HGG Advances

ARTICLE

Omics-informed CNV calls reduce false-positive rates and improve power for CNV-trait associations

Maarja Lepamets,[1,2,16,*] Chiara Auwerx,[3,4,5,6] Margit Nõukas,[1,2] Annique Claringbould,[7] Eleonora Porcu,[3,5,6] Mart Kals,[1,8] Tuuli Jürgenson,[1,9] Estonian Biobank Research Team,[1] Andrew Paul Morris,[1,10] Urmo Võsa,[1] Murielle Bochud,[6] Silvia Stringhini,[11] Cisca Wijmenga,[12] Lude Franke,[12,13] Hedi Peterson,[14] Jaak Vilo,[14] Kaido Lepik,[4,5,6,14] Reedik Mägi,[1,15] and Zoltán Kutalik[4,5,6,15,*]

Summary

Copy-number variations (CNV) are believed to play an important role in a wide range of complex traits, but discovering such associations remains challenging. While whole-genome sequencing (WGS) is the gold-standard approach for CNV detection, there are several orders of magnitude more samples with available genotyping microarray data. Such array data can be exploited for CNV detection using dedicated software (e.g., PennCNV); however, these calls suffer from elevated false-positive and -negative rates. In this study, we developed a CNV quality score that weights PennCNV calls (pCNVs) based on their likelihood of being true positive. First, we established a measure of pCNV reliability by leveraging evidence from multiple omics data (WGS, transcriptomics, and methylomics) obtained from the same samples. Next, we built a predictor of omics-confirmed pCNVs, termed omics-informed quality score (OQS), using only PennCNV software output parameters. Promisingly, OQS assigned to pCNVs detected in close family members was up to 35% higher than the OQS of pCNVs not carried by other relatives ($p < 3.0 \times 10^{-90}$), outperforming other scores. Finally, in an association study of four anthropometric traits in 89,516 Estonian Biobank samples, the use of OQS led to a relative increase in the trait variance explained by CNVs of up to 56% compared with published quality filtering methods or scores. Overall, we put forward a flexible framework to improve any CNV detection method leveraging multi-omics evidence, applied it to improve PennCNV calls, and demonstrated its utility by improving the statistical power for downstream association analyses.

Introduction

Copy-number variations (CNV) are unbalanced structural variations that alter the dosage of genomic regions via deletion and duplication events. Approximately 9.5% of the human genome is subject to CNVs,[1] which vary in length, ranging from a few dozens to several millions of base pairs (bp) in length. CNVs tend to have more severe phenotypic consequences compared with single-nucleotide variations (SNVs) as, due to their larger size, they can encompass entire coding regions.

CNVs have been associated with a number of conditions including autism,[2] schizophrenia,[3] neurodegenerative disorders,[4,5] and cancer.[6] A number of large recurrent deletions and duplications have been combined into the DECIPHER CNV syndromes list.[7] Importantly, incomplete penetrance of several syndromic CNVs has been established by studying large population biobanks, where CNV load was shown to increase the risk of obesity, physical or cognitive impairment, and congenital malformations while lowering educational attainment and socio-economic status.[8–11] In parallel, CNV genome-wide association studies (GWASs) have been conducted on numerous diagnoses[12,13] and medically relevant continuous traits,[11,14,15] including a large meta-analysis on anthropometric measurements,[16] revealing the important role of CNVs in shaping the human phenome.

Over the years, multiple CNV detection algorithms have been developed for CNV detection from SNV genotyping microarray probe intensities.[17] Currently, PennCNV[18] is the most widely used software for genotyping array-based

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia; [2]Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia; [3]Center for Integrative Genomics, University of Lausanne, Lausanne 1015, Switzerland; [4]Department of Computational Biology, University of Lausanne, Lausanne 1015, Switzerland; [5]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; [6]Center for Primary Care and Public Health (Unisanté), Department of Epidemiology and Health Systems, University of Lausanne, Lausanne 1010, Switzerland; [7]Structural and Computational Biology Unit, EMBL, Heidelberg 69117, Germany; [8]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki 00014, Finland; [9]Institute of Mathematics and Statistics, University of Tartu, Tartu 51009, Estonia; [10]Centre for Genetics and Genomics Versus Arthritis, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester M13 9PL, UK; [11]Unit of Population Epidemiology, Division of Primary Care, Geneva 1205, Switzerland; [12]University of Groningen, University Medical Center Groningen, Department of Genetics, 9713 AV Groningen, the Netherlands; [13]Oncode Institute, 3521 AL Utrecht, the Netherlands; [14]Institute of Computer Science, University of Tartu, Tartu 51009, Estonia
[15]These authors contributed equally
[16]Lead contact
*Correspondence: maarja.lepamets@ut.ee (M.L.), zoltan.kutalik@unil.ch (Z.K.)
https://doi.org/10.1016/j.xhgg.2022.100133.

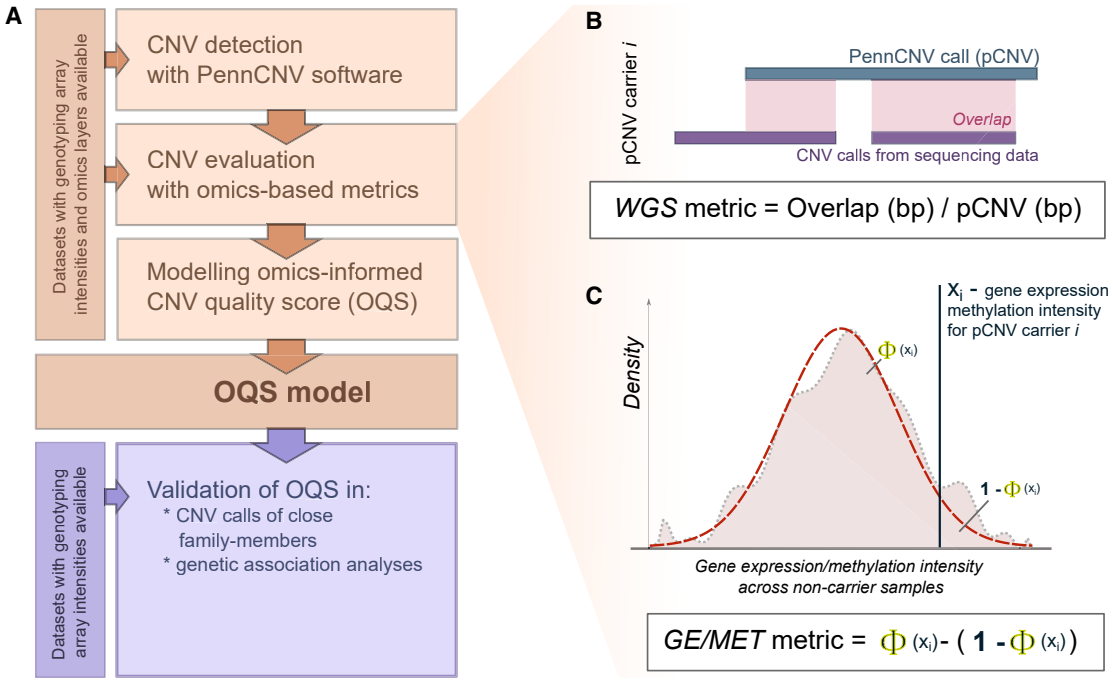

Figure 1. Workflow overview

(A) Quality estimation and modeling pipeline for PennCNV copy-number variation calls (pCNVs).

(B and C) The pCNV quality metrics are estimated based on (B) whole-genome sequencing (WGS) data and (C) gene expression (GE) and/or overall methylation (MET) intensity of genes/CpG sites overlapping the corresponding CNV calls.

(B) WGS metric is a fraction of pCNV that can be mapped to WGS CNVs of the same individual.

(C) To calculate GE/MET metrics, the reference distribution of expression/intensity based on non-carriers (pink area) is approximated to standard normal distribution (red dashed line), and the *Z* score of the expression/intensity of each pCNV carrier (x_i) is compared with it one at a time. The metric is a difference between the fraction of non-carriers with the corresponding value $\leq x_i$ and those with the corresponding value $>x_i$ and captures how extreme x_i is compared with the reference distribution of non-carriers. In case a pCNV overlaps with several genes/CpG sites, the metric values are averaged over them.

calling. For each sample, a hidden Markov model (HMM)-based algorithm uses overall signal intensity and continuous allelic intensity at polymorphic probes to estimate the probability of a hidden copy-number state at this genomic location. Unfortunately, CNV regions found by different array-based detection methods only agree in about 20% of cases,[19] indicating the high likelihood of false-positive calls. To counter this, various filtering strategies have been employed, usually by setting cut-off values to combinations of parameters including number of CNVs per sample, minimum CNV length, probe density, and PennCNV confidence score.[9,12,18,20,21] Filtering based on arbitrary thresholds is suboptimal and a continuous CNV quality score that predicts the probability that a CNV region is a consensus ($\geq$70% call overlap) between PennCNV, QuantiSNP,[22] and CNVpartition (an Illumina developed GenomeStudio software plug-in, web resources) has been proposed.[19] We refer to this as consensus-based quality score (cQS).

Still, cQS relies only on a single input dataset (i.e., microarray data). An alternative strategy to improve CNV calling can be to incorporate various types of omics datasets. Previously, software have been developed to infer CNVs from high-density DNA methylation arrays[23,24] or RNA sequencing data of highly and stably expressed genes.[25] While promising, none of these approaches were developed with the intent of performing scalable and reliable genome-wide CNV detection in large biobanks.

To fill this gap, we propose a method to improve the detection of false-positive CNV calls among PennCNV output by discriminating between high-quality (true) and low-quality (false) CNV regions based on multi-omics data (Figure 1A). Specifically, we checked if PennCNV calls (pCNVs) (1) are detectable by whole-genome sequencing (WGS), (2) alter the expression levels of overlapping genes in the expected direction (i.e., decreased by deletions, increased by duplications), and/or (3) alter the total methylation probe intensity of overlapping CpG sites in the expected direction. We built a predictor of CNV quality inferred from WGS, transcriptomics, and methylomics, solely based on PennCNV software output parameters in these samples assayed by multiple omics technologies. Predicted omics-informed quality scores (OQSs) distinguish high- from low-quality CNVs even in samples for which only SNV genotyping microarray data are available. We show that OQS reduces false discovery rate and improves CNV-trait association discovery compared with both raw pCNVs and cQSs[19] in regions with variable CNV quality.

Table 1. Overview of datasets and final sample sizes used in the analyses.

Dataset	n	Sample counts per data type			Analysis steps			
		WGS	Methyl.	RNA-seq	Omics-based metrics calculation	Model building	Model selection and validation	CNV associations
Estonian OmniExpress sample set (N = 7,750)								
EstBB-MO	1,066	983	295	382	+	+	–	–
First-degree relatives	504[a]	N/A	N/A	N/A	–	–	+	–
Lifelines deep (N = ~1,500)								
LLDeep	1,383	N/A	768	1,098	+	+	–	–
Swiss Kidney Project on Genes in Hypertension (N = 1,128)								
SkiPOGH	466	N/A	148	405	+	–	–	–
Parent-child pairs	319	N/A	N/A	N/A	–	–	+	–
Estonian Biobank GSA sample set (N = ~200,000)								
EstBB-GSA (unrelated)	89,516	N/A	N/A	N/A	–	–	–	+
MZ twins	312	N/A	N/A	N/A	–	–	+	–
First-degree relatives	79,903	N/A	N/A	N/A	–	–	+	–
UK Biobank (N = ~500,000)								
UKB (unrelated British)	331,522	N/A	N/A	N/A	–	–	–	+
MZ twins	302	N/A	N/A	N/A	–	–	+	–
First-degree relatives	42,032	N/A	N/A	N/A	–	–	+	–

N/A, not applicable.
[a]Estonian OmniExpress first-degree relatives do not overlap with EstBB-MO samples.

Material and methods

Cohorts

Estonian Biobank (EstBB; data freeze January 8, 2021; Note S1; Table 1) is an Estonian population-based cohort that consists of ~200,000 adults (≥18 years of age at recruitment).[26] About 7,750 individuals are genotyped on Illumina Infinium OmniExpress-24 genotyping array (~730,000 markers). A subset of these samples (referred to as EstBB-MO) has one or more of the following omics datasets available: 30× coverage WGS, RNA sequencing,[27] and/or methylation data (Illumina Infinium Human Methylation 450 k Beadchip). Additionally, the full EstBB cohort is genotyped on Illumina Global Screening Array (GSA; ~760,000 markers). All participants signed a broad informed consent, and analyses were carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research and data release N05 from the EstBB.

LifeLines Deep (LLDeep; Note S2; Table 1) is a deeply phenotyped ~1,500 individual subset of the Dutch population cohort LifeLines.[28] LLDeep samples are genotyped on HumanCytoSNP-12 array (~300,000 markers), and the majority of them have either RNA sequencing[29] or methylation data (Illumina Infinium Human Methylation 450 k Beadchip)[30] available. The LLDeep study was approved by the ethics committee of the University Medical Centre Groningen. All participants provided a written informed consent.

Swiss Kidney Project on Genes in Hypertension (SkiPOGH; Note S3; Table 1) is a Swiss family- and population-based cohort of 1,128 individuals from 273 families recruited to study the genetic determinants of blood pressure.[31] The samples were genotyped on Illumina 2.5 array (~2,500,000 markers). RNA sequencing and methylation array (Illumina Infinium Human Methylation 450 k) data were available for a subset of participants.[32] The study was approved by the competent institutional ethics committees in Bern, Geneva, and Lausanne. All participants signed a written informed consent.

UK Biobank (UKB; phenotype data freeze March 22, 2018; Note S4; Table 1) is a cohort of ~500,000 individuals from the UK.[33] The majority of samples (~450,000) are genotyped on Affymetrix UKB Axiom array, while the rest (~50,000) are genotyped on Affymetrix UK BiLEVE Axiom array (both arrays have ~820,000 markers). Participants signed a broad informed consent, and the data are accessed through application numbers 17085 and 16389.

