HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Nat Genet. 2023 September ; 55(9): 1589–1597. doi:10.1038/s41588-023-01449-0.

GATK-gCNV enables discovery of rare copy number variants from exome sequencing data

Mehrtash Babadi[1,*,†], **Jack M. Fu**[2,3,4,†], **Samuel K. Lee**[1,†], **Andrey N. Smirnov**[1,†], **Laura D. Gauthier**[1], **Mark Walker**[1,3], **David I. Benjamin**[1], **Xuefang Zhao**[2,3,4], **Konrad J. Karczewski**[2,5,6], **Isaac Wong**[2,3], **Ryan L. Collins**[2,3], **Alba Sanchis-Juan**[2,3,4], **Harrison Brand**[2,3,4], **Eric Banks**[1], **Michael E. Talkowski**[2,3,4,5,6,*]

[1]Data Sciences Platform, Broad Institute, Cambridge, MA, USA

[2]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA

[3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[4]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA

[5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Abstract

Copy number variants (CNVs) are major contributors to genetic diversity and disease. While standardized methods, such as the Genome Analysis ToolKit (GATK), exist for detecting short variants, technical challenges have confounded uniform large-scale CNV analyses from WES data. Given the profound impact of rare and *de novo* coding CNVs on genome organization and human disease, we developed GATK-gCNV, a flexible algorithm to discover rare CNVs from sequencing read-depth information, complete with open-source distribution via GATK. We benchmarked GATK-gCNV in 7,962 exomes from individuals in quartet families with matched genome sequencing and microarray data, finding up to 95% recall of rare coding CNVs at a resolution of more than two exons. We used GATK-gCNV to generate a reference catalog of rare coding CNVs in WES 197,306 individuals in the UK Biobank, and observed strong correlations between per-gene CNV rates and measures of mutational constraint, as well as rare

[*]**Correspondence:** mehrtash@broadinstitute.org, talkowsk@broadinstitute.org.
[†]These authors contributed equally

Author contribution statement M.B., D.I.B, and S.K.L. developed and implemented the GATK-gCNV model and the inference algorithm. A.S. contributed model enhancements and developed sample-clustering and batch-processing workflows. X.Z., A.S., and J.M.F conducted benchmarking studies of GATK-gCNV performance. A.S., M.B, and S.K.L developed WDL workflows for Terra integration and scalable analysis. M.E.T, J.M.F., E.B, H.B., S.K.L., M.W., and L.D.G. supervised aspects of this project at various stages of development. J.M.F., R.L.C., H.B., and K.J.K. contributed to association analyses. I.W. and J.M.F. generated the CNV callsets. J.M.F., I.W., R.L.C., A.S.-J., and H.B. conducted quality-control on generated callsets. M.B., J.M.F., R.L.C., H.B., and M.E.T. wrote the manuscript, which was edited by all authors.

Competing Interests Statement
The authors declare no competing interests.

CNV associations with multiple traits. In summary, GATK-gCNV is a tunable approach for sensitive and specific CNV discovery in WES, with broad applications.

Introduction

Copy number variants (CNVs) comprise duplications and deletions of genomic segments spanning ≥50 nucleotides. These gains and losses of genetic material can impact gene function and regulation with profound consequences in human disease[1,2]. While each human genome likely harbors more than 25,000 structural variants[3], most gene-disrupting CNVs, including the vast majority of clinically interpretable pathogenic CNVs, experience strong negative selection and are therefore rare in the general population[4]. Thus, the ability to discover rare and *de novo* CNVs that alter protein-coding sequences with high recall and precision can have widespread utility in human genetic research, trait association, and clinical diagnostics.

The discovery of CNVs has historically relied upon low-resolution technologies like chromosomal microarray (CMA). Despite its technical limitations, exploration of large CNVs from CMA has provided substantial insights for many diseases and remains as the first-tier diagnostic test to ascertain CNVs in children with unexplained developmental disorders[5]. However, the low resolution of CMA precludes most gene- and exon-level interpretation of CNVs, while whole exome sequencing (WES) has become a first-tier screen for coding short variants, like single-nucleotide variants (SNVs) and small (<50bp) insertions or deletions (indels) [6]. WES has revolutionized human disease research and diagnostic screening in the protein-coding sequences while being substantially lower cost than whole genome sequencing (WGS)[7,8]. In theory, WES should permit the detection of most coding CNVs with equivalent recall to WGS and represent a marked improvement in resolution beyond CMA. In practice, variability in sequencing coverage due to hybridization-based exome enrichment[9] and other biases related to WES library preparation[10] distort informative read-depth signals depending on local sequence context and the properties of each individual sample. This technical variability has presented significant challenges in balancing recall of WES-based CNV discovery with the need for high precision in many applications. Existing methods for CNV detection in WES attempt to remove systematic biases and normalize read-depth data via PCA denoising[11], regression[12], pre-clustering of samples[13,14], or accounting for genomic context such as GC-content[15]. CNVs are then detected in a second step using hidden Markov models (HMM) or nonparametric change-point detection algorithms[16]. These methods introduce a lack of self-consistency between the removal of systematic biases and the CNV calling by carrying out these two steps separately, which can inadvertently remove informative CNV signals in the former and cause decreased recall in the latter.

The generation of WES data on millions of individuals to date[17–21] provides a unique opportunity for large-scale assessment of rare CNVs across human diseases and traits. Whereas the use of the Genome Analysis Toolkit (GATK) to capture SNVs and indels in WES is well-established and ubiquitous[22], the absence of a CNV discovery tool that can be routinely applied to WES data with comparable performance, documentation, and

dedicated support to GATK's functionality for SNV/indel analysis represents a significant barrier to realizing the full potential of WES data. Here, we present GATK-gCNV, a principled Bayesian method for learning global and sample-specific biases of read-depth data from large cohorts while simultaneously detecting CNVs. Our model combines a negative-binomial factor-analysis module for learning genome-wide latent factors of technical read-depth variation together with a hierarchical hidden Markov model (HHMM) for detecting singleton and rare CNVs in WES cohorts. In addition to being packaged as part of GATK, we also provide GATK-gCNV as a cloud-enabled tool in the Terra cloud platform (https://terra.bio) for easy adoption. We provide extensive benchmarking of GATK-gCNV against gold-standard WGS and CMA data in autism quartet families, and we demonstrate the scalable utility of GATK-gCNV by generating a reference map of rare genic CNVs in 197,306 WES samples from the UK Biobank. From these data, negative selection against coding loss-of-function (LoF) variants was strongly correlated with the rates of rare deletions and duplications of individual genes as expected. We also examined rare CNV trait associations in the UKBB. These results highlight that rare gene-disruptive CNVs can be routinely captured at very large-scale for low cost in WES-based association studies and diagnostic screening.

Results

Algorithm overview

We developed an algorithm, GATK-gCNV, to jointly discover and genotype CNVs across WES datasets using read-depth information (Fig. 1). While GATK-gCNV has also been optimized for similar analyses in WGS datasets[23], the analyses presented here focus on WES methods and applications where technical sources of read-depth variation pose a major hurdle to CNV detection. The algorithm is summarized here and provided in complete detail in Methods and Supplementary Note.

GATK-gCNV begins by calculating read counts over user-defined genomic target regions (*e.g.*, exons) in each sample while excluding regions with problematic sequence content. Next, samples with technically similar read-depth profiles are clustered into batches via principal components analysis (PCA) to reduce technical biases and improve computational efficiency. After clustering, the ploidy of every chromosome is estimated for each sample. Following preprocessing, GATK-gCNV performs read-depth denoising and CNV inference within a unified probabilistic model and determines CNV boundaries via the Viterbi algorithm (Supplementary Fig. 1). GATK-gCNV can be executed in two modes: "cohort-mode" and "case-mode". Cohort-mode uses all input samples to train a read-count model while simultaneously inferring CNVs, whereas case-mode applies a pre-trained model to call CNVs for any number of additional samples. Generating CNV calls through case-mode is much faster and cheaper, as it avoids the costly step of training a new read-count model. We sub-sample up to 200 samples of each PCA-defined batch to run cohort-mode, then apply case-mode to the remaining samples in each batch, greatly saving on cost.

Benchmarking GATK-gCNV using 7,962 deeply-profiled genomes

We assessed GATK-gCNV performance to discover rare and *de novo* CNVs on WES data (~75x coverage)[24] from the Simons Simplex Collection (SSC), which is a cohort of autism spectrum disorder (ASD) families that have undergone gold-standard CNV detection and rigorous quality control, including validation rate of 97% of *de novo* CNVs using five orthogonal technologies[25–27]. This dataset consists of samples with CMA (2,591 families[26]) and WGS (2,672 families[25,27]) data. In total, we assessed 7,962 WES samples with matched WGS data[25,27,28] and 7,636 samples with matched CMA data[26] (all CMA analyses were restricted to CNVs ≥50kb reflecting the lower resolution of CMA), including 3,131 parent-child trios (1,208 families included multiple offspring). When assessing recall, we defined a CNV from WGS or CMA to be captured by GATK-gCNV if at least 50% of well-captured intervals (defined below) spanning the variant were overlapped by WES CNV predictions in at least 50% of the same samples. For precision, we deemed a GATK-gCNV variant to have WGS support if 50% of the well-captured intervals of that variant were overlapped by a matching WGS CNV called in at least 50% of the same samples.

