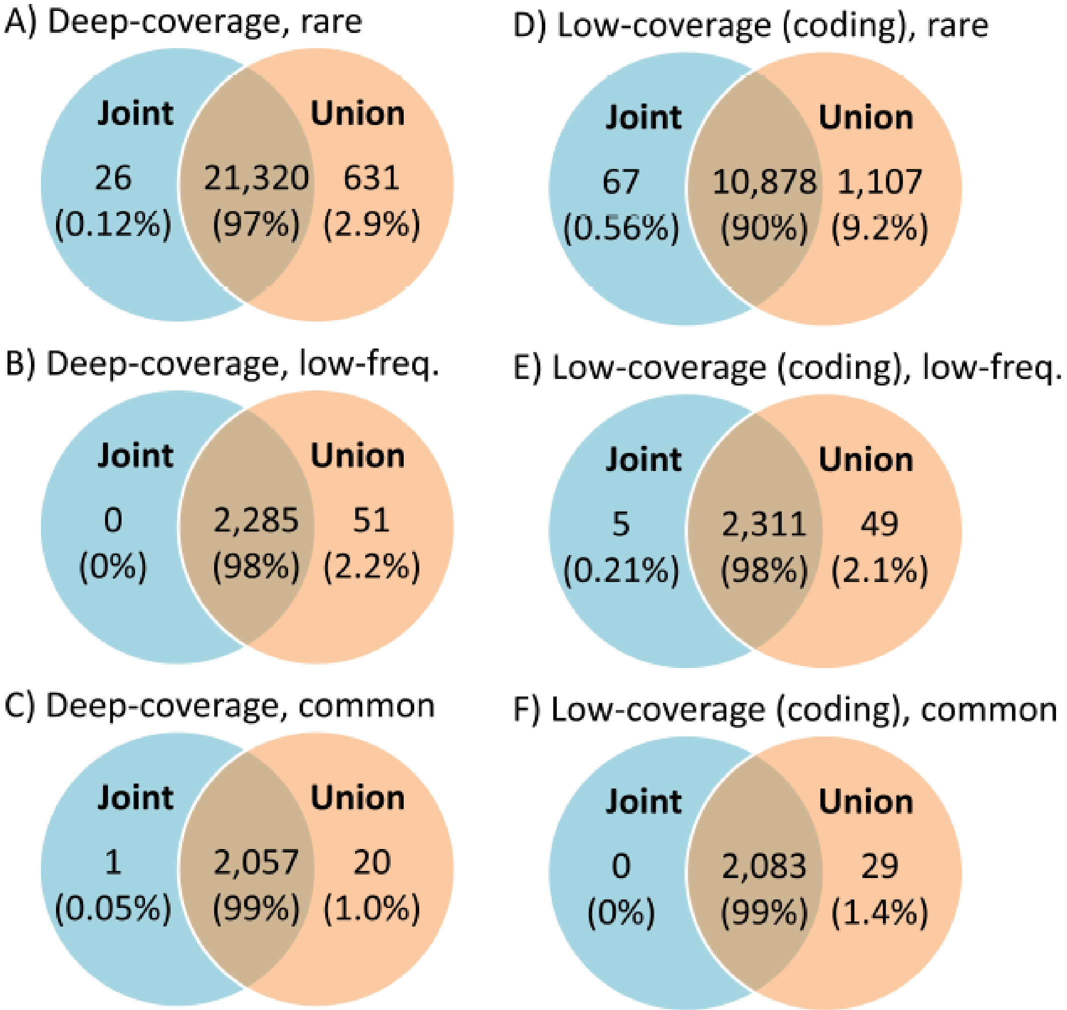

Figure 3: Comparison of single-variant association test p-values between joint and single study calling strategies for rare (MAF<0.5%) SNVs in (A-C) deep-coverage (~82X) exome sequence data and (D-F) low-coverage (~5X) genome sequence data. *Joint* refers to joint analysis of the joint callset, *meta* refers to fixed-effects meta-analysis of single-study summary statistics, and *mega* refers to joint analysis of the union callset (mega-analysis).