Data preparation

Sample sets

We included three independent datasets—LLDeep, SkiPOGH, and a subset of Estonian samples (EstBB-MO)—in CNV quality calculations and modeling. Each of these

datasets had additional omics data (WGS, methylation arrays, and/or RNA sequencing) available. The summary of PennCNV output parameters for each quality-controlled cohort are shown in Table S1. For model selection and validation steps, we extracted monozygotic (MZ) twins and first-degree relatives from the EstBB and the UKB and parent-child pairs from the SkiPOGH. Finally, we extracted unrelated quality-controlled EstBB-GSA and UKB samples for CNV association analyses. Datasets and their usage for various analyses are summarized in Table 1. Sample quality control steps are summarized in Notes S1–S4.

CNV detection

We used PennCNV[18] as our main CNV detection algorithm due to its popularity (PubMed citations: PennCNV: 885; QuantiSNP:[22] 279; Birdsuite:[34] 466; September 8, 2021). We detected putative autosomal CNV regions (pCNV) for EstBB, LLDeep, SkiPOGH, and UKB datasets as previously described[19] (Notes S1–S4). For each sample, we obtained the pCNV together with the values of four CNV-specific and nine sample-specific parameters described in Table S2. In all datasets, we filtered out samples with more than 200 pCNVs and with pCNVs larger than 10 Mbp, as these are likely to be either samples with poor genotyping quality or extreme cases that might distort the analysis. Additionally, we detected CNVs from EstBB WGS reads (WGS-CNVs) using the Genome STRiP[35] discovery pipeline (v.2.00.1611; Note S5). All genomic coordinates are in GRCh37 build version.

Methylation and RNA sequencing data preprocessing

We obtained methylation intensities (Infinium Human Methylation 450 k Beadchip) and RNA sequencing data for EstBB-MO, LLDeep, and SkiPOGH datasets. The data preparation is described in detail in Notes S6 and S7. Briefly, where applicable, after the quality control step, we corrected for age, sex, batch, blood cell counts, and population stratification based on four principal components (PCs) calculated from pruned SNP genotypes (minor allele frequency >1%). Additionally, we corrected for PCs calculated based on methylation/gene-expression data (Figures S1 and S2). Gene-expression residuals were further corrected for independent expression quantitative trait loci (eQTL) within 500 kbp of the gene.

CNV quality metrics based on multi-omics data

WGS quality metric

WGS data were available for a subset of EstBB-MO samples (n = 979). For each pCNV in these individuals, we defined a WGS metric as the fraction of the pCNV (in bps) overlapping with WGS-CNVs in the same sample (Figure 1B). Metric calculation was restricted to pCNV deletions and duplications longer than 1 and 2 kb, respectively, as we did not detect shorter WGS-CNVs (Note S5). Additionally, these samples were used to estimate the fraction and distribution of false negative (FN) and false positive (FP) pCNVs. High-confidence FN calls were defined as WGS-CNVs (copy numbers between 0 and 4) with less than 10% bp overlap with pCNVs in the same sample. pCNVs with WGS metric <0.1 were defined as high-confidence FPs.

MET quality metric

Analogously to SNP arrays, higher or lower total signal intensity captured by methylation array at a CpG site indicates excess or lack of DNA material, respectively, regardless of the methylation status of the region. Exploiting this phenomenon, CpG site intensity data can be used to validate duplications (i.e., excess genetic material leading to increased total intensity) and deletions (i.e., reduced genetic material leading to decreased total intensity). This approach was used to assess CNV quality in EstBB-MO, LLDeep, and SkiPOGH datasets. For each methylation probe passing the preprocessing steps (Note S6), we used the samples with no pCNVs overlapping the corresponding CpG site (i.e., non-carriers) to construct the approximately Gaussian reference distribution of site overall intensity (sum of the methylated and un-methylated intensities). For each carrier, we then transformed its CpG site overall intensity into a Z score by using the mean and standard deviation of the constructed reference distribution. We denoted the quality metric based on the methylation data for the i-th pCNV ($pCNV_i$) across all its overlapping CpG sites as $MET_i \in [-1;\ 1]$ and calculated its value as

$$MET_i = \frac{\sum_{j=1}^{n_i}\left(\Phi(m_{ij}) - \left(1 - \Phi(m_{ij})\right)\right)}{n_i},$$

where n_i is the total number of CpG sites overlapping $pCNV_i$, Φ is the cumulative distribution function of the standard normal distribution, and m_{ij} is the Z score calculated for the j-th CpG site overlapping $pCNV_i$ (Figure 1C). The proposed measure captures how extreme an observed methylation intensity is compared with that of the bulk of the samples (assumed to be copy neutral), equivalent to a signed two-sided tail probability. We expected $MET_i < 0$ for deletions and $MET_i > 0$ for duplications. If this was not the case, MET_i was set to zero. Finally, MET_i was converted to its absolute value such that $MET_i \in [0; 1]$.

Gene expression (GE) quality metric

GE levels from RNA sequencing data were used to assess CNV quality in the EstBB-MO, LLDeep, and SkiPOGH datasets. We extracted all the genic regions from Ensemble database (GRCh37) using biomaRt.[36] To avoid penalizing genes whose transcript levels are not affected by CNVs, we only retain genes for which expression is positively correlated (Pearson R > 0.1) to the copy number of the gene in an independent dataset.[25] After preprocessing steps (Note S7), we retained 10,786 genes. Additionally, over 80% of the genic region was required to overlap the pCNV for the gene to be included in the quality calculations of that pCNV. Requiring higher gene overlap did not show considerable improvement (Figure S3). Expression values of genes with $pCNV_i$ overlap below 80% were marked as missing. Analogously to MET_i, we constructed the expression reference distribution based on non-carriers and used its mean and standard deviation to calculate

an expression Z score for each carrier. We calculated $GE_i \in [-1;1]$, the $pCNV_i$ quality metric based on GE across all its overlapping genes (analogously to the metric for methylations), as

$$GE_i = \frac{\sum_{j=1}^{n_i}(\Phi(e_{ij}) - (1 - \Phi(e_{ij})))}{n_i},$$

where n_i is the number of genes overlapping at least 80% of the $pCNV_i$ and e_{ij} is the Z score calculated for the j-th gene overlapping $pCNV_i$ (Figure 1C). We set zero GE_i values for deletions with $GE_i > 0$ and duplications with $GE_i < 0$ and converted all GE_i scores to their absolute values such that $GE_i \in [0;1]$. Although we hypothesized that duplications could alter GE in either direction through either triplication or disruption of gene sequence,[37] this was not observed in our results (Figure S3).

Combined metric

Let us define the collection of quality metrics $\boldsymbol{Q_i} = \{MET_i, GE_i, WGS_i\}$ Further metrics can be defined by their mean, maximum, and the measure furthest away from 0.5 (i.e., most extreme; denoted as $EXTR(Q_i)$). We chose $EXTR(Q_i)$ as our final combined metric, the motivation being that if one metric clearly indicated the truth status of a pCNV, then we would use that metric (even if other metrics are unsure) (Figure S4). Note that EXTR values are only calculated for pCNVs that have at least two out of three Q_i metrics available.

CNV quality prediction models

In order to assess CNV quality in samples with no complementary omics data, we fitted prediction models for the values of the previously defined four omics metrics, summarized by $\boldsymbol{Y_i} = \{MET_i, GE_i, WGS_i, EXTR(Q_i)\}$.

The set of possible explanatory variables $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ included CNV- and sample-specific parameters from PennCNV output (CNV length, number of overlapping probes, PennCNV-specific CNV confidence score, number of pCNVs per sample [and its derivations, see Figure S5], mean and standard deviation of the allelic intensity ratios and B allele frequencies of a sample, signal waviness factor; Table S2) and their interaction terms. CNV coordinates were not included as explanatory variables. We fitted a generalized linear regression model separately on each column of $\boldsymbol{Y}$:

$$P(Y^j|\boldsymbol{X}) = f\left(\beta_0 + \sum_{k=1}^{n} X_k\beta_k\right),$$

where Y^j is the j-th column of $\boldsymbol{Y}$ (representing one of the four quality metrics), β_0 is the intercept term, and β_k is the effect estimate of explanatory variable X_k. We used a quasi-binomial link function f since instead of being strictly binary, our response was a bimodal continuous variable ranging from zero to one.

In order to choose the best subset of $\boldsymbol{X}$, we implemented a forward stepwise model selection (starting with an empty parameter set) using custom R scripts. Briefly, in each round, a parameter was added if the resulting model minimized the 10-fold cross-validation mean square error (MSE). If adding any of the remaining parameters did not improve the average MSE, the algorithm stopped and returned the existing model. We tested model building with eight different sets of conditions/parameters $\{X^s : X^s \subset \boldsymbol{X}\}$ to choose from and repeated the procedure separately for deletions and duplications. The modeling process and the eight parameter sets are characterized in detail in Note S8 and Table S3.

The model coefficients β_k can be used to predict omics-informed CNV quality scores (OQSs) as

$$OQS = \frac{1}{1 + \exp(-(\beta_0 + \sum_{k=1}^{n} X_k\beta_k))}.$$

For X_k not included in the final model, β_k is set to be equal to zero.

Selection of the best OQS and comparison with other CNV calls

CNV quality models were fitted as described above for deletions and duplication from each multi-omics dataset separately (Table 1). To determine the best models, we incorporated family information. We reasoned that the set of familial pCNVs present in at least two close family members (Jaccard index based on overlapping bp count of at least 0.9) contains a higher fraction of true-positive CNV calls than the non-familial set (no overlap in a relative). Partially overlapping pCNVs with a Jaccard index lower than 0.9 were discarded from the calculations. We predicted and compared the OQS values of all familial and non-familial pCNVs from MZ twins from the UKB and EstBB-GSA and parent-child pairs from the SkiPOGH. To select the best models, we maximized the difference of mean OQS values between the two groups, averaged over the three datasets. We validated our best models on first-degree relative pCNVs from Estonian OmniExpress samples, EstBB-GSA and UKB. To further reduce the likelihood of two relatives carrying overlapping FP calls by chance in our validation sets, we restricted the analysis to rare (frequency <0.1%) familial pCNVs, with frequency being calculated as the fraction of samples in the full cohort with a pCNV overlapping the region in question. As a comparison, we estimated the quality of non-familial and familial pCNVs using the previously published cQSs.[19]

CNV association analyses

We compared OQS with raw pCNV, four previously published PennCNV output filtering approaches[18,20,38,39] (Table S4), and cQS in an association analysis setting by incorporating them into the association models analogously to SNV dosages.[16] We used 89,516 and 331,522 quality-controlled unrelated European individuals from the EstBB-GSA and UKB data, respectively (Notes S1 and S4). We considered 21 CNV-trait pairs (Table S5) involving four continuous anthropometric traits—body mass index

(BMI), height, weight, and waist-to-hip ratio (WHR)—and 13 CNV regions that had previously[16] shown association p $<1 \times 10^{-4}$. Importantly, these associations were obtained using cQSs in the first wave of UKB genotype data samples (n = 119,873). All phenotypes were inverse normal transformed and corrected for batch, sex, age, age,[2] and PCs prior further analysis (Notes S1 and S4).

We calculated association Z statistics (estimated effect size over its standard error) and p values in a probe-by-probe manner across all probes that overlapped with >5 pCNVs inside the 13 regions of interest. The analysis was conducted separately for deletions, duplications, and mirroring effects. To model the mirroring effect (i.e., both deletions and duplications have similar effects but in opposite directions), the OQS values for deletions were negated. Associations were run using linear regression (*lm* function) in custom R scripts. We used a Bonferroni-corrected p value threshold of $0.05/21 = 2.38 \times 10^{-3}$ to determine significance. All regions containing significantly associated probes—except for the 18q21.32 region, which in the EstBB-GSA did not contain the previously reported CNV (Figure S6)—were included in the final association comparison step.

Finally, for both datasets (i.e., UKB and EstBB-GSA) and all four phenotypes, we estimated the change in explained variance when applying the OQS model, as compared with raw PennCNV values, four filtering approaches, or cQS model. We started by clumping probes originating from significant CNV regions with $R^2 > 0.5$ using *snp_clumping* from the *bigsnpr* R package.[40] Clumping usually prioritizes probes based on association summary statistics or allele frequencies, which in our case are heavily dependent on the applied CNV quality measure. To avoid any bias, we generated a random probe priority order instead. If after clumping we retained m probes, we calculated

$$F = \frac{Z_2' S^{-1} Z_2}{Z_1' S^{-1} Z_1},$$

where Z_2 is an array of Z statistics (of clumped probes) from association analysis using OQS, Z_1 is the corresponding array from the comparison analysis (either raw PennCNV, filtering approaches, or cQS), and S is a probe correlation matrix calculated based on raw pCNV. Under the null hypothesis, F follows an F distribution with both degrees of freedom equal to m. Since F depends on the probes retained after the randomized clumping process, we repeated the random clumping 20 times and used the average F value.

Results

Omics-based metrics for CNV quality

We estimated pCNV quality based on methylation (MET), GE, and WGS (only available in EstBB-MO dataset) data, which resulted in up to three independent omics-based CNV quality metrics in three independent datasets (EstBB-MO, LLDeep, and SkiPOGH; Tables 1 and S6). Within all datasets, the metrics were positively correlated with each other (Figures 2A, 2B, S7, and S8). Both MET and GE metrics had high correlations (Pearson R ≥ 0.7) with the WGS metric. The correlations between MET and GE metrics ranged between 0.59 and 0.80 for deletions and 0.33 and 0.57 for duplications. The correlations between all three metrics and previously published cQSs[19] ranged between 0.17 and 0.55 for deletions and 0.21 and 0.63 for duplications, depending on the dataset.

All three metrics had bimodal distributions with modes near 0 and 1, which indicates clear differentiation between true and false calls for the majority of pCNVs (Figures 2C, S7, and S8). To retain just one quality metric per pCNV (i.e., combined metric; Figure 2D), we retained the metric that was furthest from 0.5 (denoted as EXTR in material and methods). Detailed composition of this metric is characterized in Figure S10.

We estimated the precision of pCNV based on the WGS and the combined metric. In the EstBB-MO dataset, out of 3,496 pCNVs evaluated with the combined metric (1,750 deletions, 1,746 duplications), 47.3% of deletions and 47.5% of duplications had values inferior to 0.1, most likely reflecting FP calls. In LLDeep and SkiPOGH, the percentages corresponding to high-confidence FP calls were 50.5%/28.3% and 70.9%/59.4% for deletions/duplications, respectively (Table S7). When considering a larger EstBB-MO set of 15,063 deletions and 8,914 duplications with the WGS metrics available, 31.6% (n = 4,762) of deletions and 53.2% (n = 4,745) of duplications could be labeled as FPs. These results illustrate the need for CNV quality filtering prior further analyses.