We applied GATK-gCNV to all SSC samples using the cloud-based Terra platform for biomedical research (http://terra.bio/) and have deployed a demonstration workspace as a resource (Methods). We implemented PCA-based sample batching based on a set of 7,981 curated intervals that differentiated common WES capture technologies (Fig. 2a,b, Methods). This approach subdivided the SSC WES samples into 14 batches of approximately 722 samples each (Interquartile range [IQR]=466; Fig. 2c). To further harmonize different exon-capture targets across studies, we restricted all analyses to protein-coding exons from canonical transcripts in GENCODE v33[29] and merged overlapping regions to derive a consensus set of 190,488 autosomal exons. We filtered out regions of extreme GC-content, repetitive sequence content, and poor mappability, and subdivided large exons to produce a final set of 330,526 intervals for CNV discovery (median size=384 bp, IQR=518; Methods). All analyses presented here were conducted with this set of intervals to ensure direct comparison. Within each PCA-defined batch of samples, we further filtered intervals based on low sequencing coverage (median <10 reads per sample). On average, this batch-specific coverage filtering retained 187,804 (IQR=55,732) intervals for CNV discovery per batch, corresponding to 169,442 exons on average (IQR=17,492). Hereafter, we refer to these intervals as “well-captured”.

We executed GATK-gCNV in cohort-mode on random subsets of 200 samples from each PCA-defined batch, training a CNV-discovery model tailored to each batch. GATK-gCNV cohort-mode ran for a median of 9:05 hours wall clock time to train and call each batch. For each batch, GATK-gCNV processed 12,500 intervals at a time across 14 parallel preemptible compute instances, each with 4 CPU cores and 24GB memory total, costing $0.037 per sample. Following training CNV discovery models for the sample batches, we conducted CNV discovery on all remaining samples using GATK-gCNV case-mode by batch, which required a median of 7:42 hours wall clock time and $0.021 per sample, again with every 200 samples running on an instance of 4 cores and 24GB memory. By leveraging the highly parallelized computing possible on cloud-based platforms like Terra, we processed 7,962 SSC samples in less than 24 hours of wall time at $0.026 per sample.

By design, the unfiltered output of GATK-gCNV is extremely sensitive to allow for exhaustive searches of candidate CNVs, producing an average of 6.3 rare (variant site frequency < 1%) CNV calls per sample (2.4 deletions and 3.9 duplications) at a resolution of more than two well-captured exons. At this resolution, the raw GATK-gCNV output achieved 95% recall in 7,962 SSC samples with matching WES and WGS data (Supplementary Fig. 2a, Table 1), but precision is low (22%). We developed a series of sample- and variant-level filters to define high-confidence CNVs for applications where high precision is critical, such as trait association studies or *de novo* CNV prediction. For variant-level filtering, we leveraged a quality metric (QS) emitted by GATK-gCNV for each CNV, which models the Phred-scaled probability that at least one interval within the CNV event locus was consistent with the estimated copy number state. We assigned a dynamic minimum QS threshold that scales with increasing CNV size, as described in Methods. For sample-level filtering, we found that the total number of CNV calls per sample correlated with the overall reliability and calibration of that sample, and that thresholds of >200 raw CNV calls or >35 CNVs with QS>20 were able to isolate and exclude poor-quality samples.

Applying these *post hoc* filters in the SSC WES data retained 89% (7,116/7,962) of all samples, yielding a callset of 9,246 autosomal CNV calls corresponding to 3,119 unique variants spanning more than two well-captured exons, or an average of 1.3 CNVs per sample (0.47 deletions and 0.83 duplications). In this high-quality callset, deletions had a median size of 6 exons and duplications a median size of 10 exons, while 72% of samples carried at least one such CNV (37% carried a deletion, and 55% carried a duplication). Benchmarking these high-quality CNV calls against matched WGS data revealed high precision (90%) with good recall (96% without WES filtering, 86% after WES filtering; Fig. 2d,e, Table 1, Supplementary Fig. 3). The QS threshold can be further raised for increased precision, where for example a threshold of QS>1000 produces extremely high precision (96%) for all CNVs, at the cost of reduced sensitivity (Supplementary Fig. 2c, Table 1). We also evaluated the performance of our high-quality GATK-gCNV callset versus rare CNVs (<1% site frequency) identified by CMA in 7,157 SSC samples for which we had matching ES and CMA data. After restricting to large (>50 kilobases & >2 exons), high-confidence CNVs from CMA (probability pCNV < 10^{-9} from Sanders et al.[26]), the high-quality GATK-gCNV callset achieved 97% recall (Supplementary Fig. 4). These benchmarks indicate that GATK-gCNV is sufficiently sensitive to displace CMA in diagnostic screening for protein-coding CNVs, with WES providing the added benefit of simultaneously capturing all coding SNVs and indels.

We next benchmarked the accuracy of our GATK-gCNV pipeline in identifying *de novo* CNVs in the offspring of SSC families. We predicted the transmission for each high-quality CNV identified in ES samples whose parents were both also present in our GATK-gCNV callset and identified 99 high-quality *de novo* CNVs (56 deletions and 43 duplications) among 3,097 children (mean = 0.032 *de novo* CNVs per child), which ranged in size from 3 to 667 exons (mean = 143 exons; median = 112 exons). We assessed the accuracy of these 99 *de novo* CNVs by comparing against matched WGS-derived *de novo* CNVs from the same samples[27]. We found that GATK-gCNV achieved 97% precision across all sizes of *de novo* CNVs, while maintaining 86% and 80% recall for 56 *de novo* deletions and 64 *de novo*

duplications in the gold-standard WGS dataset spanning more than 2 well-captured exons (Fig. 2f,g).

Finally, we compared GATK-gCNV results to four existing CNV tools: XHMM, CONIFER, cn.mops, and ExomeDepth. XHMM leverages a PCA denoising step followed by an HMM based calling step and was used to generate the largest publicly available exome-derived CNV reference to date[4,30]. CONIFER uses Singular Value Decomposition to normalize ES read-count variability followed by a threshold heuristic for CNV calling[31]. cn.mops uses a mixture Poisson model for read-depth denoising prior to segmentation[15]. ExomeDepth uses a sample-optimized panel of subjects to apply a beta-binomial model for read-depth denoising[32]. All evaluated CNV tools received as input the set of 330,526 intervals described above. We processed 96.3% (7,665/7,962 with accessible ES CRAMs) of the SSC samples using both XHMM and CONIFER, and all 7962 samples with cn.mops and ExomeDepth. Samples were analyzed across the same batches as in our GATK-gCNV implementation to minimize the impact of batching (Supplementary Note). Sample- and call-level filtering were conducted according to published best-practices, including the removal of low-quality samples, intervals, and calls. Using the set of high-quality samples (Supplementary Note) and evaluating on the basis of all unfiltered, non-overlapping GENCODE v33 exons, GATK-gCNV achieved recall and precision of 81% and 90%, respectively; XHMM 75% and 50%; CONIFER 47% and 49%; cn.mops 16% and 4% ; ExomeDepth 79% and 74%, all at a resolution of >2 exons (Fig. 2h,i). GATK-gCNV also generated copy number estimates ranging from 0–5, with 5 encompassing loci with copy state of at least 5. We validated these copy number estimates in the SSC using WGS data and found excellent accuracy, with 93% of GATK-gCNV estimated copy numbers within 0.2 copies of normalized WGS copy numbers (Supplementary Fig. 5a). Additionally, for 23/25 (92%) loci that harbor multiple copy number states across samples, GATK-gCNV was able to accurately ascertain all of the samples' different copy numbers within 0.2 copies compared to WGS normalized copy number (Supplementary Fig. 5b).

A rare CNV resource across 197,306 UK Biobank participants

Having established the accuracy of GATK-gCNV on >7,000 WES samples using gold-standard WGS and CMA matching callsets, we subsequently applied this method to two large cohorts to study the contribution of rare coding CNVs to human disorders and phenotypic variation. The first, which was recently published, analyzed and integrated such CNV data for more than 60,000 individuals in the study of ASD, significantly improving discovery of ASD associated loci[19]. Secondly, we applied GATK-gCNV to the UK Biobank (UKBB)[33] collection of more than 200,000 samples with WES data, where computational efficiency, cost, and performance are all important factors when conducting variant discovery.