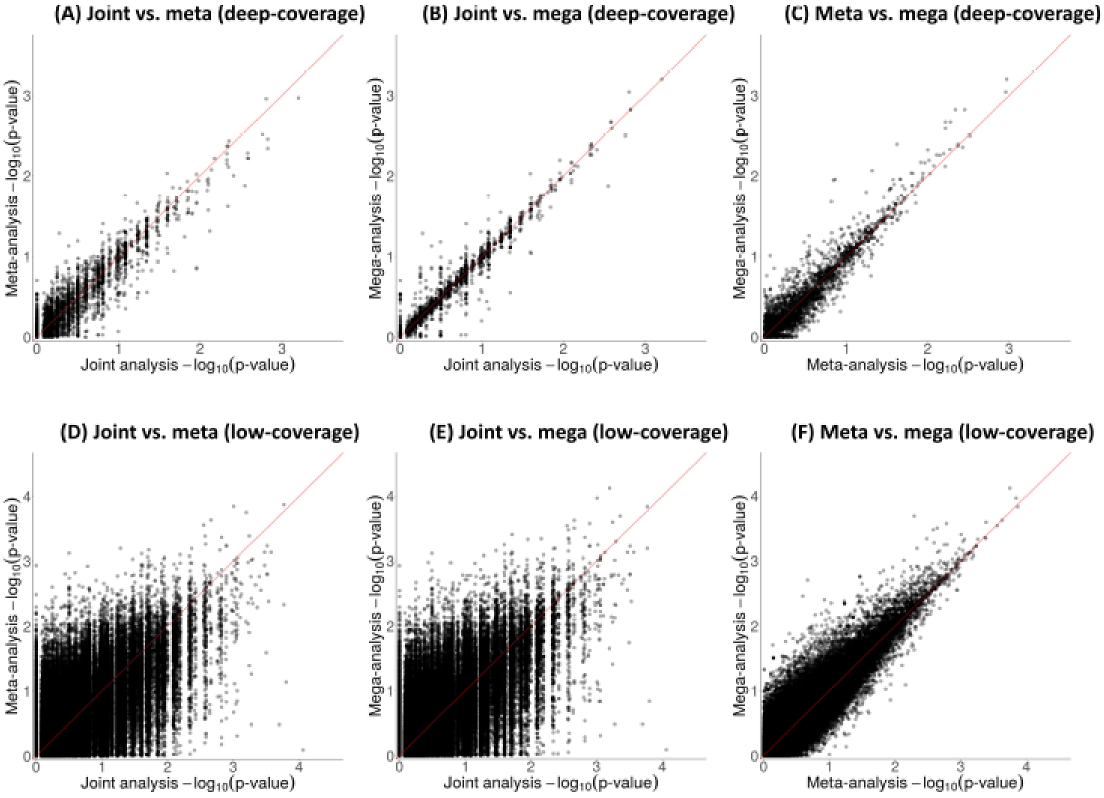

Table 1. Overlap in variant detection for the union callset

Data type	**Variants detected by only one study**	**Variants detected by 2 to 4 studies**	**Variants detected by all 5 studies**
Deep-coverage			
Rare (MAF <0.5%)	17,128 (78.0%)	4,316 (20%)	507 (2.3%)
Low-frequency (MAF 0.5-5%)	28 (1.2%)	435 (19%)	1,873 (80%)
Common (MAF >5%)	1 (0.05%)	26 (1.3%)	2,050 (99%)
Low-coverage (coding regions)			
Rare (MAF <0.5%)	9,262 (77%)	2,563 (21%)	160 (1.4%)
Low-frequency (MAF 0.5-5%)	38 (1.6%)	890 (38%)	1,432 (61%)
Common (MAF >5%)	5 (0.24%)	123 (5.8%)	1,984 (94%)

Note. The union callset pools variant calling results from the five single-study cohorts. Numbers in table refers to SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data and low-coverage (~5X) genome sequence data restricted to coding regions.

Table 2. Comparison of genotype calls for rare SNVs from deep-coverage exome sequence data

Single-study variant calling (union callset)	Joint variant calling (joint callset)				
	Missing	Homozygous reference	Heterozygous	Homozygous alternate	Total
Missing	0	13,093,060 (64%)	68 (0.00033%)	0	13,093,128 (64%)
Hom. ref.	0	**7,135,459 (35%)**	9 (0.000044%)	0	7,135,468 (35%)
Heterozygous	0	31 (0.00015%)	**25,862 (0.13%)**	0	25,893 (0.13%)
Hom. alt.	0	0	4 (0.000020%)	**211,507 (1.0%)**	211,511 (1.0%)
Total	0	20,228,550 (99%)	25,943 (0.13%)	211,507 (1.0%)	20,466,000 (100%)

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 9,096 rare (MAF <0.5%) SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Table 3. Comparison of genotype calls for rare SNVs from low-coverage genome sequence data (coding regions)

Single-study variant calling (union callset)	Joint variant calling (joint callset)				
	Missing	**Homozygous reference**	**Heterozygous**	**Homozygous alternate**	**Total**
Missing	0	14,240,248 (70%)	5,966 (0.029%)	399 (0.002%)	14,246,613 (70%)
Hom. ref.	0	**5,981,638 (29%)**	1,855 (0.009%)	2 (0.000010%)	5,983,495 (29%)
Heterozygous	0	3,687 (0.02%)	**21,073 (0.10%)**	99 (0.00048%)	24,859 (0.12%)
Hom. alt.	0	0	37 (0.00018%)	**210,996 (1.0%)**	211,033 (1.0%)
Total	0	20,225,573 (99%)	28,931 (0.14%)	211,496 (1.0%)	20,466,000 (100%)

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 9,096 rare (MAF <0.5%) SNVs from chromosome 2 in low-coverage (~5X) genome sequence data restricted to coding regions. Concordant calls between the two strategies are highlighted in bold.

Table 4. Non-reference genotype accuracy for joint and single-study calling strategies

Data type	**Genotype concordance for joint callset**	**Genotype concordance for union callset**
Deep-coverage		
Rare (MAF <0.5%)	99.7% (91,457/91,756)	99.7% (91,456/91,756)
Low-frequency (MAF 0.5-5%)	99.3% (171,939/173,131)	99.3% (171,930/173,131)
Common (MAF >5%)	99.3% (1,712,741/1,72,4873)	99.2% (1,711,385/1,724,873)
Low-coverage (all regions)		
Rare (MAF <0.5%)	99.7% (3,563,500/3,575,402)	99.3% (3,550,178/3,575,402)
Low-frequency (MAF 0.5-5%)	99.6% (6,837,310/6,866,584)	99.1% (6,807,530/6,866,584)
Common (MAF >5%)	99.6% (112,966,946/113,401,131)	99.4% (112,694,329/113,401,131)

Note. Genotype concordance for joint and single-study calling strategies in deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data. The “truth” set of high confidence genotypes being compared against comes from Illumina HumanOmni 2.5 array data and deep exome sequence in the GoT2D integrated panel. Raw genotype counts are displayed in parentheses.