Using this EstBB-MO set, we studied the distribution of FP calls across the genome (Figure S11). Overall, 120,395 probes (17% of all Illumina OmniExpress autosomal probes, 77.5% of probes overlapping pCNV) had an FP rate (FPR) >0. There was a modest negative correlation between FPR and pCNV frequency (Pearson R = −0.12) indicating that if a pCNV is detected in multiple samples, it is more likely to be true. Still, for 939 probes, FP pCNVs were discovered in ≥10 samples (FP frequency >1% in 979 samples) with 816 (86.9%) of them having a FPR >0.9. These FP "hot spots" contributed to 5.1% (488/9,507) of all FP calls.

Additionally, we estimated the fraction of FN pCNV calls based on the overlap with WGS-CNVs to be 97.3%. When only considering CNVs that overlapped ≥3 genotyping array probes (n = 69,889), thus meeting the minimum requirement of the PennCNV discovery algorithm, this number dropped to 75.7% (66.6% for deletions only, and 84.1% for duplications only). These percentages further decreased with increasing number of overlapping probes, remaining considerably higher for duplications, compared with deletions (Figure S12). We further studied the genome-wide distribution of FNs on a restricted set on WGS-CNVs with a ≥3 probe overlap (Figure S13). We observed that despite a large fraction of FNs, only 12,065

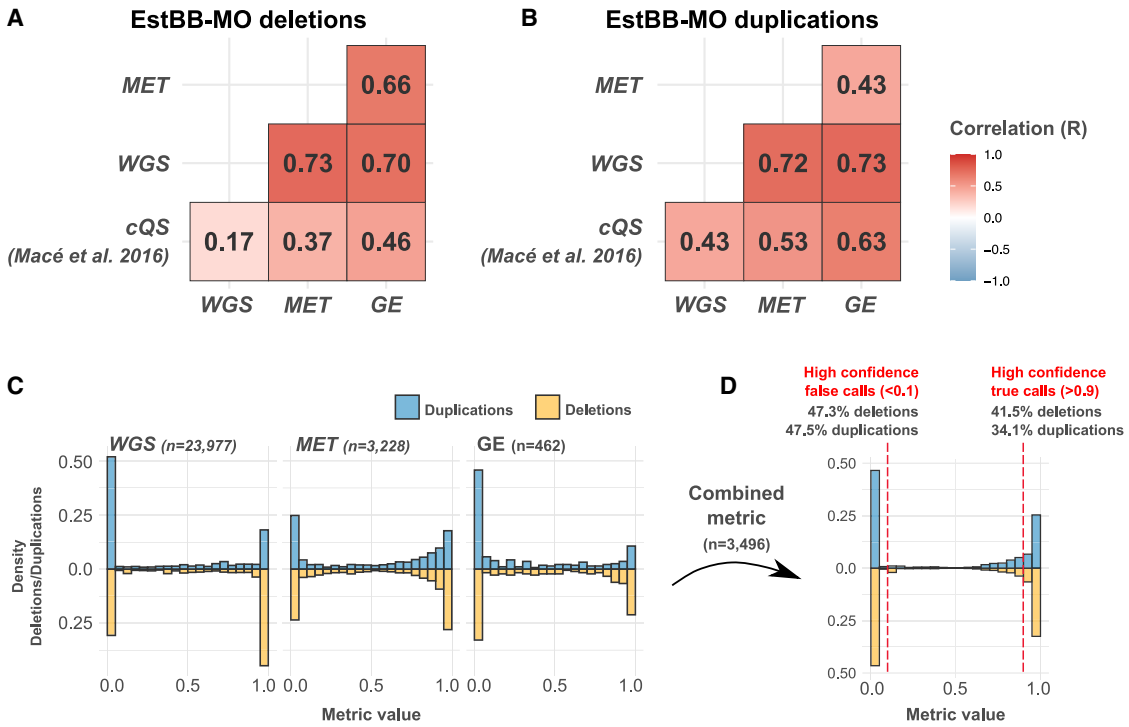

Figure 2. Overview of CNV quality metrics in EstBB-MO

(A and B) Omics-based metrics— WGS, MET, and GE—and cQS[19] Pearson correlations for EstBB-MO deletions (A) and duplication (B). Note that the number of pCNVs used in correlation calculations is not identical in each group of metric pairs (Figure S9).

(C and D) Bimodal distribution of WGS, MET, and GE metrics (C), as well as their combined metric (see material and methods) (D) for duplications (blue) and deletions (yellow). The combined metric is calculated for pCNVs that have at least two omics-based metrics available (n = 3,496) and the fractions of high-confidence false (combined metric <0.1) and true (combined metric >0.9) calls are reported.

probes (1.7% of all probes, 40.0% of probes overlapping WGS-CNVs) had an FN rate (FNR) >0. Unlike the FPR, the FNR was positively correlated to WGS-CNV frequency (Pearson R = 0.19), and CNVs with >50% frequency contributed to 31.8% (22,196/69,889) of all FN CNVs.

Prediction models for omics-informed CNV quality scores (OQSs)

We built logistic regression models to predict the previously calculated CNV quality metrics based solely on PennCNV output parameters to enable pCNV evaluation in samples lacking multi-omics measurements. Due to a smaller fraction of true-positive calls when compared with other datasets, we omitted SkiPOGH from the model-building step but retained it for model validations. Models were evaluated based on their ability to discriminate between pCNVs that were shared (familial) or not (non-familial) between MZ twins in the UKB and EstBB-GSA and parent-child pairs in the SkiPOGH (Tables S8 and S9), as pCNVs detected across multiple family members are less likely to be FPs and, thus, should act as a set of likely true-positive calls suitable for model selection and validation. For both deletions and duplications, the best model was built based on the LLDeep dataset using the combined metric. Models are characterized in Tables S10 and S11. We refer to the CNV quality measure (ranging from 0 to 1) predicted by the best models as the omics-informed CNV quality score (OQS).

To validate the OQS, we performed a familial versus non-familial pCNV comparison on first-degree relatives from the Estonian OmniExpress, EstBB-GSA, and UKB that did not overlap with individuals used for the CNV quality estimation and model-building steps (i.e., samples with other omics data; Figures 3 and S14). The average OQS for familial calls ranged between 0.67 and 0.82 for deletions and between 0.48 and 0.70 for duplications, which was significantly higher (paired Wilcoxon test $p < 1.4 \times 10^{-21}$) than for cQS (0.27–0.32 in deletions and 0.42–0.53 in duplications). As some genomic regions are more prone to FP pCNVs, resulting in shared false calls between close relatives by chance, we executed a similar analysis using only rare (frequency <0.1%) familial pCNVs. This further increased the average OQS values, which ranged between 0.76 and 0.83 for deletions and between 0.56 and 0.76 for duplications. In all cases except EstBB-GSA duplications, we observed significantly higher score values for rare pCNVs with OQS compared with cQS (0.33–0.40 for deletions, 0.56–0.66 for duplications; Wilcoxon $p < 0.046$). Furthermore, OQS distinguished well between familial and non-familial pCNVs. The difference in OQSs between two groups were between 0.22 and 0.35 depending on the dataset (0.16–0.25 for deletions and 0.12–0.48 for duplications; Wilcoxon $p < 3.0 \times 10^{-90}$). Only in the

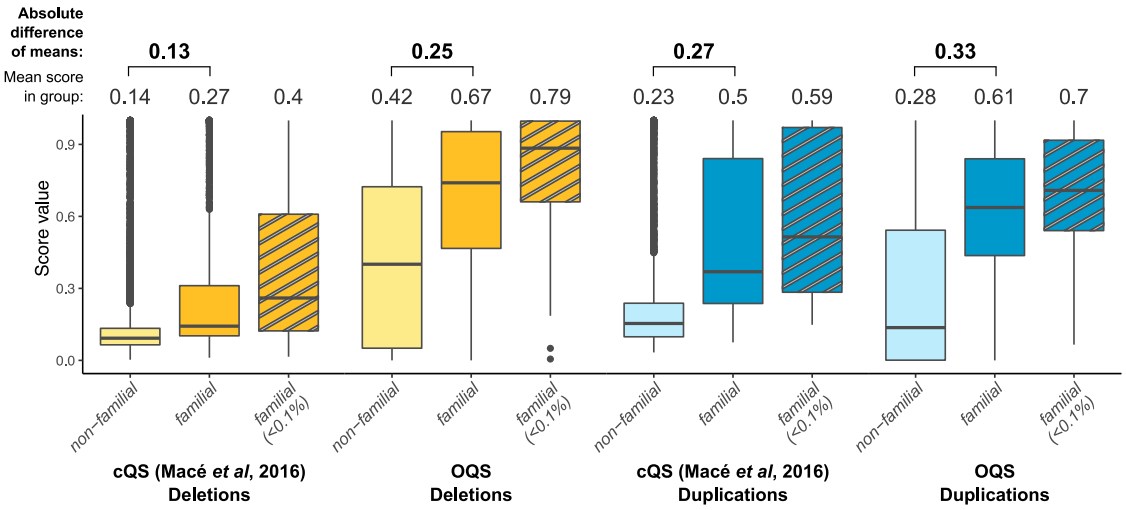

Figure 3. Comparison of quality scores on pCNVs of closely related Estonian samples
Consensus-based (cQS)[19] and omics-informed (OQS) CNV quality scores of non-familial and familial (found in two or more family members) deletions (yellow) and duplications (blue) calculated on a subset of Estonian OmniExpress samples (n = 504; do not overlap with EstBB-MO). Familial pCNVs are likely true positives, while non-familial group contains both true and false positives. We included rare (frequency <0.1%, striped background) familiar pCNVs as a subset of CNVs less likely to validate in a relative by chance. The mean score of each pCNV group and their pairwise difference are shown on top of the figure. Compared with cQS, the OQS shows higher values for familial pCNVs and larger differences between non-familial and familial pCNV quality. All differences for both scores are significant with $p < 1 \times 10^{-16}$ (Wilcoxon test).

case of EstBB-GSA duplications was the average difference larger with cQS.

As the best models were built on the LLDeep dataset, we could use EstBB-MO for out-of-sample validations (Figure S15). We found that Pearson correlation coefficients between the combined metric and predicted OQSs were 0.70 and 0.57 for deletions and duplication, respectively. The area under the receiver operating characteristic curve (AUC) values were 0.91 for deletions and 0.87 for duplications (Figure S16).

Associations between CNV and anthropometric traits

We compared the association results obtained using raw pCNV, four quality filtering parameter sets,[18,20,38,39] cQS,[19] and OQS. Of 21 previously established associations between CNVs and four anthropometric traits (BMI, height, weight, and WHR),[16] we replicated ($p < 2.38 \times 10^{-3}$) 10 in the EstBB-GSA and 18 in the UKB cohort (Tables S5, S12, and S13). For both datasets, we calculated the change in variance explained per phenotype when using OQS compared with the other six approaches.

First, we tested mirror-type associations where deletions and duplications have similar effects but opposite effect directions. We found that in the EstBB-GSA, OQS led to a relative increase of 2%–34% and 23%–55% in the explained variance compared with raw PennCNV and cQS, respectively, depending on the phenotype (Figure 4A; Table S14). A good example is an association between the 16p11.2 BP4-BP5 CNV status and BMI (Figure 4B), for which alone the relative variance explained increased by 26% and 40% compared with raw PennCNV and cQS, respectively. Compared with published quality filtering approaches, the relative increase of variance explained was between 3% and 56%, depending on the approach and the phenotype (except for Wang et al.[18] and weight, for which the explained variance decreased). For deletion-only associations, the relative increase was equally good, up to 33%, 42%, and 46% compared with raw PennCNV, filtering approaches, and cQS, respectively (Figure S17). For duplication-only analysis, only one BMI-associated region was included, and it showed an up to 71% relative increase in explained variance compared with the other approaches. In the UKB, OQS showed improvement compared with raw PennCNV in three out of four phenotypes and compared with conventional filtering approaches in all four phenotypes. The greatest improvements were over Palta et al.[38] with >100% gain in explained variance and Chettier et al. [20] with, in some cases, even >400% gain in explained variance. However, compared with the cQS, the explained variance was decreased in most cases. This was to be expected, as the associations incorporated in this study were originally detected using the cQS in a dataset where over 60% of samples were from UKB.[16] None of the changes were statistically significant, as the number of independent CNV regions per phenotype was very low, ranging from one to seven.