The UKBB is one of the world's largest population-based biobanks with WES data linked to deep electronic health information. At the time of these analyses, a total of 200,624 WES samples from the UKBB were available to the research community. Several trait association studies have already been conducted from CMA and WES in these samples[17,34]. However, the patterns of rare coding CNVs in the UKBB at the resolution of individual exons and

genes remain unknown. We therefore sought to demonstrate the utility of GATK-gCNV by generating a uniform, high-quality rare CNV resource from the UKBB WES data.

We processed 200,624 UKBB exomes using GATK-gCNV with the method described above. Samples were clustered into 110 batches (median 1,687 samples per batch, IQR=1,300). We randomly selected 200 samples from each cluster to train a model in cohort-mode, with the remainder of samples in each cluster processed in matching case-mode, and applied the same sample- and variant-level filtering as used in the SSC cohort. The entire UKBB callset was processed in 60.05 hours of wall clock time, spread across 16,069 parallel CPU hours for 110 cohort-mode runs and 110 matching case-mode runs. The total cost to process all 200,624 samples was $6,423.44 ($0.032 per sample), including $1,002.43 for 22,000 samples in cohort-mode ($0.046 per sample) and $4,184.07 for 178,624 samples in case-mode ($0.023 per sample).

After applying all sample- and variant-level quality filters as described above, only 1.7% (3,318) of samples failed to meet our stringent sample-level thresholds. Across all 197,306 high-quality samples, we discovered 207,017 high-confidence rare CNV calls corresponding to 38,731 unique variants spanning >2 exons (Fig. 3a). Most samples (64%) carried at least one rare coding CNV: 31% and 49% of samples carried at least one rare deletion and duplication of >2 exons, respectively (Fig. 3b). As expected, we found that coding deletions were smaller on average (median size: 6 exons) than coding duplications (median size: 12 exons), which likely reflects a combination of stronger purifying selection on large coding deletions[28] and the comparatively higher technical difficulty for sensitive discovery of small duplications. We have returned these high-quality CNVs to the UKBB for dissemination to qualified researchers through the UKBB's data-release procedure.

In the absence of gold-standard WGS data on all UKBB samples, we assessed the quality of the UKBB CNV callset generated by GATK-gCNV versus existing UKBB CMA datasets[35]. First, we conducted systematic *in silico* confirmation of high-quality variants from GATK-gCNV using the Intensity Rank Sum (IRS) test from the GenomeSTRiP software package[13]. We applied the IRS test to 33,679 high-quality sites from GATK-gCNV that (i) overlapped at least 10 CMA probes and (ii) exhibited site frequencies between 0.01% and 1% in the subset of 177,158 UKBB samples that had matching WES and CMA data. For each variant, the IRS test determines if the raw CMA probe intensity rankings align with detected WES CNVs. This approach revealed that 95.7% of tested, high-quality CNVs from GATK-gCNV had orthogonal support from raw CMA intensity data at a nominal IRS p-value <0.01 (Fig. 3c, Supplementary Fig. 6). As a second, independent quality assessment of our WES-based UKBB CNV callset, we compared the rates of 49 genomic disorder (GD) CNVs[35]—large, disease-associated CNVs often formed by non-allelic homologous recombination—in our callset versus previously published rates from CMA analyses of the UKBB[35]. We found that the CNV frequency estimates at these 49 GD loci were highly concordant with prior CMA analyses (Fig. 3d, R^2=0.95; p=1.5×10^{-23}, Pearson correlation test).

We next assessed whether the rates of CNVs in the UKBB correlated with established metrics of negative selection against LoF variation. Several prior population-based studies have shown that negative selection against CNVs correlates with evolutionary constraint

against LoF variation[4,28], as measured by metrics like the LoF Observed over Expected Upper-bound Fraction (LOEUF[36]) from gnomAD or the probabilities of haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) recently proposed by a large-scale CNV meta-analysis[37]. Encouragingly, we observed severe depletion of high-quality deletions in our GATK-gCNV callset that overlapped constrained genes as measured by both LOEUF (Fig. 3e) and pHaplo (Supplementary Fig. 7), as well as strong linear relationships between the number of deletions observed in UKBB per gene (defined as >10% deletion of exonic base pairs) and the constraint scores of those genes in percentiles (Spearman's correlation=0.97 and =−0.90, respectively). Similarly, when examining the set of high-quality rare duplications from the GATK-gCNV callset, we found severe depletion in the number of CNVs that impact genes (defined as >75% duplication of exonic base pairs) to be triplosensitive by pTriplo (Fig. 3f, Spearman's correlation=−0.93), as well as a similarly strong linear relationship between the number of duplications and pHaplo (correlation=−0.85, Supplementary Fig. 8). Lastly, while the functional consequences of intragenic exonic duplications (IEDs) are context-specific and less readily predictable *in silico*, we nevertheless found depletion of putative IEDs correlating LOEUF (Fig. 3g, cor=0.33), consistent with previous observations in gnomAD[28].

Finally, as a demonstration of the utility of GATK-gCNV for trait association, we conducted a CNV-phenotype association analysis across 171,549 UKBB samples of European ancestry with high-quality CNVs for a curated set of 478 traits (median: 168,643 samples per trait)[38]. We tested each phenotype for association against deletions and duplications of genes and against 46 previously reported GD loci[19]. After restricting to sites with at least 5 overlapping CNVs, we applied the Sequence Kernel Association Test (SKAT[39]) and adjusted for the top 20 SNP-based principal components[38], sex, and age. At a conservative multiple-testing corrected threshold of 1.5×10^{-8} (Supplementary Note), we found 84 significant associations (Supplementary Table 1), including a recapitulation of known GD-phenotype associations, such as the canonical 16p11.2 deletions with body mass index[40] (BMI, $p=1.9\times10^{-17}$, Fig. 3h, Supplementary Table 1). Outside of known GD loci, we also identified associations in established pathogenic deletions, such as deletion of the hemoglobin gene cluster (encompassing *HBM, HBA2, HBA1, HBQ1*) which was previously associated with alpha thalassemia[41]. We also identified several gene-resolution CNV-phenotype associations, including recapitulation of an association between deletions overlapping *PDZK1* and urate levels[43] (Fig. 3i, $p=1.6\times10^{-15}$, Supplementary Table 1). We also recapitulated a dosage-dependent relationship between *CST3* copy number and cystatin C levels in blood[44] ($p=7.4\times10^{-17}$) as well as a corresponding decrease in estimated glomerular filtration rate (eGFR, $p=1.2\times10^{-21}$). The decrease in eGFR tracked with increasing copy number (Fig 3j, Supplementary Table 1), providing support for the validity of this association. Curiously, we observed that individuals carrying *CST3* duplications presented with eGFR comparable to individuals in the UKBB with documented renal failure (n=5,455), although none of the *CST3* duplication carriers themselves were documented as having any renal-related disease phenotypes. Knowledge of these duplications could be clinically significant for these patients, sparing them the stress and follow-up testing for kidney diseases indicative of decreased eGFR levels.

Discussion

Despite the widespread usage of WES in clinical and research applications, the overwhelming majority of research studies using WES have not evaluated or leveraged CNVs, which can decrease power for discovering novel disease-associated genes[45]. The advances in WES-based CNV discovery using GATK-gCNV presented here will provide significant added value to WES studies. We find that the recall and precision of GATK-gCNV to be tunable for use-cases ranging from association studies to sensitive diagnostic screening at a resolution of >2 exons when compared to gold-standard WGS CNVs. The critical feature of GATK-gCNV, which motivated its development, is its ability to maintain high accuracy for applications that require low false-positive rates, such as family-based research studies and clinical applications.

An advantage of GATK-gCNV is the ability to use a previously trained model (from the "cohort-mode") to call rare CNVs in other well-matched samples (in the "case-mode"), significantly saving time and resources. This feature could also be leveraged to analyze single samples, provided that a well-matched cohort-mode model can be found. Matching samples on exome probe hybridization kit design is critical to this process, while other factors such as sequencing center and platform are also important. This can be evaluated by repeating the PCA batching process to determine how close new samples lie to the cohort-mode training samples, while also ascertaining sample-level QC metrics on the number of CNV events to identify poorly-matched samples. By comparing the UKBB and ASD data studied in this manuscript, we readily observe appreciable differences in well-captured intervals and a larger degree of heterogeneity in the SSC data that was generated over a longer period of time from multiple platforms and sites relative to the UKBB dataset (Supplementary Fig. 9).

Despite the relative value added to standard WES applications, there remains several limitations for GATK-gCNV studies. First, GATK-gCNV performance decays rapidly for CNVs smaller than three exons. Single-exon analyses are routinely performed by visualization in many settings and single-exon CNVs (in particular deletions) are often readily accessible with such an approach, but the sensitivity against WGS would be insufficient for large-scale association studies. Extracting read-depth data at a higher resolution and incorporating statistics beyond read-depth in the GATK-gCNV probabilistic model may improve accuracy for smaller events in the future. Second, we have optimized GATK-gCNV for the detection of rare CNVs at a site frequency <1%; common CNVs (frequency >1%) can be assessed but it becomes challenging to disentangle true polymorphic CNVs segregating in the general population from the technical biases introduced by probe-based hybridization capture. For these variants, the performance of GATK-gCNV in the present implementation is lower than for rare CNVs (Supplementary Table 2). It is possible that these challenges may be mitigated in the future by incorporating prior weights on the distribution of population copy numbers at a given locus based on large, WGS-based databases of CNVs such as gnomAD[36] or the UKBB[33]. Additionally, there have been methods recently developed to nominate common copy number dosage associations with phenotypes using the UKBB exome data. In one example, the CNest workflow applies a linear model directly to normalized log-ratio read-depth data to discover

phenotypic associations in the UKBB data without directly performing CNV discovery and genotyping each individual sample.[46] This method captured many common CNV to phenotype associations that were previously tagged by SNPs and such approaches can provide complementary value to the rare CNV discovery and genotyping described herein for GATK-gCNV. Third, all WES-based analyses are necessarily restricted in resolution to well-captured exons, and thus CNV breakpoint resolution is variable depending on local gene density. It is possible that leveraging off-target reads[47], which are commonly found in WES daata and are presently ignored by GATK-gCNV, may serendipitously allow extending the detection range of CNVs beyond the exome.

The GATK-gCNV tool is fully accessible via the GATK software package (https://gatk.broadinstitute.org; default parameters Supplementary Table 3), where it can be deployed across local machines, high-performance enterprise computing clusters, and distributed cloud-computing environments (*e.g.*, Google Cloud Platform, Amazon Web Services, Microsoft Azure). In addition, GATK-gCNV is fully supported via the GATK User Forum, which provides tutorials and example cloud workspaces. Using GATK-gCNV is relatively cost-efficient per WES sample and could be further optimized through improved scaling using techniques such as amortized inference and subsampling. As a demonstration of the utility of GATK-gCNV, we applied it to 200,624 WES samples from the UKBB. These analyses serve to provide a resource of rare coding CNVs that we have released for use by the biomedical community (Data Availability). We demonstrated that patterns of CNV selection are in concordance with orthogonal genic constraint metrics in GATK-gCNV callsets, and as one initial exploration of the myriad potential uses of this rare CNV resource in the UKBB, we demonstrated correlations between rare coding CNVs and several traits. We anticipate that the dissemination of these methods and data resources will catalyze new discoveries and deepen our understanding of the contribution of rare coding CNVs to a wide range of human traits and disorders.

Methods

GATK-gCNV probabilistic approach to CNV detection

GATK-gCNV employs a generative model of sequencing read-depth data that accounts for both a) copy number variation and b) technical variation associated with differences in sample extraction, library preparation, enrichment, sequencing, and mapping. The method takes as input the read-depth data from a collection of samples over a set of genomic intervals, and learns to disentangle CNV events from technical factors by modeling both on an equal footing. Conceptually, disentanglement is made possible due to the discreteness and rarity of germline CNV events relative to the continuity and ubiquity of technical variation. Our proposed generative model consists of two main compartments, a model for read-depth

Code Availability
GATK-gCNV is distributed as part of the GATK jar release. For an example workspace on Terra, with recommended parameters, please see: https://app.terra.bio/#workspaces/help-gatk/Germline-CNVs-GATK4
GATK-gCNV evaluation and benchmarking code is available at: https://github.com/broadinstitute/gatk-gcnv-evaluation
CMA-CNV Validation code consists of: https://github.com/talkowski-lab/cnv-validation
GENOMESTRiP version 2.00.1982
MoChA version 2022-01-14 WDL https://software.broadinstitute.org/software/mocha/mocha.20220114.wdl

likelihood given the copy number states, and a hierarchical hidden Markov model (HHMM) that encodes copy number prior probabilities and state transitions along the genome. The read-depth likelihood compartment is a negative-binomial linear latent-factor model that accounts for technical read-depth variation in terms of a small number of learnable and predetermined bias factors over the genomic intervals. Individual samples share statistical power by determining the shared bias factors together. The copy number state HHMM compartment models the copy number structure, both at the level of individual samples and at the population level, and accounts for the genomic correlation of copy number states and the higher state-to-state transition rate within CNV loci that are determined to be polymorphic. Model parameters and latent variables, including copy number states and read-depth bias factors, are inferred simultaneously within a variational inference framework. We will describe the GATK-gCNV model and inference in the following sections. Further technical details, in particular those pertaining to implementation and the inference algorithm, are provided in the Supplementary Note.

Likelihood model for read depth conditioned on copy number

We consider the integer read-depth matrix n_{st}, with rows $s = 1,2,\ldots,S$ and columns $t = 1,2,\ldots,T$ denoting samples and genomic intervals, respectively (Supplementary Fig. 1a). Our goal is to model the conditional distribution $P(n_{st}|c_{st})$, where c_{st} is an integer copy number state matrix. At a fundamental level, the observed counts are typically obtained by sequencing a random subsample of the short-read hybrid capture-based library. As such, we expect random sampling noise (i.e., Poisson noise) to set the lower bound on the count dispersion. In practice, this fundamental noise is far outweighed by other sources of systematic noise, such as amplification artifacts and sequencing biases that are difficult to explicitly model. We take a data-driven approach and model n_{st} as a negative-binomial (NB) distributed random variable with rate $\lambda_{st} \geq 0$ and overdispersion $\Phi_{st} \geq 0$:

$$n_{st} \sim \text{Negative Binomial}(\lambda_{st}, \Phi_{st}) \tag{1}$$

In our choice of NB parameterization, $\mathbb{E}[n_{st}] = \lambda_{st}$ and $Var[n_{st}] = \lambda_{st} + \Phi_{st}\lambda_{st}^2$. Our general approach to modeling is to capture the generalizable patterns of read-depth variation in the NB rate λ_{st}, and to allow the NB overdispersion Φ_{st} to absorb the residual variance. To this end, we structure the NB rate λ_{st}, into multiplicative contributions arising from sequencing depth, copy number, and capture bias, as well as a small additive contribution from read-mapping errors:

$$\lambda_{st} = d_s c_{st} \mu_{st} + d_s \epsilon_M \tag{2}$$

where $d_s \sim \text{LogNormal}(\mu_d, \sigma_d)$ is the sample-specific sequencing depth with prior mean μ_d and standard deviation σ_d as model hyperparameters, c_{st} is the integer copy number matrix, ϵ_M is a small mapping-error rate hyperparameter, and μ_{st} is the multiplicative bias factor matrix. We model the latter as a low-rank linear latent-factor model with an exponential link function:

$$\log(\mu_{st}) = m_t + \sum_{v=1}^{D} W_{tv} z_{vs} + \sum_{v=1}^{K} \overline{W}_{tv} \bar{z}_{vs},$$

$$m_t \sim \mathrm{N}(0, \sigma_m),$$

$$W_{tv} \sim \mathrm{N}\left(0, \alpha_v^{-1}\right),$$

$$z_{vs} \sim \mathrm{N}(0, 1),$$

$$\bar{z}_{vs} \sim \mathrm{N}(0,1)\,. \tag{3}$$

Our model for the bias matrix μ_{st} comprises three terms: (I) The first term is an interval-specific mean bias m_t that is shared across all samples and has a normal prior with scale σ_m as a model hyperparameter; (II) The second term is a product of D learnable bias factors $W_{tv} \sim \mathrm{N}\left(0, \alpha_v^{-1}\right)$, $v = 1,2, \ldots, D$ and their corresponding sample-specific loadings $z_{vs} \sim \mathrm{N}(0,1)$. This structure can be thought of as a factor analysis sub-model. During model-fitting, all samples contribute to learning the same bias factors W_{tv}, whereas each sample uses ("loads") the factors to varying degrees. We set the number of bias factors D to an estimated upper bound, e.g. $D \sim 10 - 20$, and tune the prior scale of each factor α_v^{-1} to maximize model evidence. Known as automatic relevance determination (ARD), this empirical Bayes procedure shrinks the prior scale of unnecessary bias factors to zero and automatically selects the appropriate number of bias factors from the data. (III) Finally, the last term is a product of K predetermined bias factors $\overline{W}_{tv}$, $v = 1, 2, \ldots, K$ and their corresponding sample-specific loadings $\bar{z}_{vs}$. This provision allows us to explicitly include known read-depth bias factors into the model and accelerate model training. In practice, we found it beneficial to treat the GC-content of interval genomic intervals as predetermined bias factors. To this end, we set a lower and upper bound on the GC-content according to our interval filtering criteria and binned the allowed range uniformly into N_{GC} equally-sized bins. We determined the GC-content of each genomic interval t as a preprocessing step, constructed a mapping $\varphi_{\mathrm{GC}}: t \to 1, \ldots, N_{\mathrm{GC}}$ from each genomic interval to the best-matching GC-content bin, and set the GC bias factors as $\overline{W}_{tv} \equiv \delta(\varphi_{\mathrm{GC}}(t), v), v = 1, \ldots, N_{\mathrm{GC}}$. Intuitively, $\overline{W}_{tv}$ selects all genomic intervals with similar GC contents and the inferred sample-specific loadings $\bar{z}_{vs}$ can be thought of as the conventional "GC curves". We did not include any other hand-crafted bias factors in our implementation and therefore, $K = N_{\mathrm{GC}}$.