Discussion

Genotyping microarray data are frequently used for CNV calling and analyses, but up to 48% of the calls from

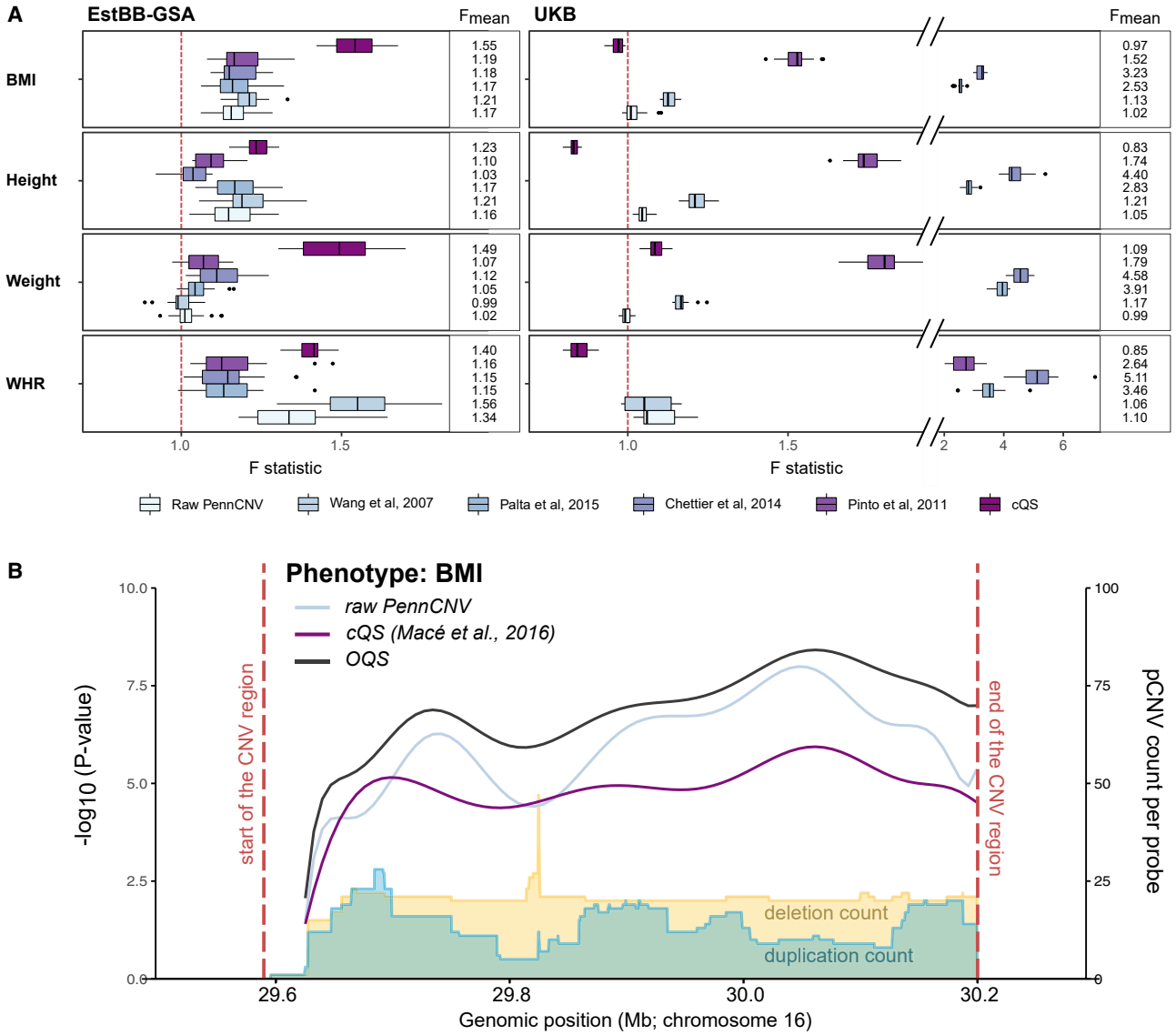

Figure 4. Impact of OQS on CNV-trait associations

(A) Change of variance explained in mirror-type model when using OQS over raw PennCNV, four published quality filtering approaches,[18,20,38,39] or cQS[19] in the EstBB-GSA and UKB, depicted as distribution of F statistics calculated by randomizing the probe pruning priority order 20 times (see material and methods). Explained variance is increased when F >1 and decreased when F <1. Larger F values indicate greater improvement in statistical power when using OQS over the given reference approach.

(B) Locus plot of a CNV region in 16p11.2 BP4-BP5 (red dashed lines: chr16:29,590,000–30,200,000 in GRCh37) associated with BMI in EstBB-GSA dataset. The lines indicate the –log10 association p values using mirror model with raw PennCNV calls (light blue), cQS (purple), and OQS (black). The yellow and blue areas illustrate the frequency of PennCNV deletion and duplication counts, respectively, across the region.

commonly used software, such as PennCNV, are not supported by other omics measures and are, therefore, likely FPs. To counter this, a quality score based on results overlap between three detection software tools has been developed (cQS).[19] We aimed at improving the discriminatory capacity of this score by devising an omics-informed CNV quality score—OQS—that incorporates independent omics-based sources of evidence to identify high-quality PennCNV CNV calls (pCNVs).

Datasets included in the development of our OQS include GE levels from RNA sequencing (GE metric) and summed methylated and unmethylated intensities at CpG sites (MET metric), as well as CNVs detected from WGS reads (WGS metric). Each of these three approaches yielded a quality metric between 0 and 1 for every pCNV, all of which showed high concordance. We found that the correlation between WGS and the other two metrics was $\geq$0.7, suggesting that the use of MET and GE data for CNV quality assessment is a suitable alternative if WGS data are not available.

Still, out of three omics layers used in our study, GE is the noisiest, as the expression changes can have various

biological and technical causes. We assumed deletions to always decrease and duplication to increase the expression levels. Duplications, however, can also disrupt the gene sequence, resulting in decreased expression levels instead.[37] To minimize the contribution of this scenario to our analysis set, we required the gene (1) to be almost fully (>80%) covered by a CNV region and (2) to show positive correlation (Pearson $R > 0.1$) to its copy number in an independent study.[25] Note that no significant improvement in results was observed when requesting higher gene-duplication overlap or allowing duplications to alter expression in both directions.

Interestingly, the correlation between WGS and previously published cQSs is quite low for comparison (0.17 for deletions, 0.43 for duplications), illustrating the potential benefit of incorporating omics data in CNV quality assessment compared with simple overlap between several detection software (which are all prone to the same weaknesses, such as prioritization of longer and discarding shorter CNV regions).

Using GE, MET, and WGS metrics, we built predictive models relying only on output parameters of PennCNV that allow estimating CNV quality in datasets where no additional omics data are available. Although larger omics data sizes can lead to better CNV quality models, we believe that even modest sample sizes can be used in case the assessed set of CNVs are a good representative of the final CNV set in the analysis. In our study, the best-quality models were built only on 441 pCNVs from an LLDeep dataset having both GE and MET metrics calculated.

In validation sets of close relatives, OQS clearly discriminated between familial(true positives) and non-familial pCNVs, the former being attributed to a higher OQS compared with cQS. This effect was consistent over all independent tested datasets. Based on out-of-sample AUC and correlations, predicting quality of deletions was easier than predicting that of duplications. Possible explanations include better detection of deletions by PennCNV due to larger relative difference in allelic intensity ratio between one and two copies compared with two and three copies or stronger effect of deletions on GE and MET. In a second step, we compared OQS with raw pCNV, four previously published sets of CNV quality filtering thresholds, and cQS through an association analysis exercise aiming at replicating previously established CNV-trait associations. We found that OQS systematically increases (up to 34% in the EstBB-GSA and 10% in the UKB) the amount of explained variance when compared with raw PennCNV. The increase was even higher (up to 56% in the EstBB-GSA and >400% in the UKB) when compared with sets of filtering thresholds. This indicates that, especially in the UKB, conventional filtering methods are too strict and result in a major loss of power. Compared with cQS, we observed a strong improvement in explained variance in the EstBB-GSA (up to 55%) but not in the UKB (down by 18%). As the associations we aimed at replicating were originally detected in the UKB using cQS approach, cQS-based associations suffer from winner's curse, which distorts the effect magnitudes in favor of the cQS. Alternatively, different quality scores might perform better in different datasets, and combining the two might be a good option (e.g., by incorporating the maximum of the two scores in the analyses).

It is to be expected that the improvements offered by the OQS is small when studying strong associations in a well-known and -detectable genomic region, as we have done. We expect to see greater improvement in intermediate-quality CNV regions for which previous studies have lacked statistical power for CNV-trait association detection. As observed, when excluding a small number of FP hotspot regions, false pCNVs are distributed randomly and uniformly across most of the genome and, thus, only introduce a modest amount of noise per probe/region. However, given the difficulty to detect CNVs, which themselves tend to be rare, even slight improvement in statistical power to detect CNV associations can be beneficial.

Although we strived to optimize our models for different genotyping array types and densities, our results may still be specific to the arrays we explored. Furthermore, while the OQS helps to reduce FP calls, it does not improve the FNR, which remains high, especially for shorter CNVs and CNVs in regions with low array probe density. Still, unlike FP, FN load mainly originates from a few specific high-frequency CNV regions.

Using only PennCNV output parameters as predictors is a limiting factor in itself, as their prediction ability can vary from dataset to dataset. Furthermore, the PennCNV detection algorithm considers each sample separately and does not exploit between-sample similarities, which was shown to improve the detection of short (and frequent) CNVs considerably.[15] Still, our omics-informed CNV quality assessment approach is not limited to PennCNV but can be used with any CNV detection method that produces multiple output parameters.

In conclusion, we developed a modular and customizable omics-based quality score framework that can be used for both genome-wide and smaller-scale CNV analyses. The OQS developed in the current study is independent of CNV coordinates or genome build and can be applied directly to filter out high-confidence FP pCNVs using a hard cutoff (i.e., OQS <0.5) or plugged into dosage-based association models, eliminating the need for an arbitrary CNV quality threshold. In turn, lower FNRs increase statistical power to detect associations between CNVs and complex traits. Alternatively, with at least one suitable multi-omics measurement available for a subset of the samples, researchers can use our framework to build their own custom models, which could be applied to any CNV detection software, leading to further improved results for follow-up analyses.

Data and code availability

Access to the UKB Resource is available by application (http://www.ukbiobank.ac.uk/). LLDeep RNA sequencing

and MET data can be accessed via European Genome-Phenome Archive (accession code EGAS00001001077). EstBB, SkiPOGH, and LLDeep individual-level data are available upon request. Custom R code and Genome STRiP pipeline commands are available on GitHub: https://github.com/maarjl/CNV_OQS.

Consortia

The members of Estonian Biobank Research Team are Andres Metspalu, Tõnu Esko, Mari Nelis, and Lili Milani.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100133.

Acknowledgments

We thank participants of EstBB, UKB, SkiPOGH, and LLDeep for their data provision. The scholarships for PhD student mobility of M.L. and K.L. to work with SkiPOGH (University of Lausanne, Switzerland) and LLDeep (University of Groningen, the Netherlands) data and funding for EstBB data acquisition were provided by the European Regional Development Fund (project nos. 2014-2020.4.01.16-0125 and 2014-2020.4.01.15-0012 and EXCITE). This project has benefited from the funding of European Union's Horizon 2020 Research and Innovation Programme under grant agreement nos. 101016775 and 633666 and of the Estonian Research Council grant PUT (PRG687, PRG555, PRG1095, and PUT JD817). Z.K. was supported by the Department of Computational Biology at the University of Lausanne. M.N. was supported by Jacobs Foundation Research Fellowship (grant 2016 1217 09, to Dr. Katrin Männik, Health 2030 Genome Center, Geneva, Switzerland). The SKIPOGH study was supported by grants from the Swiss National Science Foundation (33CM30-124087 and 33CM30-140331). EstBB and part of the UKB computations were carried out on the High-Performance Computing (HPC) Center, University of Tartu. UKB association study was carried out on the JURA server, University of Lausanne. SkiPOGH computations were carried out on the HPC1 computational server of the University Hospital of Lausanne. We thank UMCG Genomics Coordination Center, the UG Center for Information Technology, and their sponsors BBMRI-NL and TarGet for storage and compute infrastructure for LLDeep data. Finally, we thank Natàlia Pujol Gualdo and Triin Laisk for critical reading of the manuscript.

Declaration of interests

The authors declare no competing interests.

Received: February 22, 2022
Accepted: July 26, 2022

Web Resources

CNVPartition, https://support.illumina.com/downloads/genomestudio-2-0-plug-ins.html.
Genome STRiP, https://software.broadinstitute.org/software/genomestrip/.
PennCNV, http://penncnv.openbioinformatics.org/en/latest/.
QTLtools v.1.1, https://qtltools.github.io/qtltools/.

References

1. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.
2. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.
3. Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41, 321 subjects. Nat. Genet. *49*, 27–35.
4. Chapman, J., Rees, E., Harold, D., Ivanov, D., Gerrish, A., Sims, R., Hollingworth, P., Stretton, A., GERAD1 Consortium, and Holmans, P., et al. (2013). A genome-wide study shows a limited contribution of rare copy number variants to Alzheimer's disease risk. Hum. Mol. Genet. *22*, 816–824.
5. Gentile, G., La Cognata, V., and Cavallaro, S. (2021). The contribution of CNVs to the most common aging-related neurodegenerative diseases. Aging Clin. Exp. Res. *33*, 1187–1195.
6. Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. Genome Med. *1*, 62.
7. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. Am. J. Hum. Genet. *84*, 524–533.
8. Männik, K., Mägi, R., Macé, A., Cole, B., Guyatt, A.L., Shihab, H.A., Maillard, A.M., Alavere, H., Kolk, A., Reigo, A., et al. (2015). Copy number variations and cognitive phenotypes in unselected populations. JAMA *313*, 2044–2054.
9. Huguet, G., Schramm, C., Douard, E., Jiang, L., Labbe, A., Tihy, F., Mathonnet, G., Nizard, S., Lemyre, E., Mathieu, A., et al. (2018). Measuring and estimating the effect sizes of copy number variants on general intelligence in community-based samples. JAMA Psychiatr. *75*, 447–457.
10. Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K.M., Rees, E., Pardiñas, A.F., Einon, M., Escott-Price, V., Walters, J.T.R., O'Donovan, M.C., et al. (2019). Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. J. Med. Genet. *56*, 131–138.
11. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Porcu, E., Reymond, A., and Kutalik, Z. (2021). The individual and global impact of copy number variants on complex human traits. Preprint at medRxiv. https://doi.org/10.1101/2021.08.10.21261839.
12. Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide burden of copy-number variation in the UK biobank. Am. J. Hum. Genet. *105*, 373–383.
13. Li, Y.R., Glessner, J.T., Coe, B.P., Li, J., Mohebnasab, M., Chang, X., Connolly, J., Kao, C., Wei, Z., Bradfield, J., et al. (2020). Rare copy number variants in over 100, 000 European ancestry subjects reveal multiple disease associations. Nat. Commun. *11*, 255.
14. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila,