Finally, we allow the likelihood model to capture the variance that is not accounted for by the described bias-factor model using the NB overdispersion Φ_{st}. We propose the following parametric decomposition of the overdispersion into interval-specific and sample-specific contributions:

$$\log(1+\Phi_{st}) = \Psi_s + \Psi_t,$$
$$\Psi_s \sim \mathrm{Exp}(\sigma_S),$$
$$\Psi_t \sim \mathrm{Exp}(\sigma_T),$$

(4)

where σ_S and σ_T are model hyperparameters. The NB overdispersion can be thought of as a stopgap mechanism to prevent overfitting. Without this mechanism, model misspecification will lead either to learning non-generalizable bias factors, or worse, exploitation of the copy number state variables c_{st} as well as the genomic-region class τ_t latent variables (defined below) to account for the residual variance. Eq. (4) induces a heavy-tailed distribution over Φ_{st}, and this permissive prior allows the bias latent-factor model to "fail fast," effectively preventing overfitting and ultimately increasing the precision of the detected CNVs.

Hierarchical Hidden Markov Model for copy number states

We model the copy number state prior probabilities via a two-level hierarchical hidden Markov model (HHMM) as shown in Supplementary Fig. 1c. The top-level, primary Markov chain dictates the "class" of a genomic region as active (highly polymorphic) and/or silent (mostly copy-neutral); this binary determination, in turn, sets the prior probability and the state-to-state transition matrix of the secondary Markov chains. Active regions are given permissive copy number priors (i.e. uniform, Supplementary Fig. 1d), while silent regions have priors heavily weighted on the copy-neutral state (Supplementary Fig. 1e). Adjacent genomic regions are more likely to belong to the same region class, and we model this using a "sticky" region-to-region transition matrix.

The second hierarchy comprises a group of Markov chains, one for each sample, and conditionally independent of one another given top-level region-class variables. These secondary chains model the state-to-state copy number transitions along the genome separately for each sample. Again, genomic regions within a characteristic length scale tend to have similar copy number states, which we also model using a "sticky" copy number state-to-state transition matrix.

We describe both levels of the hierarchy in more detail in the following sections.

Top-level Markov chain: genomic-region classes—To model highly polymorphic (active) and mostly diploid (silent) genomic regions in a unified model, we introduce a per-interval binary random variable $\tau_t \in \{\text{active,silent}\}(t = 1, \ldots, T)$. The region class of the first interval $t = 1$ is sampled from a Bernoulli distribution, $\tau_1 \sim \text{Bernoulli}(\pi_{\text{region}})$, where:

$$\pi_{\text{region}}(\tau) = \begin{cases} p_{\text{active}} & \tau = \text{active}, \\ 1 - p_{\text{active}} & \tau = \text{silent}, \end{cases}$$

(5)

where p_{active} is a model hyperparameter. The region classes of subsequent loci are conditionally sampled according to the following transition matrix:

$$p(\tau_{t+1}|\tau_t) = \exp\left(-\frac{\Delta_{t,t+1}}{d_T}\right)\delta(\tau_t, \tau_{t+1}) + \left[1 - \exp\left(-\frac{\Delta_{t,t+1}}{d_\tau}\right)\right]\pi_{\text{region}}(\tau_{t+1}), \tag{6}$$

Where $\Delta_{t,t+1}$ is the genomic distance between the midpoints of region t and t+1, d_τ is a model hyperparameter that determines the typical correlation length of region classes, and $\delta(\tau_t, \tau_{t+1})$ is the Kronecker delta function, namely 1 if $\tau_t = \tau_{t+1}$, and 0 if $\tau_t \neq \tau_{t+1}$. Eq. (7) models the "sticky" behavior advertised earlier and is best understood by considering two limiting cases: (1) in the limit $\Delta_{t,t+1} \ll d_\tau$, we obtain $p(\tau_{t+1}|\tau_t) \approx \delta(\tau_t, \tau_{t+1})$, i.e. the next region inherits the state of the previous region; (2) in the limit $\Delta_{t,t+1} \gg d_\tau$, we obtain $p(\tau_{t+1}|\tau_t) \approx \pi_{\text{region}}(\tau_{t+1})$, i.e. the previous state is forgotten and the next region is sampled from the prior.

Secondary Markov chains: sample-specific copy number states —Given a determination of the genomic-region classes from the top-level chain, the copy number states of each sample (i.e. the rows of the copy number matrix, see Supplementary Fig. 1c) are independent of one another. We set an upper bound on the largest detectable copy number, C, as a model hyperparameter. We further assume being given a matrix of baseline copy number states for each sample and at each genomic region, κ_{st}. For a diploid organism, $\kappa_{st} = 2$ in the autosome (except for samples with aneuploidy) and $\kappa_{st} = 0, 1, 2$ for sex chromosomes (depending on the per-sample sex-chromosome ploidy). We interpret the copy number matrix c_{st} as a small perturbation of the baseline copy number matrix κ_{st}. We define the prior copy number distributions for the silent and active classes as follows:

$$\pi_{\text{silent}}(c)\Big|\kappa = \begin{cases} p_{alt} & c \neq \kappa, \\ 1 - C p_{alt} & c = \kappa, \end{cases}$$
$$\pi_{\text{active}}(c) = \frac{1}{C+1}\Big(\text{independent of } c\Big) \tag{7}$$

These priors are schematically shown in Supplementary Fig. 1d,e. Note that p_{alt} is another model hyperparameter that determines the permissiveness of having a non-baseline (e.g., non-copy-neutral) copy number state in silent regions. The prior distribution is assumed to be flat in active regions, that is, all $C + 1$ copy number states are assumed to be equally likely.

At the first interval $t = 1$, the copy number state in sample S is sampled from the prior:

$$c_{s1}|\tau_1 \sim \text{Categoritcal}(\pi_{\tau 1}). \tag{8}$$

For the subsequent intervals, the copy number state is sampled according to the following transition matrix:

$$p(c_{s,t+1}|c_{st},\tau_{t+1}) = \exp\left(-\frac{\Delta_{t,t+1}}{d_{\mathrm{CNV}}}\right)\delta(c_{st},c_{s,t+1}) + \left[1-\exp\left(-\frac{\Delta_{t,t+1}}{d_{\mathrm{CNV}}}\right)\right]\pi_{\tau_{t+1}}(c_{s,t+1}),$$

(9)

where $\Delta_{t,t+1}$ is the genomic distance between the midpoints of region t and $t+1$ as before, and d_{CNV} is a model hyperparameter that determines the typical correlation length of CNV events. Eq. (10) models the "sticky" behavior advertised earlier and is again best understood by considering two limiting cases: (1) in the limit $\Delta_{t,t+1} \ll d_{\mathrm{CNV}}$, we obtain $p(c_{s,t+1}|c_{st},\tau_{t+1}) \approx \delta(c_{st},c_{s,t+1})$, i.e. the next region inherits the copy number state of the previous region; (2) in the limit $\Delta_{t,t+1} \gg d_{\mathrm{CNV}}$, we obtain $p(c_{s,t+1}|c_{st},\tau_{t+1}) \approx \pi_{\tau_{t+1}}(c_{s,t+1})$, i.e. the previous state is forgotten and the next region is sampled from the prior.

Determining chromosomal baseline copy number states

The generative model for copy number states requires the knowledge of the chromosomal baseline copy number matrix κ_{st} for each sample $S = 1,\ldots,S$ at genomic interval $t = 1,\ldots,T$. By definition, the baseline copy number is the most prevalent copy number state at the scale of chromosomes (e.g., 2 for diploid, 3 for trisomy, etc.), and its determination serves to unify the treatment of diploid and aneuploid samples, as well as sex chromosomes in mixed-sex sample cohorts. All genomic regions belonging to the same chromosome $j = 1,\ldots,J$ have the same baseline copy number and therefore, it is sufficient to determine a copy number matrix, κ_{sj}, at the resolution of chromosomes instead of fine-grained genomic intervals. We define the per-chromosome read-depth as:

$$n_{sj} = \sum_{t \in \mathrm{chr}j} n_{st},$$

(10)

and like before, model it as negative-binomial distribution:

$$\begin{aligned}
&n_{sj} \sim \mathrm{NegativeBinomial}(\lambda_{sj}, \Phi_{sj}),\\
&\lambda_{sj} = (1-\epsilon_M)\frac{T_j m_j \kappa_{sj}}{\sum_{j'=1}^{J} T_{j'} m_{j'} \kappa_{sj'}} n_s + \epsilon_j n_s,\\
&\kappa_{sj} \sim \mathrm{Categorical}(\pi_{\mathrm{ploidy}}),\\
&\log(1+\Phi_{sj}) = \Psi_s + \Psi_j,\\
&m_j \sim \mathrm{PositiveNormal}(1,\sigma_m),\\
&\Psi_s \sim \mathrm{Exp}(\sigma_S),\\
&\Psi_j \sim \mathrm{Exp}(\sigma_J).
\end{aligned}$$

(11)

Here, $T_j = |\{t : t \in \mathrm{chr}j\}|$ is the number of genomic intervals spanning chromosome j, ϵ_M is a mapping error rate, $n_s = \sum_t n_{st}$ is the sample-wide total read-depth, and $\epsilon_j = \epsilon_M T_j / \sum_{j'=1}^{J} T_{j'}$ is the fractional mapping error rate for chromosome j. The multiplicative bias m_j accounts for chromosome-to-chromosome bias in read-depth. Like the fine-grained read-depth model, we account for the unexplained chromosome-scale read-depth variance as a sum of sample-

specific Ψ_s and chromosome-specific Ψ_j contributions. Finally, π_{ploidy} is the chromosome-scale ploidy prior.

GATK-gCNV model fitting using variational inference

The structured Bayesian model we described above captures key aspects of the phenomenology of sequencing read-depth variation and germline CNV events in a unified manner. However, the complexity of this hierarchical model and the lack of simplifying Bayesian conjugacy relationships implies that an exact inference algorithm is likely to be out of reach. Practical approximate-inference strategies include sampling-based Markov chain Monte Carlo (MCMC) methods and variational inference (VI). Here, we pursue VI as a more attractive option for the following reasons: (I) VI typically allows faster convergence times compared to MCMC-based strategies; (II) the flexibility of VI allows us to perform exact inference on certain sectors of the model (i.e., copy number HMMs); (III) recent advances in machine-learning software and probabilistic programming languages (PPLs) allow us to perform automated VI over the continuous sector of the model (i.e., the read-depth likelihood compartment) with little effort. We describe the details of our variational-inference approach in Supplementary Note. Operationally, we adopt a mean-field approximation and neglect posterior correlations between continuous latent variables $Z_{\text{continuous}}$ (e.g., sequencing depth, bias factors, loadings, etc.) and discrete latent variables Z_{discrete} (e.g., copy number states and genomic-region class indicators). We further assume a fully-factorized mean-field posterior for $Z_{\text{continuous}}$ and neglect posterior correlations between top-level and secondary Markov chains in the HHMM compartment. We leverage the PyMC3[48] PPL to perform incremental variational updates of the continuous posterior. These updates are interleaved with updates of the discrete posterior distributions, which are made tractable by exploiting the emergent linear conditional random field (CRF) structure that follows from mean-field factorization. An annealed entropy-regularization strategy is used throughout to avoid poor local minima in the early stages of model fitting, and convergence is assessed by testing the stability and self-consistency of variational posteriors within specified error tolerances.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Lee Lichtenstein, Yossi Farjoun, Benjamin Neale, and Niall Lennon for insightful discussions at various stages of this project, and Shadi Zaheri for carefully reviewing and providing feedback on the manuscript.

This work was supported by grants from the Simons Foundation for Autism Research Initiative (#573206); the SPARK project and SPARK analysis projects (#606362 and #608540); the National Institutes of Health (MH115957, HD081256, HD105266, HD099547, HD104224, HG008895, HG011755, MH123155, and HG011450). J.M.F. was supported by an Autism Speaks Postdoctoral Fellowship and R.L.C. was supported by NSF GRFP #2017240332.

Data Availability

The SSC benchmarking raw sequencing data can be accessed through NHGRI AnVIL; accession ID: phs000298; Databank URL: https://anvilproject.org/data. SSC CNVs can be accessed through SFARIBase(base.sfari.org), Accession IDs: SFARI_DS340921 (CNVs). Approval by the Simons Foundation for Autism Research Initiative (SFARI) is required.

Access to the UK Biobank raw sequencing data and the CNV data generated here will be provided by the UK Biobank.

GENCODE V33 annotation can be found at https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_33/gencode.v33.annotation.gtf.gz

References

1. Marshall CR et al. Structural variation of chromosomes in autism spectrum disorder. Am. J. Hum. Genet 82, 477–488 (2008). [PubMed: 18252227]
2. Egolf LE et al. Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma. Am. J. Hum. Genet 105, 658–668 (2019). [PubMed: 31474320]
3. Ebert P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372, (2021).
4. Ruderfer DM et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat. Genet 48, 1107–1111 (2016). [PubMed: 27533299]
5. Miller DT et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet 86, 749–764 (2010). [PubMed: 20466091]
6. Srivastava S. et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet. Med 21, 2413–2421 (2019). [PubMed: 31182824]
7. Gnirke A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol 27, 182–189 (2009). [PubMed: 19182786]
8. Ng SB et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276 (2009). [PubMed: 19684571]
9. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA & Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum. Mutat 36, 815–822 (2015). [PubMed: 25973577]
10. Benjamini Y. & Speed TP Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40, e72 (2012). [PubMed: 22323520]
11. Fromer M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet 91, 597–607 (2012). [PubMed: 23040492]
12. Jiang Y, Oldridge DA, Diskin SJ & Zhang NR CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 43, e39 (2015). [PubMed: 25618849]
13. Handsaker RE et al. Large multiallelic copy number variations in humans. Nat. Genet 47, 296–303 (2015). [PubMed: 25621458]
14. Packer JS et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics 32, 133–135 (2016). [PubMed: 26382196]
15. Klambauer G. et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 40, e69 (2012). [PubMed: 22302147]
16. Olshen AB, Venkatraman ES, Lucito R. & Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572 (2004). [PubMed: 15475419]

17. Backman JD et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599, 628–634 (2021). [PubMed: 34662886]
18. Lek M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016). [PubMed: 27535533]
19. Fu JM et al. Rare coding variation illuminates the allelic architecture, risk genes, cellular expression patterns, and phenotypic context of autism. bioRxiv (2021) doi:10.1101/2021.12.20.21267194.
20. Singh T. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 604, 509–516 (2022). [PubMed: 35396579]
21. Flannick J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature 570, 71–76 (2019). [PubMed: 31118516]
22. McKenna A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010). [PubMed: 20644199]
23. Byrska-Bishop M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185, 3426–3440.e19 (2022). [PubMed: 36055201]
24. De Rubeis S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215 (2014). [PubMed: 25363760]
25. Werling DM et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet 50, 727–736 (2018). [PubMed: 29700473]
26. Sanders SJ et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87, 1215–1233 (2015). [PubMed: 26402605]
27. Belyeu JR et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. Am. J. Hum. Genet 108, 597–607 (2021). [PubMed: 33675682]
28. Collins RL et al. A structural variation reference for medical and population genetics. Nature 581, 444–451 (2020). [PubMed: 32461652]
29. Frankish A. et al. GENCODE 2021. Nucleic Acids Res. 49, D916–D923 (2021). [PubMed: 33270111]
30. Fromer M. & Purcell SM Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr. Protoc. Hum. Genet 81, 7.23.1–21 (2014).
31. Krumm N. et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 22, 1525–1532 (2012). [PubMed: 22585873]
32. Plagnol V. et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28, 2747–2754 (2012). [PubMed: 22942019]
33. Sudlow C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015). [PubMed: 25826379]
34. Canela-Xandri O, Rawlik K. & Tenesa A. An atlas of genetic associations in UK Biobank. Nat. Genet 50, 1593–1599 (2018). [PubMed: 30349118]
35. Owen D. et al. Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. BMC Genomics 19, 867 (2018). [PubMed: 30509170]
36. Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). [PubMed: 32461654]
37. Collins RL et al. A cross-disorder dosage sensitivity map of the human genome. Cell 185, 3041–3055.e25 (2022). [PubMed: 35917817]
38. Pan UKBB. https://pan.ukbb.broadinstitute.org.
39. Wu MC et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet 89, 82–93 (2011). [PubMed: 21737059]
40. Auwerx C. et al. The individual and global impact of copy-number variants on complex human traits. Am. J. Hum. Genet 109, 647–668 (2022). [PubMed: 35240056]
41. Tamary H. & Dgany O. Alpha-Thalassemia. in GeneReviews® (eds. Adam MP et al.) (University of Washington, Seattle, 2005).

42. Sabath DE et al. Characterization of Deletions of the HBA and HBB Loci by Array Comparative Genomic Hybridization. J. Mol. Diagn 18, 92–99 (2016). [PubMed: 26612711]
43. Anzai N. et al. The multivalent PDZ domain-containing protein PDZK1 regulates transport activity of renal urate-anion exchanger URAT1 via its C terminus. J. Biol. Chem 279, 45942–45950 (2004). [PubMed: 15304510]
44. Sinnott-Armstrong N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet 53, 185–194 (2021). [PubMed: 33462484]
45. Fu JM et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet 1–12 (2022). [PubMed: 35022602]
46. Fitzgerald T. & Birney E. CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. Cell Genomics 2, 100167 (2022). [PubMed: 36779085]
47. Laver TW et al. SavvyCNV: Genome-wide CNV calling from off-target reads. PLoS Comput. Biol 18, e1009940 (2022).
48. Salvatier J, Wiecki TV & Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science vol. 2 e55 Preprint at 10.7717/peerj-cs.55 (2016).