H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194.
15. Hujoel, M.L.A., Sherman, M.A., Barton, A.R., and Mukamel, R.E. (2021). Influences of rare copy number variation on human complex traits. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.21.465308.
16. Macé, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Nõukas, M., Sapkota, Y., Schick, U., Porcu, E., Rüeger, S., et al. (2017). CNV-association meta-analysis in 191, 161 European adults reveals new loci associated with anthropometric traits. Nat. Commun. *8*, 744.
17. Dellinger, A.E., Saw, S.-M., Goh, L.K., Seielstad, M., Young, T.L., and Li, Y.-J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. *38*, e105.
18. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.
19. Macé, A., Tuke, M.A., Beckmann, J.S., Lin, L., Jacquemont, S., Weedon, M.N., Reymond, A., and Kutalik, Z. (2016). New quality measure for SNP array based CNV detection. Bioinformatics *32*, 3298–3305.
20. Chettier, R., Ward, K., and Albertsen, H.M. (2014). Endometriosis is associated with rare copy number variants. PLoS One *9*, e103968.
21. Kendall, K.M., Rees, E., Escott-Price, V., Einon, M., Thomas, R., Hewitt, J., O'Donovan, M.C., Owen, M.J., Walters, J.T.R., and Kirov, G. (2017). Cognitive performance among carriers of pathogenic copy number variants: analysis of 152, 000 UK biobank subjects. Biol. Psychiatr. *82*, 103–110.
22. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. *35*, 2013–2025.
23. Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G.A., Thirlwell, C., Morris, T.J., Flanagan, A.M., Teschendorff, A.E., Kelly, J.D., and Beck, S. (2014). Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. *15*, R30.
24. Marzouka, N.-A.-D., Nordlund, J., Bäcklin, C.L., Lönnerholm, G., Syvänen, A.C., and Carlsson Almlöf, J. (2016). CopyNumber450kCancer: baseline correction for accurate copy number calling from the 450k methylation array. Bioinformatics *32*, 1080–1082.
25. Talevich, E., and Hunter Shain, A. (2018). CNVkit-RNA: Copy number inference from RNA-Sequencing data.
26. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. Int. J. Epidemiol. *44*, 1137–1147.
27. Lepik, K., Annilo, T., Kukuškina, V., eQTLGen Consortium, Kisand, K., Kutalik, Z., Peterson, P., and Peterson, H. (2017). C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. PLoS Comput. Biol. *13*, e1005766.
28. Tigchelaar, E.F., Zhernakova, A., Dekens, J.A.M., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M.A., Muñoz, A.M., Deelen, P., Cénit, M.C., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open *5*, e006772.
29. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.
30. Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., et al. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. *49*, 131–138.
31. Alwan, H., Pruijm, M., Ponte, B., Ackermann, D., Guessous, I., Ehret, G., Staessen, J.A., Asayama, K., Vuistiner, P., Younes, S.E., et al. (2014). Epidemiology of masked and white-coat hypertension: the family-based SKIPOGH study. PLoS One *9*, e92522.
32. Carmeli, C., Kutalik, Z., Mishra, P.P., Porcu, E., Delpierre, C., Delaneau, O., Kelly-Irving, M., Bochud, M., Dhayat, N.A., Ponte, B., et al. (2021). Gene regulation contributes to explain the impact of early life socioeconomic disadvantage on adult inflammatory levels in two cohort studies. Sci. Rep. *11*, 3100.
33. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
34. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. *40*, 1253–1260.
35. Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.
36. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.
37. Newman, S., Hermetz, K.E., Weckselblatt, B., and Rudd, M.K. (2015). Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. Am. J. Hum. Genet. *96*, 208–220.
38. Palta, P., Kaplinski, L., Nagirnaja, L., Veidenberg, A., Möls, M., Nelis, M., Esko, T., Metspalu, A., Laan, M., and Remm, M. (2015). Haplotype phasing and inheritance of copy number variants in nuclear families. PLoS One *10*, e0122713.
39. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. *29*, 512–520.
40. Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M.G.B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics *34*, 2781–2787.

HGGA, Volume *3*

Supplemental information

Omics-informed CNV calls

reduce false-positive rates

and improve power for CNV-trait associations

Maarja Lepamets, Chiara Auwerx, Margit Nõukas, Annique Claringbould, Eleonora Porcu, Mart Kals, Tuuli Jürgenson, Estonian Biobank Research Team, Andrew Paul Morris, Urmo Võsa, Murielle Bochud, Silvia Stringhini, Cisca Wijmenga, Lude Franke, Hedi Peterson, Jaak Vilo, Kaido Lepik, Reedik Mägi, and Zoltán Kutalik

Supplemental information

Contents

Supplemental Notes

S1. Estonian Biobank (EstBB): overview, CNV detection and quality control

Dataset overview

EstBB[1] is an Estonian population-based cohort of over 200,000 adult (aged ≥18 at recruitment) individuals recruited in years 2002-2020. All samples were genotyped using Illumina Global Screening Array (GSA) v1.0, GSA v2.0, GSA v2.0_ESTChip, and GSA v3.0_ESTChip2 arrays in 12 batches at the Core Genotyping Lab of the Institute of Genomics, University of Tartu. A subset of ~7,750 individuals recruited before 2012 have additionally been genotyped using Illumina Infinium OmniExpress-24 genotyping array. Another subset of ~2,500 individuals have whole-genome sequencing (WGS) data available[2]. The overlap between OmniExpress and WGS samples is ~1,000 individuals. Additionally, as part of smaller initiatives, RNA sequencing[3] and methylation (Infinium Human Methylation 450k Beadchip) data have been generated for subsets of the OmniExpress samples. Altogether, 1,066 OmniExpress samples have at least one out of WGS, methylation and RNA sequencing datasets available and are referred to as the EstBB multi-omics set or EstBB-MO.

Quality control of GSA and OmniExpress CNV samples

For both the OmniExpress and GSA datasets, we excluded samples with genotype call rate <98%, Hardy-Weinberg equilibrium test P-value $<1\times10^{-4}$ or mismatched sex based on chromosome X heterozygosity. Only one of each duplicated samples was retained. We created the intensity files (log R ratios (LRR) and B allele frequencies (BAF)) with Illumina GenomeStudio v2.0.4. For GSA samples, genotypes were re-clustered by manual realignment of cluster locations.

Autosomal copy number variations (CNVs) were called for 7,509 OmniExpress samples and 193,844 GSA samples using PennCNV detection software[4]. We denote the (true and false CNV) regions output by PennCNV as pCNV. The `hhall.hmm` (included in PennCNV software package) was chosen for the Hidden Markov Model (HMM) file and GC model file was created from `hg19.gc5Base.txt` file that was downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu). For OmniExpress, the population frequency of the B allele (PFB) file was created using all samples. Adjacent CNV calls were merged. We excluded all OmniExpress samples with >200 pCNV calls and with pCNV length >10Mb. For GSA, the PFB file was created based on 1,000 randomly selected individuals from the first batch. Again, adjacent CNV calls were merged. Two full genotyping batches were found to be outliers based on the intensity signal compared to the rest of the batches and, thus, were excluded.

Genotyping plates with >3 samples with either >200 called pCNVs or a total length of pCNV calls >10 Mb were excluded. Individual samples meeting these criteria were further removed. In total, 7,396 OmniExpress and 156,254 GSA samples were retained after filtering. The size of the sample overlap between the two datasets was 3,881 (out of which 39 samples belonged to the EstBB-MO subset).

Relatives extraction and exclusion

Relative pairs were detected using SNV genotypes with KING-robust kinship estimator[5] included in the PLINK 2.0 software (https://www.cog-genomics.org/plink/2.0/). We extracted monozygotic (MZ) twins (N=312 individuals; KING coefficient >0.354) and a subset of samples consisting of first-degree relatives (N=79,903 individuals; KING coefficient 0.177-0.354) from GSA samples. We excluded samples included in the EstBB-MO set from the OmniExpress samples and extracted the relative pairs with KING coefficient >0.177 (N=504 individuals). In order to create a GSA association study dataset, we excluded one sample of each pair with kinship coefficient >0.0884. This resulted in 89,516 unrelated GSA samples identical to what is used in[6].

Phenotype data

We extracted four anthropometric measurements (body mass index (BMI), height, weight and waist-to-hip ratio (WHR)) for each of the GSA association study samples. BMI, height and weight were collected during the biobank recruitment, while WHR was parsed from health registries and doctors' notes. If several WHR measurements were available, the most recent was retained. Phenotypes were inverse normal transformed and residualised on sex, age, age^2, genotyping batch and first 20 principal components.

S2. Lifelines Deep (LLDeep)

LLDeep is a subset of ~1,500 unrelated deeply phenotyped samples from the Dutch Lifelines cohort. All samples were genotyped using HumanCytoSNP-12 array and have RNA sequencing[7] and methylation (Infinium Human Methylation 450k Beadchip)[8] data available. The genotype dataset, as well as the quality control steps of omics data, are described elsewhere[9]. CNVs were detected in two batches of 865 and 522 samples. The `hhall.hmm` was chosen for the HMM file and GC correction was applied. The full dataset was included in the analyses as no individuals had >200 pCNVs or pCNVs longer than 10Mb.

S3. Swiss Kidney Project on Genes in Hypertension (SkiPOGH)

SkiPOGH is a family and population-based study of genetic determinants of blood pressure and renal function[10,11]. The samples were collected between 2009 and 2013 from the cantons

of Bern and Geneva, and from the city of Lausanne. It contains 1,128 adult (aged ≥18) samples from 274 families (mostly trios) and is part of a larger European Project on Genes in Hypertension (EPOGH) study. All samples were genotyped using a dense Illumina 2.5 array. CNVs were detected in one batch of 675 samples using PennCNV software. The `hhall.hmm` was chosen for the HMM file and GC correction was applied. All carriers with >200 pCNVs or pCNVs longer than 10Mb were excluded resulting in 466 samples. Out of these, 405 had gene expression and 148 had methylation data available[12]. A separate set was compiled with all parent-child pairs that meet the CNV filtering criteria (319 samples from 102 families).

S4. UK Biobank (UKB): overview, CNV detection and filtering

Dataset overview

UKB is a population-based cohort of ~500,000 individuals from United Kingdom aged between 40 and 69 at recruitment. The majority of the samples (~450,000) are genotyped on Affymetrix UK Biobank Axiom array while the rest (~50,000) are genotyped on Affymetrix UK BiLEVE Axiom array. The dataset and general genotype quality control steps have been described by [13]. Although (exome) sequencing data is now available for a subset of UKB samples, this was not the case at the time of most of our analyses.

CNV detection and quality control for familial analysis

CNVs were detected in 106 batches using PennCNV. Individual specific intensity files containing B allele frequencies (BAF) and log R ratios (LRR) per probe were created using PennCNV-Affy conversion pipeline prior CNV detection. The PFB files were created for each batch separately, using 250 randomly selected samples per batch. The `hhall.hmm` was used and no GC correction was performed. Finally, adjacent CNV calls were merged. All samples with >200 pCNVs or pCNV total length >10Mb were excluded. Altogether, 401,571 samples were retained. Relative pairs were calculated using KING-robust kinship estimator[5]. Analogously to EstBB-GSA samples, we extracted MZ twins (N=302; coefficient >0.354) and first-degree relatives (N=42,032; KING coefficient 0.177-0.354).

CNV detection and quality control for association analyses

The subset of UKB samples and anthropometric phenotypes used for association analysis was prepared in a separate pipeline as described in [6]. Namely, the CNVs were called in 106 batches using Affymetrix genome-wide 6.0 array HMM file and GC correction. Samples genotyped on plates with mean pCNV count per sample >100 were excluded. Samples with >200 pCNVs or a single pCNV >10Mb were further removed. After these quality control steps, 331,522 unrelated British samples were retained. Four anthropometric traits were extracted

and inverse normal transformed prior correction for sex, age, age^2, genotyping batch, and PC1-40.

S5. CNV detection from whole-genome sequencing reads

The Genome STRiP pipeline[14] was used to call CNVs in five separate batches for EstBB-MO samples. Eleven samples with excessive number of calls (#calls/#samples > median (across all samples) + 3 median absolute deviation) were removed. The union of the discovered sites was genotyped with Genome STRiP SVGenotyper module in all batches separately and merged. Duplicate calls were removed using the standard Genome STRiP duplicate removal settings, involving site overlap greater than 50% and duplicate score (logarithm of odds (LOD) score of genotype concordance at most discordant sample) greater than zero. In addition, low-quality (LQ) CNVs and CNVs with call rate less than 90% were excluded. Additionally, the pipeline excludes deletions shorter than 1,000bp and duplications shorter than 2,000bp.

S6. Preparations of methylation data

Methylation data from Infinium Human Methylation 450k Beadchip was collected for EstBB-MO, LLDeep and SkiPOGH samples. We extracted the methylation data matrices containing methylated and un-methylated intensities from sample-specific *idat* files using R package *minfi*[15] and summed those matrices to an overall intensity matrix. We eliminated all Type I methylation probes (N=135,476), since these cannot be used for capturing overall summed intensity, and only used Type II probes (N=350,036) in our study. All intensities that had detection P-values > 1×10^{-16} were marked as missing and all probes with >5% missingness were filtered out. We corrected the overall intensity of each probe for age and sex. In EstBB-MO and SkiPOGH we additionally corrected for first four genotype principal components (PCs).

To exclude as much noise as possible while retaining the effect of CNVs on methylation data, we tested correcting the data for several numbers of methylation intensity PCs, ranging from 0 to 200 (using the following fixed categories: 0, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100, 150 and 200). CNV quality metrics based on methylation data (*MET* metric; as described in the main text) were calculated for each set of PCs. In each dataset, we chose the best set of PCs by maximising correlation with previously published consensus-based CNV quality score (cQS) [16] or *WGS* metric in EstBB-MO samples. The best number of PCs for *MET* metric was 30 for EstBB-MO deletions, 10 for EstBB-MO duplications, 100 for LLDeep deletions, 30 for LLDeep duplications, 1 for SkiPOGH deletions and 4 for SkiPOGH duplications (see **Figure S1**). Note that results obtained with different numbers of PCs did not vary substantially.

S7. Preparations of gene expression data

RNA-sequencing counts were obtained for EstBB-MO, LLDeep and SkiPOGH samples. We further processed the data as follows: (1) we removed genes with low or no expression in the majority of individuals by requiring for each gene to have ≥ 5 individuals with a count per million (cpm) value greater than 0.5; (2) we normalized remaining genes by weighted trimmed mean of M-values; (3) we calculated the log(cpm) values of each gene, and Z-transformed these values; (4) we corrected the resulting gene expression measures for covariates: age, sex, blood component values, and batch. In EstBB-MO and SkiPOGH we additionally corrected for genotyping PC1-4.