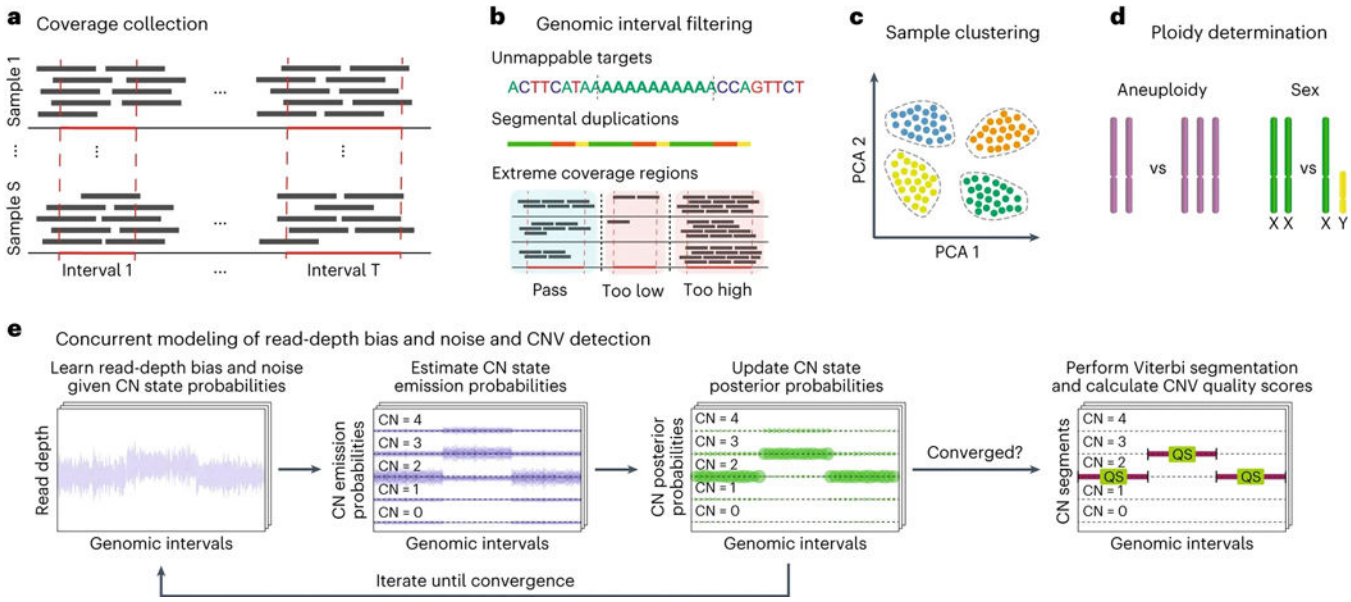

Fig. 1.

GATK-gCNV pipeline steps. **a,** Coverage information is collected from genome-aligned reads over a set of predefined genomic intervals. **b,** The original interval list is filtered to remove coverage outliers, unmappable genomic sequence, and regions of segmental duplications. **c,** Samples are clustered into batches based on read-depth profile similarity and each batch is processed separately. **d,** Chromosomal ploidies are inferred using total read-depth of each chromosome. **e,** The GATK-gCNV model learns read-depth bias and noise and iteratively updates copy number state posterior probabilities until a self-consistent state is obtained; after convergence, constant copy number segments are found using the Viterbi algorithm along with segmentation quality scores.

Abbreviations: CN - copy number; QS - quality score.

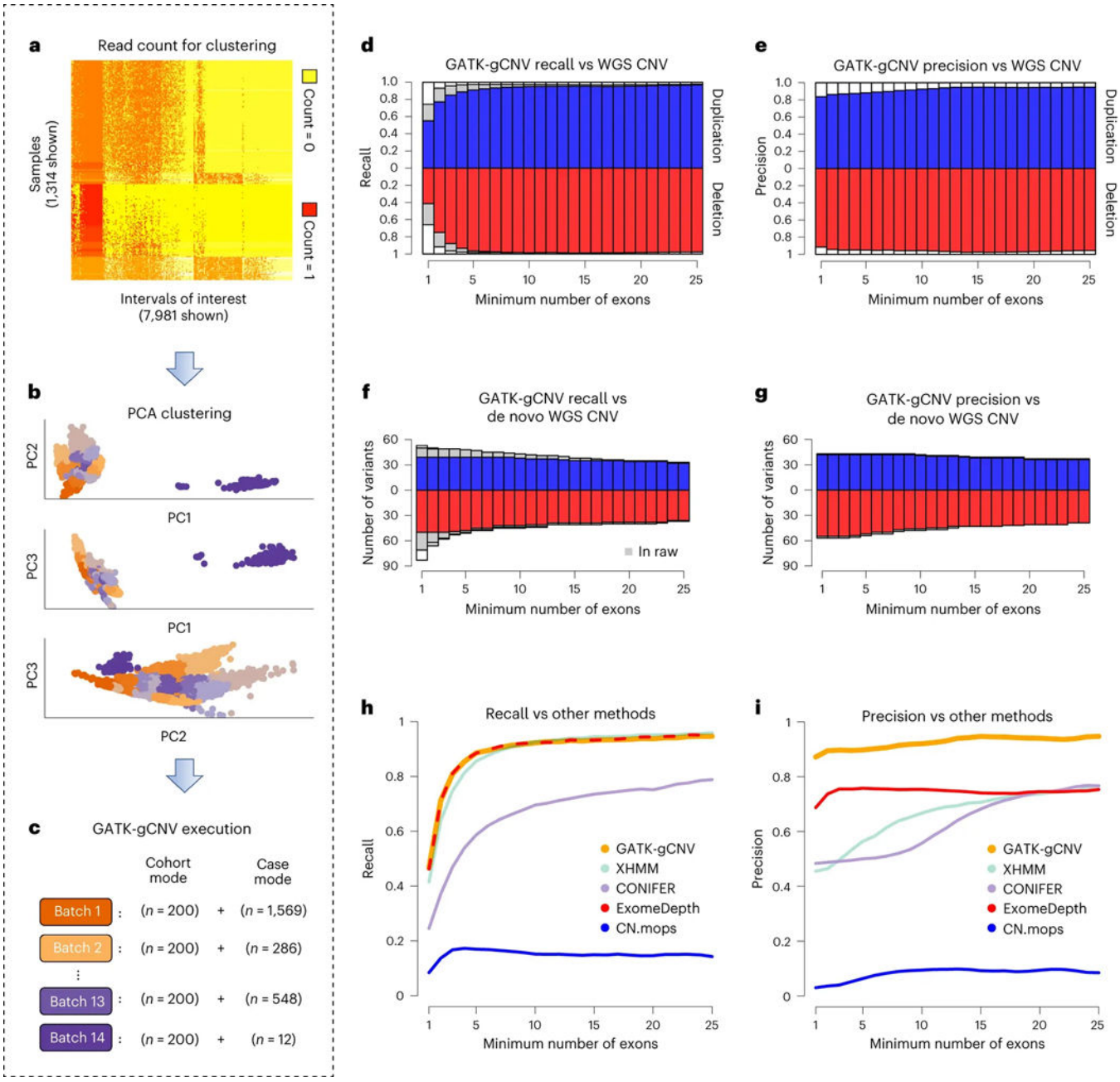

Fig. 2.

Calling and benchmarking of GATK-gCNV callset in a cohort of more than 7,000 samples with matching deep WGS sequencing. **a,** A heatmap illustration of the distinct read-count signal of the 7,981 intervals chosen for the batch creation procedure. **b,** After normalizing for median read count, the first three PCs are clustered to determine which samples will be processed together with GATK-gCNV, colored by the assigned batch. **c,** For each of the 14 batches generated, a random subset of 200 samples was chosen to generate a read-count model using cohort-mode; the remaining samples were processed in case-mode. **d,** The recall (and **e,** precision) of rare CNVs in GATK-gCNV ES CNVs compared to WGS gold-standard CNVs as a function of the number of exons the variant spans. **f,** The recall

(and **g,** precision) of de novo CNVs in GATK-gCNV compared to gold-standard WGS CNVs as a function of the number of exons. **h,** The recall (and **i,** precision) of rare CNVs in GATK-gCNV, XHMM, CONIFER, cn.mops, and ExomeDepth WES CNVs compared to WGS gold-standard CNVs as a function of the number of exons the variant spans. **Abbreviations**: PCA - principal component analysis; WES - exome sequencing WGS - whole genome sequencing.

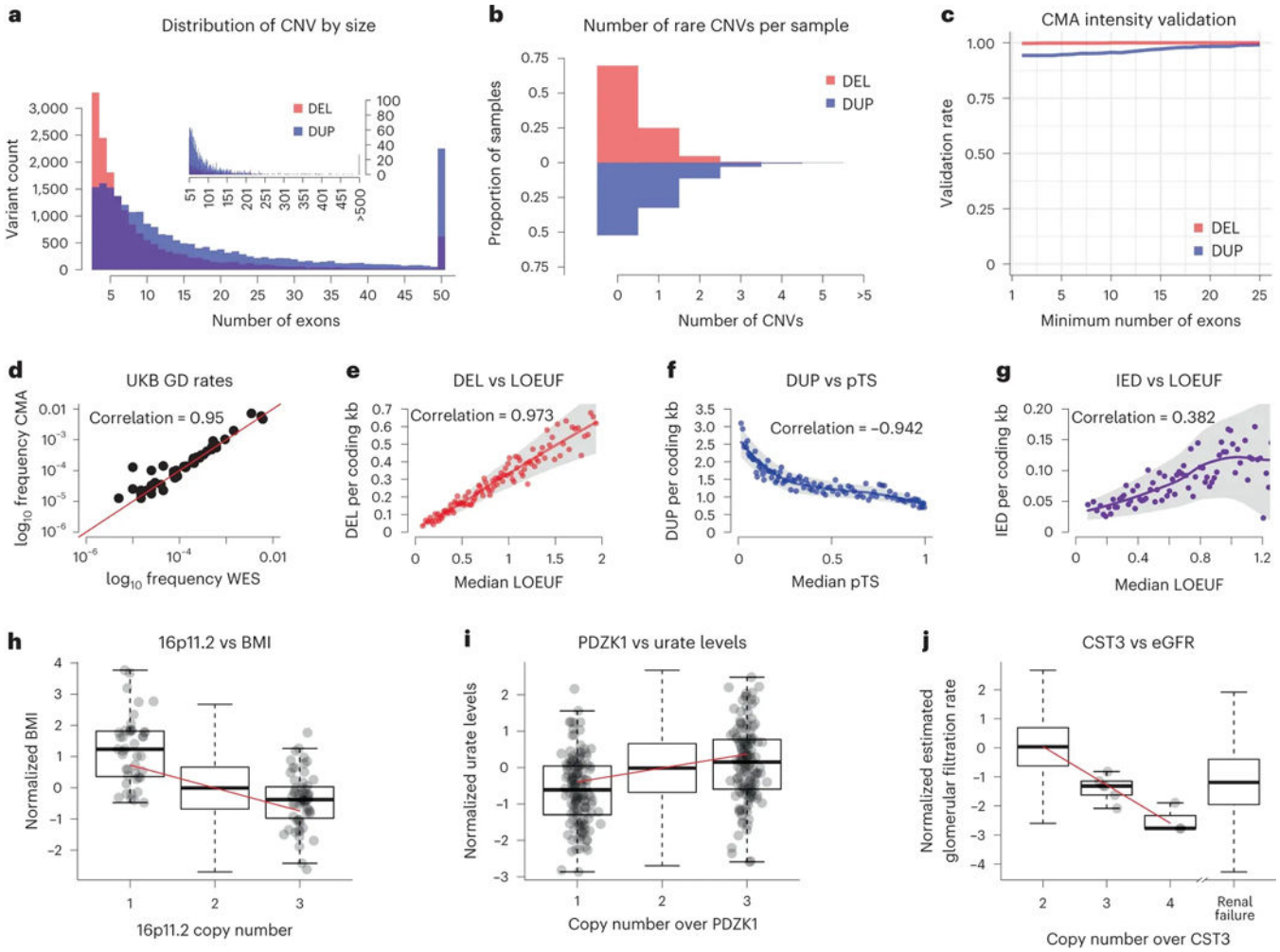

Fig. 3.

A high-quality rare CNV callset was generated on 200,624 exomes from the UK Biobank (UKBB) using GATK-gCNV **a,** The variant-size distribution of high-quality, rare CNVs in the UKBB as a function of the number of exons each variant spans. **b,** The distribution of the number of rare, high-quality CNVs per-sample in the UKBB. **c,** Using 177,158 UKBB samples with matching CMA data, we find excellent validation of high-quality GATK-gCNV WES calls using Genome STRiP Intensity Rank Sum testing. **d,** GD CNV rates in the UKBB GATK-gCNV WES callset were highly concordant with rates from previous reports based on UKBB CMA data. **e,** The number of rare deletions observed over a gene in the UKBB GATK-gCNV callset is tightly correlated with LOEUF, with grey band representing LOESS smoothing of the 95% confidence intervals on corresponding point estimates. **f,** The number of rare duplications observed over a gene in the UKBB GATK-gCNV callset is also strongly correlated with the pTriplo score measuring intolerance to duplications, with grey band representing LOESS smoothing of the 95% confidence intervals on corresponding point estimates. **g,** The number of high-confidence duplications (IED) with both breakpoints within the boundaries of a gene are also correlated with LOEUF, with grey band representing LOESS smoothing of the 95% confidence intervals on

corresponding point estimates. **h,** 16p11.2 deletions are associated with a significant increase in normalized BMI (n=41 carried a CN=1 deletion, n=61 carried a CN=3 duplication, and 169,711 individuals copy normal; boxplot corresponding to first, second, and third quartile of data, with whiskers denoting 1.5x interquartile range). **i,** PDZK1 deletions are associated with a significant decrease in normalized urate levels (n=145 carried a CN=1 deletion overlapping, n=143 carried a duplication of CN=3 overlapping, and 161,773 individuals copy norma; boxplot corresponding to first, second, and third quartile of data, with whiskers denoting 1.5x interquartile range l). **j,** CST3 duplications are significantly associated with decreased normalized eGFR values (n=6 carried a CN=3 duplication overlapping, n=3 carried a CN=3 duplication overlapping, and 162,666 individuals copy normal; boxplot corresponding to first, second, and third quartile of data, with whiskers denoting 1.5x interquartile range), on par with eGFR of individuals with renal failure (n=5,455).
Abbreviations: CNV - copy number variation; DEL - deletion; DUP - duplication; CMA - chromosomal microarray; UKBB - UK Biobank; LOEUF - loss-of-function observed over expected upper bound fraction; pTriplo - probability of triplosensitivity; IED - intragenic exonic duplication; GD - genomic disorder; WES - exome sequencing; BMI - body mass index; eGFR - estimated glomerular filtration rate.

Table 1.

Performance comparison of GATK-gCNV at different filtering thresholds, demonstrating flexibility of the method for varying performance levels.

Filtering level	CNV	Minimum number of exons	Recall (n)	Precision (n)	Mean variant count per exorne
NoQS filtering, <1% site frequency, WGS-passing samples	DEL	1	0.65 (4,246)	0.06 (47,404)	13.7
		3	0.96 (1,233)	0.21 (6,233)	1.13
		10	0.99 (338)	0.39 (908)	0.14
	DUP	1	0.73 (3,333)	0.07 (40,688)	13.2
		3	0.95 (1,895)	0.23 (8,864)	2.14
		10	0.97 (989)	0.57 (1,802)	0.46
No QS filtering, <1% site frequency, WES-passing samples, WGS-passing samples	DEL	1	0.66 (3,992)	0.09 (31,739)	13.7
		3	0.96(1,165)	0.41 (2,964)	1.13
		10	0.99 (318)	0.87 (280)	0.14
	DUP	1	0.74(3,151)	0.11 (24,545)	13.2
		3	0.96 (1,802)	0.40 (4,829)	2.14
		10	0.97 (939)	0.84 (1,147)	0.46
Recommended QS filtering: (If CN=0, QS≥min (1,000, max (400,10×n^{Int})) If CN=1, QS≥min (1,000, max(100,10×n^{Int})) If CN>2, QS≥min (400, max(50,4×n^{Int}))), <1% site frequency, WES-passing samples, WGS-passing samples	DEL	1	0.41 (3,992)	0.92 (1,986)	0.81
		3	0.88 (1,165)	0.95 (1,150)	0.47
		10	0.99 (318)	0.96 (347)	0.12
	DUP	1	0.55 (3,151)	0.84 (2,353)	1.00
		3	0.85 (1,802)	0.87 (1,963)	0.83
		10	0.95 (939)	0.92 (1,024)	0.41
QS>1,000, <1% site frequency, WES-passing samples, WGS-passing samples	DEL	1	0.07 (3,992)	0.96 (336)	0.11
		3	0.24(1,165)	0.97 (328)	0.11
		10	0.67 (318)	0.97 (231)	0.08
	DUP	1	0.08 (3,151)	0.95 (251)	0.07
		3	0.13 (1,802)	0.95 (250)	0.07
		10	0.24(939)	0.95 (250)	0.07

Abbreviations: QS - quality score output by GATK-gCNV used for call-level filtering; GS - genome sequencing; ES - exome sequencing; CN - copy number; N^{Int} - number of well-captured intervals.