Analogously to methylation data, we corrected the remaining residual expression of each gene for up to 200 expression PCs (calculated on the residuals) to further clean the signal. Additionally, we corrected expression residuals of each gene for the gene's independent SNP *cis*-eQTL effects (P-value < 0.05) in EstBB-MO and LLDeep. For EstBB-MO set the *cis*-eQTL analysis was run on SNPs within 500kbp proximity to the transcription start site using QTLtools v1.1[17]. For LLDeep we used the already published eQTLs[7]. In SkiPOGH we did not correct for eQTLs. We calculated the scores both before and after eQTL corrections and saw that correlations obtained after corrections were slightly higher. We also tested only including genes that showed R>0.1 correlation with copy number in [18].

Gene expression based CNV quality metric (*GE* metric) was calculated for each set of PCs and filters. We only considered genes that had at least 80% overlap with a CNV. The best set was chosen by maximising the correlation between *GE* metric and previously published cQS[16] or *WGS* metric in EstBB-MO samples. The number of selected PCs for *GE* metric was 30 for EstBB-MO deletions, 4 for EstBB-MO duplications, 30 for LLDeep deletions, 50 for LLDeep duplications, 40 for SkiPOGH deletions and 30 for SkiPOGH duplications (**Figure S2**).

S8. CNV quality modelling

We used the stepwise regression with forward selection as an algorithm to pick the best parameters into the final omics-informed CNV quality score (OQS) model. Our first objective was to compile the initial parameter set. We did that using the PennCNV output, which contains a set of characteristics and parameters for each of the detected CNV regions. All parameters are described in **Table S2**. Altogether, we tested eight different models with slightly varying initial parameter sets. First two initial parameter sets we tested included all PennCNV output variables as was done before[16]. The remaining initial sets were compiled using subsets of PennCNV output variables and additional variables calculated based on PennCNV output. Initial sets per model are shown in **Table S3**. The models were built for three metrics (GE, MET and Combined) in LLDeep dataset and four metrics (WGS, GE, MET and

Combined) in EstBB-MO dataset, resulting in 24 and 32 (not necessarily unique) models, respectively, for both deletions and duplications.

During each model-building step of the stepwise forward selection process, the following parameters remaining in the initial parameter set were tested:

a. PennCNV output (or derived) parameters not yet in the model;
b. interactions between a CNV-specific and a sample-specific parameter already in the model;
c. (if allowed) higher order terms of parameters already in the model.

An extra term was added to the model if it minimized the average mean square error (MSE) from *k*-fold cross-validation. If no term minimized the MSE, the algorithm stopped and returned the existing model. Models were built separately for deletions and duplications.

References

1. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.

2. Mitt, M., Kals, M., Pärn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., Esko, T., et al. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur. J. Hum. Genet. *25*, 869–876.

3. Lepik, K., Annilo, T., Kukuškina, V., eQTLGen Consortium, Kisand, K., Kutalik, Z., Peterson, P., and Peterson, H. (2017). C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. PLoS Comput. Biol. *13*, e1005766.

4. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

5. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

6. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Porcu, E., Reymond, A., and Kutalik, Z. (2021). The individual and global impact of copy number variants on complex human traits (medRxiv).

7. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

8. Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., et al. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. *49*, 131–138.

9. Tigchelaar, E.F., Zhernakova, A., Dekens, J.A.M., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M.A., Muñoz, A.M., Deelen, P., Cénit, M.C., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open *5*, e006772.

10. Alwan, H., Pruijm, M., Ponte, B., Ackermann, D., Guessous, I., Ehret, G., Staessen, J.A., Asayama, K., Vuistiner, P., Younes, S.E., et al. (2014). Epidemiology of masked and white-coat hypertension: the family-based SKIPOGH study. PLoS One *9*, e92522.

11. Rousson, V., Ackermann, D., Ponte, B., Pruijm, M., Guessous, I., d'Uscio, C.H., Ehret, G., Escher, G., Pechère-Bertschi, A., Groessl, M., et al. (2021). Sex- and age-specific reference intervals for diagnostic ratios reflecting relative activity of steroidogenic enzymes and pathways in adults. PLoS One *16*, e0253975.

12. Carmeli, C., Kutalik, Z., Mishra, P.P., Porcu, E., Delpierre, C., Delaneau, O., Kelly-Irving, M., Bochud, M., Dhayat, N.A., Ponte, B., et al. (2021). Gene regulation contributes to explain the impact of early life socioeconomic disadvantage on adult inflammatory levels in two cohort studies. Sci. Rep. *11*, 3100.

13. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

14. Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.

15. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

16. Macé, A., Tuke, M.A., Beckmann, J.S., Lin, L., Jacquemont, S., Weedon, M.N., Reymond, A., and Kutalik, Z. (2016). New quality measure for SNP array based CNV detection. Bioinformatics *32*, 3298–3305.

17. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat. Commun. *8*, 15452.

18. Talevich, E., and Hunter Shain, A. (2018). CNVkit-RNA: Copy number inference from RNA-Sequencing data.

19. Palta, P., Kaplinski, L., Nagirnaja, L., Veidenberg, A., Möls, M., Nelis, M., Esko, T., Metspalu, A., Laan, M., and Remm, M. (2015). Haplotype phasing and inheritance of copy number variants in nuclear families. PLoS One *10*, e0122713.

20. Chettier, R., Ward, K., and Albertsen, H.M. (2014). Endometriosis is associated with rare copy number variants. PLoS One *9*, e103968.

21. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. *29*, 512–520.

22. Macé, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Nõukas, M., Sapkota, Y., Schick, U., Porcu, E., Rüeger, S., et al. (2017). CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat. Commun. *8*, 744.

Supplemental Tables

Table S1. Mean (and standard deviation (SD)) of raw PennCNV parameters per cohort/platform

Table characterise the quality controlled dataset (autosomes only) after the exclusion of samples with >200 CNV calls or calls larger than 10Mbp.

PARAMETERS	EstBB-MO	EstBB-GSA	LLDeep	SkiPOGH	UKB
Array	Illumina Infimum Omniexpress-24	Illumina Global Screening Array	Illumina Infimum Human CytoSNP-12	Illumina 2.5	Affimetrix UK BiLEVE/ Biobank Axiom
# of array markers	730k	760k	300k	2.5M	820k
mean LRR	-0.0064 (0.019)	-0.0027 (0.0076)	-0.0042 (0.013)	-0.0022 (0.0068)	0.0049 (0.0096)
SD of LRR	0.13 (0.052)	0.11 (0.020)	0.11 (0.038)	0.16 (0.021)	0.28 (0.040)
mean BAF	0.50 (0.0027)	0.50 (0.0012)	0.50 (0.0048)	0.50 (0.00080)	0.50 (0.00047)
SD of BAF	0.040 (0.010)	0.036 (0.0044)	0.033 (0.013)	0.037 (0.0028)	0.029 (0.0036)
BAF drift	0.00066 (0.0033)	0.00035 (0.00023)	0.00017 (0.00069)	0.00044 (0.00024)	0.00008 (0.00004)
absolute waviness factor	0.011 (0.0077)	0.0073 (0.0022)	0.014 (0.0063)	0.024 (0.0051)	0.014 (0.0030)
DELETIONS					
CNV count per sample	17.029 (13.99)	9.32 (12.17)	5.46 (5.74)	39.83 (31.07)	2.57 (3.18)
CNV length	33,583 (82,122)	31,994 (86,139)	64,992 (127,927)	14,773 (51,723)	98,336 (187,287)
number of probes	9.81 (18.11)	10.12 (17.83)	7.33 (10.14)	17.54 (28.39)	32.94 (42.49)
confidence value	26.45 (51.68)	27.28 (43.75)	23.05 (29.80)	29.61 (48.11)	34.99 (50.85)
DUPLICATIONS					
CNV count per sample	9.25 (15.13)	8.04 (5.29)	3.74 (3.32)	51.91 (31.56)	1.52 (2.74)
CNV length	58,401 (184,153)	61,117 (170,842)	97,586 (179,278)	40,991 (128,057)	251,993 (388,058)
number of probes	15.04 (51.76)	14.50 (29.15)	11.16 (17.88)	21.15 (44.20)	88.97 (86.32)
confidence value	32.16 (67.74)	26.63 (39.21)	33.27 (54.71)	21.57 (51.95)	33.19 (45.12)

Table S2. **Parameters used for CNV quality modelling**

BAF – B allele frequency, LRR – log R ratio.

Parameter	**Sample/variant-specific**	**Obtained**	**Variable name**
mean BAF	sample	from PennCNV output	`BAF_mean`
standard deviation of BAF	”	”	`BAF_SD`
BAF drift	”	”	`BAF_drift`
mean LRR	”	”	`LRR_mean`
standard deviation of LRR	”	”	`LRR_SD`
absolute waviness factor	”	”	`WF`
number of variant calls	”	”	`NumCNV`
variant confidence	variant	”	`Max_Log_BF`
variant length	”	”	`Length_bp`
number of probes	”	”	`No_probes`
variant length per probe	”	calculated as `Length_bp / No_probes`	`Length_per_Probe`
number of variant calls corrected for genotyping array size	sample	calculated as `NumCNV /` number of probes on array since number of variants per sample directly depends on array density	`NumCNV_corr`
corrected and dichotomised number of variant calls	sample	We noticed in several independent datasets that the CNV quality starts to decline rapidly for samples with number of CNVs (corrected for array density) over a certain threshold. We used ROC curves in order to pinpoint this threshold and found that in case of deletions, its value falls into a narrow range of 3.8×10^{-5}...3.9×10^{-5} dependent on the dataset (**Figure S4**). Therefore, we compiled a parameter for deletions that equals zero if `NumCNV_corr` is below and one if it is above the given threshold.	`NumCNV_bin`

Table S3. Initial sets of explanatory variables included in the step-wise model selection algorithm

The explanatory variables are either output parameters of PennCNV software or directly calculated from these output parameters (**Table S2**). Two out of eight models allow higher order terms. All models allow interactions between sample-specific and CNV-specific parameters. Model building is described in **Supplemental Note S8**. The final set of parameters selected to the model depends on the dataset.

Model No.	Initial parameter set	Higher terms
1	Sample-specific parameters, CNV confidence score, NumCNV, Length_bp, No_Probes	No
2	Sample-specific parameters, CNV confidence score, NumCNV, Length_bp, No_Probes	Yes
3	Sample-specific parameters, CNV confidence score, NumCNV, Length_per_Probe	Yes
4	Sample-specific parameters, CNV confidence score, NumCNV	No
5	Sample-specific parameters, CNV confidence score, NumCNV_corr, Length_bp, No_Probes	No
6	Sample-specific parameters, CNV confidence score, NumCNV_bin, Length_bp, No_Probes	No
7	Sample-specific parameters, CNV confidence score, NumCNV_bin, Length_per_Probe	No
8	Sample-specific parameters, CNV confidence score, NumCNV_bin	No

Table S4. CNV quality filtering thresholds used for PennCNV output in publications

	Wang *et al.*[4]	Palta *et al.*[19]	Chettier *et al.*[20]	Pinto *et al.*[21]
LRR sd	<0.3	≤0.25	<0.24	≤0.27
BAF drift	<0.01	≤0.002		
BAF sd		≤0.05		≤0.13
|WF|	<0.05	≤0.04	<0.05	
No. CNVs	<100			
Call rate			>0.99	≥0.98
Confidence		≥5		
No. Probes			≥10	≥5
Length		≥1 kb		≥1 kb

Table S5. 21 CNV region–phenotype pairs analysed in the EstBB-GSA and UKB cohorts

All pairs reached an association P-value < 1×10^{-4} in previously published study[22], using cQS[16], in at least one analysis type (mirror/ deletion only/ duplication only; flagged with "+"). In the EstBB-GSA and UKB columns, the "+" sign indicates that the corresponding association was also significant (for at least one probe in the CNV region) in the respective analysis of our current study. In this case the significance threshold was set to P-value < 0.05/21 = 2.38×10^{-3} (using raw PennCNV calls). *In the EstBB-GSA, the CNV region on chromosome 18 contained a probe with P < 2.38e-3 but the CNV structure in the region and the association with BMI/weight was clearly not the same as reported before [22]. Therefore, this region was not included in the EstBB-GSA follow-up analyses (**Figure S6**).

Phenotype	CNV region (hg37)	Macé et al. (2017)			EstBB-GSA			UKB		
		Deletions	Mirror	Duplications	Deletions	Mirror	Duplications	Deletions	Mirror	Duplications
BMI	16:28820000-29040000	+	+			+		+	+	
	16:29590000-30200000	+	+		+	+		+	+	
	18:57660000-57900000	+	+		+*	+*		+	+	
	22:19030000-20310000		+	+			+		+	+
	22:20740000-21000000		+	+					+	+
Weight	1:146530000-147430000	+	+					+	+	
	2:111580000-111780000		+						+	
	2:112680000-112780000		+						+	
	3:196220000-196420000		+							
	16:28820000-29040000	+	+			+		+	+	
	16:29590000-30200000	+	+		+	+		+	+	
	18:57660000-57900000	+	+		+*	+*		+	+	
	22:20740000-21000000		+	+		+			+	+
Height	1:146530000-147430000	+	+		+	+		+	+	
	3:196220000-196420000		+						+	
	3:196710000-196920000		+						+	
	11:27010000-27230000	+	+		+	+				
	15:22780000-23070000	+	+					+	+	
	16:29590000-30200000	+	+		+	+		+	+	
WHR	7:73020000-73110000		+	+						
	16:29590000-30200000	+	+		+	+		+	+	

Table S6. Number of PennCNV calls evaluated per omics layer

The number and percentage of PennCNV calls (pCNVs) that could be evaluated using methylation (*MET*), gene expression (*GE*) and whole-genome sequencing (*WGS*) based quality metrics in EstBB-MO, LLDeep and SkiPOGH datasets. The combined metric (*EXTR*) requires the availability of at least two out of three omics based metrics. The percentage is calculated as the fraction of pCNVs evaluated amongst all autosomal pCNVs of the set of individuals with respective omics dataset available (after sample exclusion in quality control). In case of *GE* metric, 80% gene-pCNV overlap is required.

	EstBB-MO	**LLDeep**	**SkiPOGH**
***MET* metric**	3,228 (41.8%)	3,476 (68.7%)	5,048 (44.9%)
***GE* metric**	462 (4.7%)	706 (7.1%)	1,497 (3.8%)
***WGS* metric**	23,977 (100%)	-	-
Combined metric (*EXTR*)	3,496	441	410

Table S7. Number of pCNVs per combined (*EXTR*) metric range in three datasets

Metric range	EstBB-MO		LLDeep		SkiPOGH	
	Deletions	Duplications	Deletions	Duplications	Deletions	Duplications
[0;0.1]	828	829	110	63	163	107
(0.1;0.2]	36	33	9	2	13	8
(0.2;0.3]	10	15	1	2	1	6
(0.3;0.4]	7	19	0	0	0	3
(0.4;0.5]	2	6	0	1	2	0
(0.5;0.6]	2	9	1	0	0	0
(0.6;0.7]	23	33	0	2	0	3
(0.7;0.8]	45	77	3	6	1	5
(0.8;0.9]	71	129	2	20	0	11
(0.9;1.0]	726	596	92	127	50	37

Table S8. Mean scores for familial and non-familial pCNVs calculated using models built on Estonian OmniExpress samples

The models are tested on monozygotic twins of UK Biobank and EstBB-GSA samples, as well as parent-child pairs form the SkiPOGH dataset. Altogether we tested eight models (**Table S3**) for three omics-based metrics (*WGS* – whole-genome sequencing metric, *MET* – methylation metric, *GE* – gene expression metric) and a combined metric (*EXTR*). The best model was selected as the one that maximises the difference between familial and non-familial pCNVs over three independent datasets. (Table in a separate .xlsx file)

Table S9. Mean scores for familial and non-familial pCNVs calculated using models built on LLDeep samples.

Models are tested on monozygotic twins of UK Biobank and EstBB-GSA samples, as well as parent-child pairs form SkiPOGH dataset. Altogether we tested eight models (**Table S3**) for two omics-based metrics (*MET* – methylation metric, *GE* – gene expression metric) and a combined metric (*EXTR*). The best model was selected as the one that maximises the difference between familial and non-familial pCNVs over three independent datasets. The best model is highlighed in blue. (Table in a separate .xlsx file)

Table S10. Description of best omics-informed CNV quality model for deletions

All parameters are explained in **Table S2**. 'X:Y' denotes the interaction term between parameters X and Y.

Parameter	Effect size (beta)
(Intercept)	-2.12350638759242
Max_Log_BF	0.138074328580999
LRR_mean	192.33129681258
NumCNV_bin	-1.17079719833943
Max_Log_BF:LRR_mean	-6.07975553625751

Table S11. Description of best omics-informed CNV quality model for duplications

All parameters are explained in **Table S2**. 'X:Y' denotes the interaction term between parameters X and Y.

Parameter	Effect size (beta)
(Intercept)	-114.039850181999
Max_Log_BF	-0.0368788122813245
Length_bp	-9.76978925613445e-06
LRR_SD	-34.7848931008816
BAF_mean	233.943409217459
BAF_drift	-2032.95959074771
Max_Log_BF:LRR_SD	0.60672686014698
Length_bp:LRR_SD	0.000111986438009925

Table S12. CNV associations analysis results in EstBB-GSA dataset

Results for seven settings (raw PennCNV, four published quality filtering approaches[4,19–21], cQS[16] and omics-informed quality score (OQS)) in EstBB-GSA dataset. 21 CNV region-phenotype pairs ($P<1\times10^{-4}$; [22]) are included in the analysis (**Table S5**). (Table in a separate .xlsx file)

Table S13. CNV associations analysis results in UKB dataset

Results for seven settings (raw PennCNV, four published quality filtering approaches[4,19–21], cQS[16] and omics-informed quality score (OQS)) in UKB dataset. 21 CNV region-phenotype pairs ($P<1\times10^{-4}$; [22]) are included in the analysis (**Table S5**). (Table in a separate .xlsx file)

Table S14. Average F statistic values per phenotype and dataset

F statistics characterise the change in explained variance when comparing OQS to raw PennCNV, four filtering setting used in publications[4,19–21] and cQS[16] in two datasets (EstBB-GSA / UKB). F>1 means an increase and F<1 a decrease in explained variance (i.e. F=1.16 indicates a 16% increase in explained variance). The values presented in this table are calculated as an average over 20 runs. None of the values reach even nominal statistical significance (P<0.05) due to the low number of independent associated regions per phenotype.

Phenotype	Analysis type	F (raw PennCNV)	F (Wang et al, 2007[4])	F (Palta et al, 2015[19])	F (Chettier et al, 2014[20])	F (Pinto et al, 2011[21])	F (cQS (Macé et al, 2016[16]))
BMI	mirror	1.17 / 1.02	1.21 / 1.13	1.17 / 2.53	1.18 / 3.23	1.19 / 1.52	1.55 / 0.97
Height	mirror	1.16 / 1.05	1.21 / 1.21	1.17 / 2.83	1.03 / 4.40	1.10 / 1.74	1.23 / 0.83
Weight	mirror	1.02 / 0.99	0.99 / 1.17	1.05 / 3.91	1.12 / 4.58	1.07 / 1.79	1.49 / 1.09
WHR	mirror	1.34 / 1.10	1.56 / 1.06	1.15 / 3.46	1.15 / 5.11	1.16 / 2.64	1.40 / 0.85
BMI	deletion-only	1.19 / 1.02	1.27 / 1.15	1.23 / 2.64	1.12 / 3.32	1.18 / 1.44	1.30 / 0.93
Height	deletion-only	1.08 / 0.97	1.11 / 1.27	1.13 / 3.35	1.01 / 5.60	1.07 / 1.74	1.09 / 0.96
Weight	deletion-only	1.33 / 1.09	1.39 / 1.10	1.34 / 2.65	1.17 / 3.43	1.33 / 1.47	1.46 / 0.88
WHR	deletion-only	1.13 / 1.10	1.42 / 1.14	1.11 / 3.98	1.04 / 4.76	1.10 / 1.81	1.23 / 0.82
BMI	duplication-only	0.85 / 1.00	0.76 / 1.06	1.71 / 5.37	0.93 / 4.85	0.91 / 2.08	1.43 / 1.01
Weight	duplication-only	. / 1.03	. / 0.97	. / 4.69	. / 2.76	. / 1.99	. / 1.02

Supplemental Figures

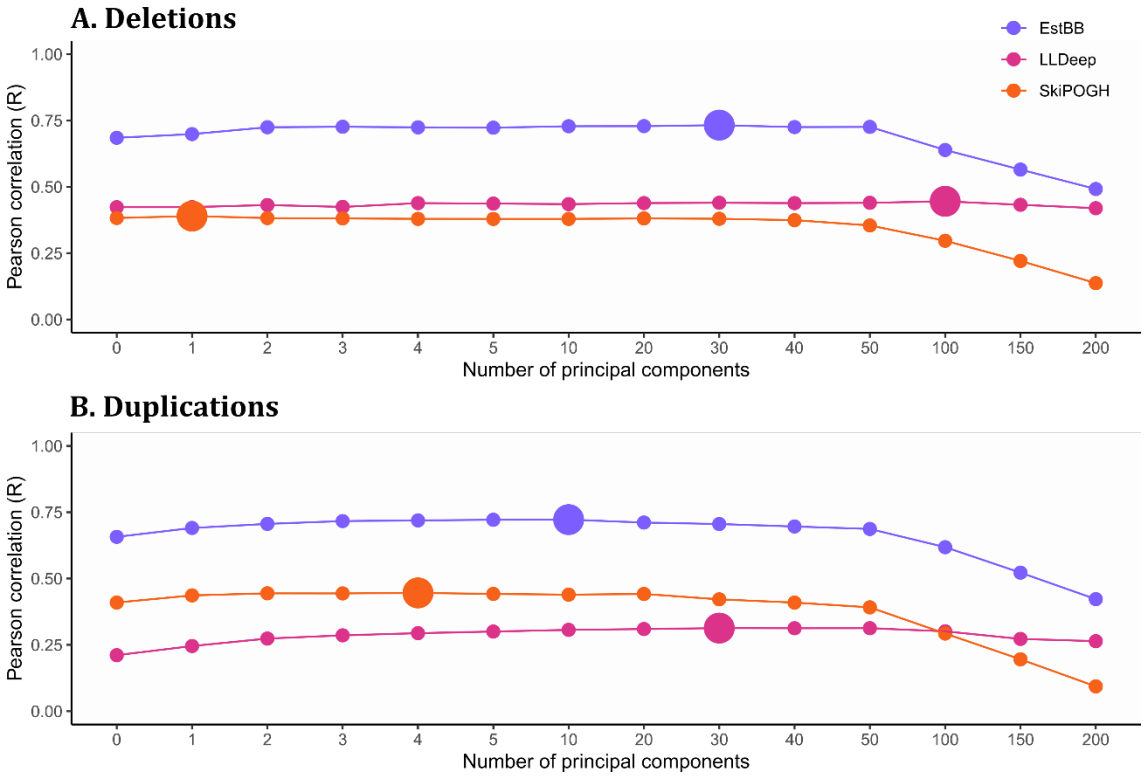

Figure S1. Methylation metric correlations per number of principal components

Methylation metrics for (A) deletion and (B) duplication quality (*MET*) calculated using 0 to 200 methylation intensity principal components (PCs; **Supplemental Note S6**) and correlated (Pearson's correlation) with *WGS* metric in EstBB-MO and cQS[16] in LLDeep and SkiPOGH. The number of PCs chosen for further analyses are indicated with large points. We used *MET* values corrected for 30 PCs for EstBB-MO deletions, 10 PCs for EstBB-MO duplications, 100 PCs for LLDeep deletions, 30 PCs for LLDeep duplications, 1 PCs for SkiPOGH deletions and 4 PCs for SkiPOGH duplications, as these maximised the corresponding correlations in the respective datasets.

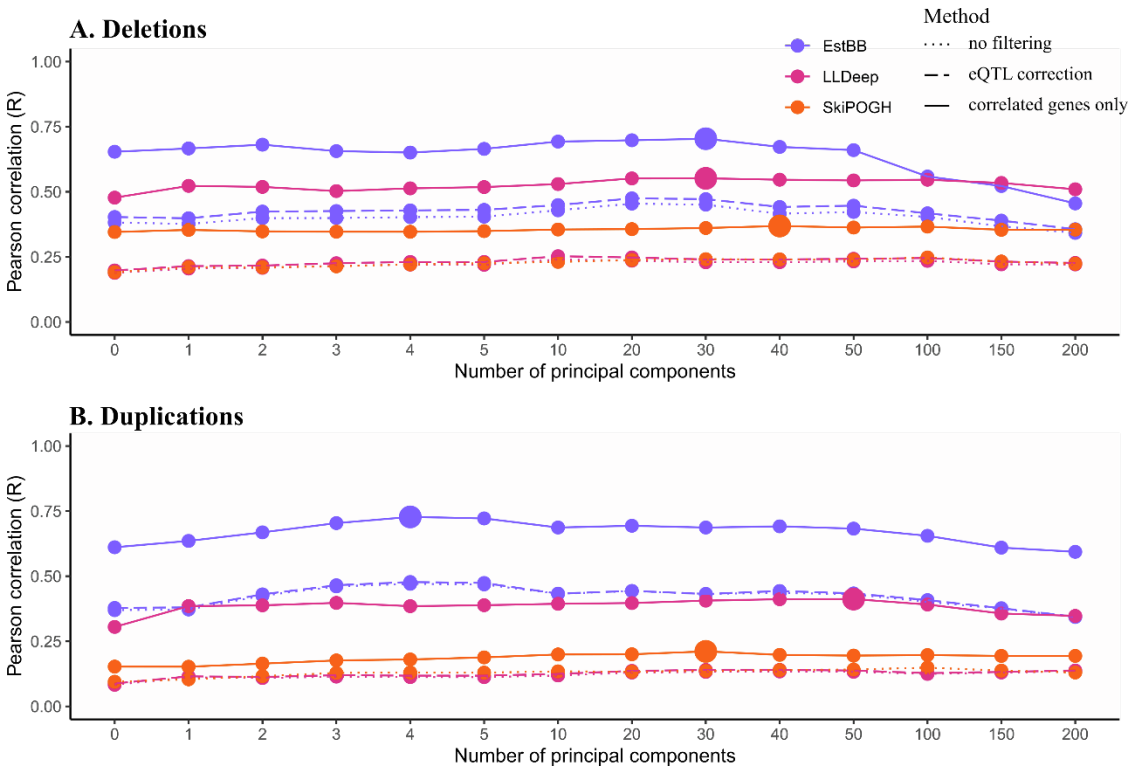

Figure S2. Gene expression metric correlations per number of principal components

Gene expression metrics for (A) deletion and (B) duplication quality (*GE*) calculated using 0 to 200 expression principal components (PCs; dotted line). Additionally we tested correcting gene expression for eQTLs prior score calculations (dashed line) and only including genes which showed correlation to CNV status in [18] (solid line; **Supplemental Note S7**). We correlated (Pearson's correlation) *GE* values with *WGS* metric in EstBB-MO and cQS[16] in LLDeep and SkiPOGH. In most cases, correcting for eQTLs slightly increased the correlation while gene filtering improved the correlations significantly. The number of PCs chosen for further analyses are indicated with large points. We used *GE* metric values corrected for eQTLs and 30 PCs for EstBB-MO deletions, 4 PCs for EstBB-MO duplications, 30 PCs for LLDeep deletions, 50 PCs for LLDeep duplications, 40 PCs for SkiPOGH deletions and 30 PCs for SkiPOGH duplications, as these maximised the corresponding correlations in the respective datasets.

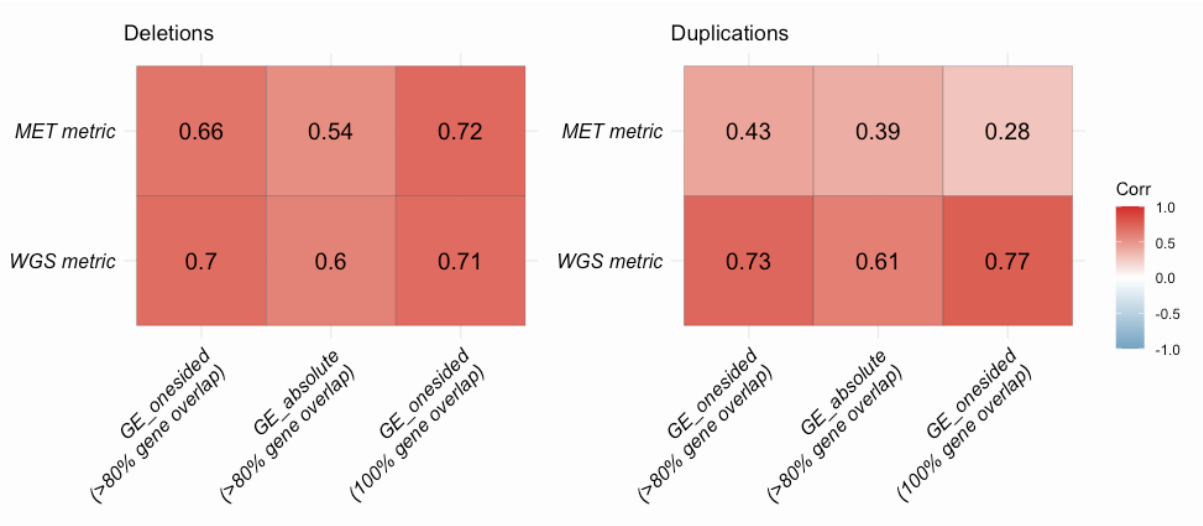

Figure S3. Pearson correlations between WGS/MET metrics and GE metrics calculated using three different approaches

GE_onesided refers to the approach for GE calculation where GE>0 for deletions and GE<0 for duplications were set to zero. *GE_absolute* refers to the approach where GE can take both positive and negative values and its absolute values are used in the correlation calculations.

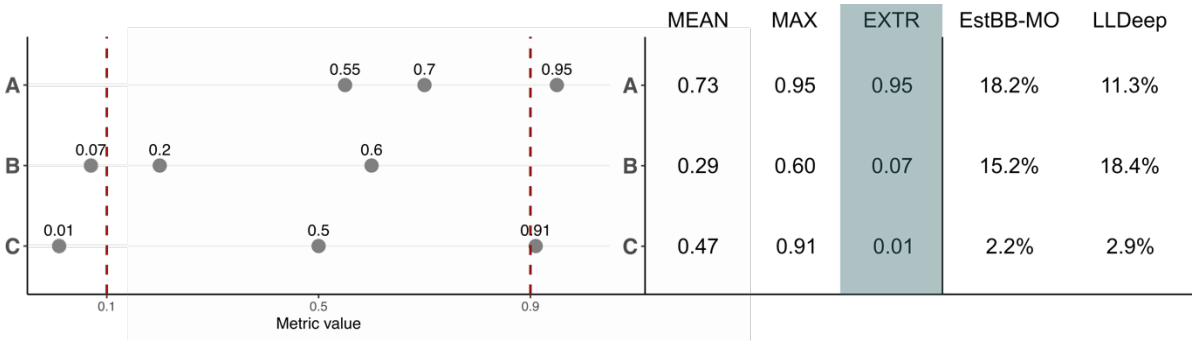

Figure S4. Comparison of omics-based combined metrics

For each PennCNV call (pCNV), we calculated up to three omics-based CNV quality metrics (*GE*, *MET* and *WGS*). Our aim was to combine them into one metric per pCNV. We considered using the mean (*MEAN*), maximum (*MAX*) and the most extreme (i.e., furthest from 0.5, *EXTR*) metric. We divided the metric value range into three subranges: false positive [0; 0.1), ambiguous [0.1; 0.9] and true positive (0.9, 1]. For the majority of pCNVs (64.4% in EstBB-MO (full set N=3,496) and 67.3% in LLDeep (N=441)) all omics-based metrics were in close agreement with all values falling into the same subrange. For these cases the exact choice of combined metric makes little difference. Here we bring three toy examples of cases where the metrics do not agree: (A) at least one metric indicates a true positive CNV while others are ambiguous; (B) at least one metric indicates a false positive CNV while others are ambiguous; (C) at least one metric indicates a false positive CNV and at least one other a true positive CNV. Each example is followed by the values of corresponding combined metrics and the percentages on pCNVs that fall under these example categories in EstBB-MO and LLDeep. The usage of *MEAN* metric dilutes the effect of omics-based metrics that clearly indicate the presence of a true or false positive CNV, which is only reasonable in example C (2.2%-2.9% of cases). The usage of *MAX* metric is only intuitive when we expect the CNV to be true positive (example A; 11.3%-18.2% of cases). The usage of *EXTR* metric is intuitive in both examples A and B (29.7%-33.4% of cases) and even in example C it produces correct estimations in roughly half of the cases. Therefore, we chose to use *EXTR* as our combined metric per pCNV.

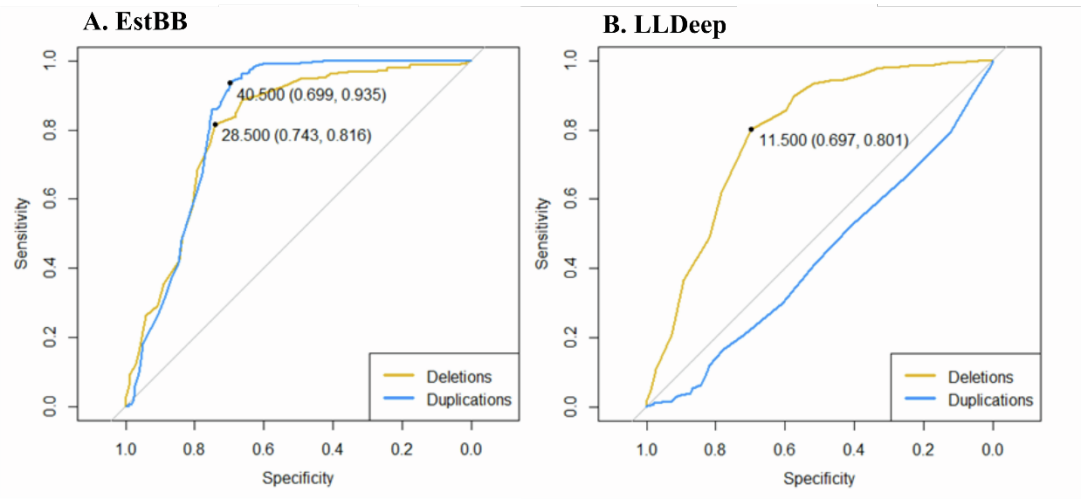

Figure S5. Number of pCNVs per sample as pCNV quality predictor

ROC curves built on (A) EstBB-MO and (B) LLDeep pCNVs with number of pCNVs per person as predictor. The true and false calls are defined as methylation metric *MET* >0.9 and <0.1, respectively. We used *MET* metric as it was present in both datasets in greater extent that combined metrics (the results for combined metric did not differ significantly). For duplications (blue) in EstBB-MO and deletions (yellow) in both datasets, there is a clear cut-off for the number of pCNVs per sample that can be used to discriminate between true and false calls. These cut-offs values are presented as dots on the ROC curves (followed by the corresponding specificity and sensitivity values). The exact cut-off is not generalisable as it is dependent on array density (~730k for EstBB-MO and ~300k for LLDeep). After correction for array density, the values for deletions are comparable: 3.9×10^{-5} in EstBB and 3.8×10^{-5} in LLDeep. We used these values to create a binary variable out of number of CNVs corrected for array density.

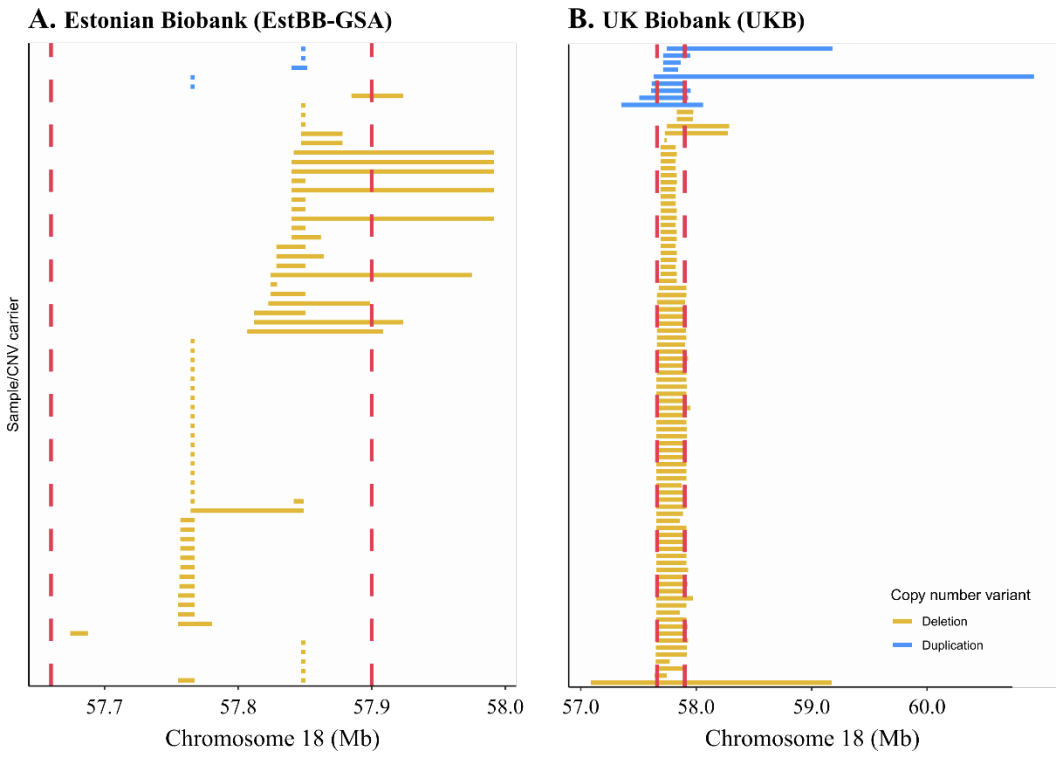

Figure S6. 18q21.32 CNV region in EstBB-GSA and UKB datasets

In EstBB-GSA analyses we excluded the 18q21.32 CNV previously associated with body mass index and weight[22]. Although there were EstBB-GSA CNVs (A) overlapping the region of interest (indicated with red dashed lines), they are not the CNVs for which the association was detected. In UKB (B) this region was included in the association tests.

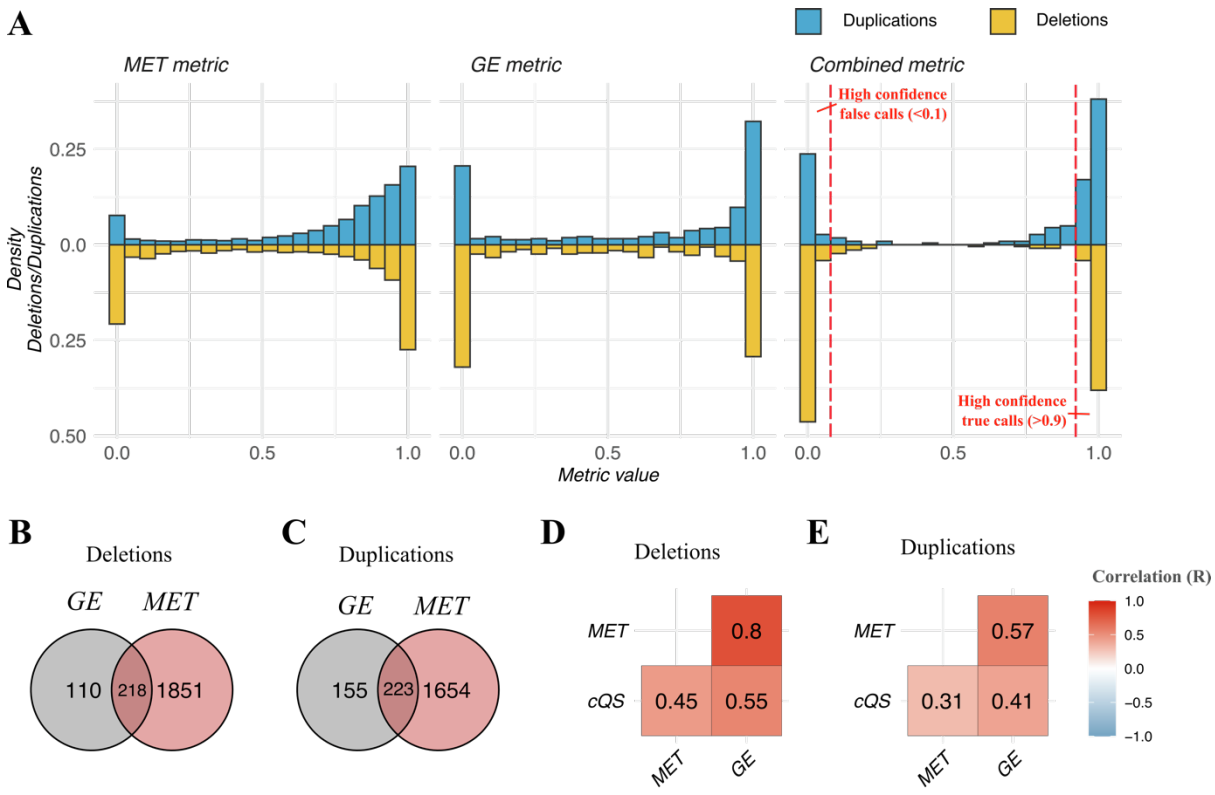

Figure S7. Overview of CNV quality metrics in LLDeep

(A) Distributions of CNV quality metrics based on methylation (*MET*) and gene expression (*GE*) as well as their combination metric in LLDeep cohort for duplications (blue) and deletions (yellow). Altogether, 4,211 pCNVs have either *MET* or *GE* metric calculated. The combined metric is calculated for 441 pCNVs. 50.5% of deletions and 28.3% of duplications are high confidence false calls based on the combined metric (value <0.1), 42.2% of deletions and 57.0% of duplications are high confidence true calls (value >0.9). Number of deletions (B) and duplications (C) in LLDeep dataset that could be evaluated with *MET* and/or *GE* metrics. Pearson correlations between different quality metrics for deletions (D) and duplications (E), including with previously published cQS[16].

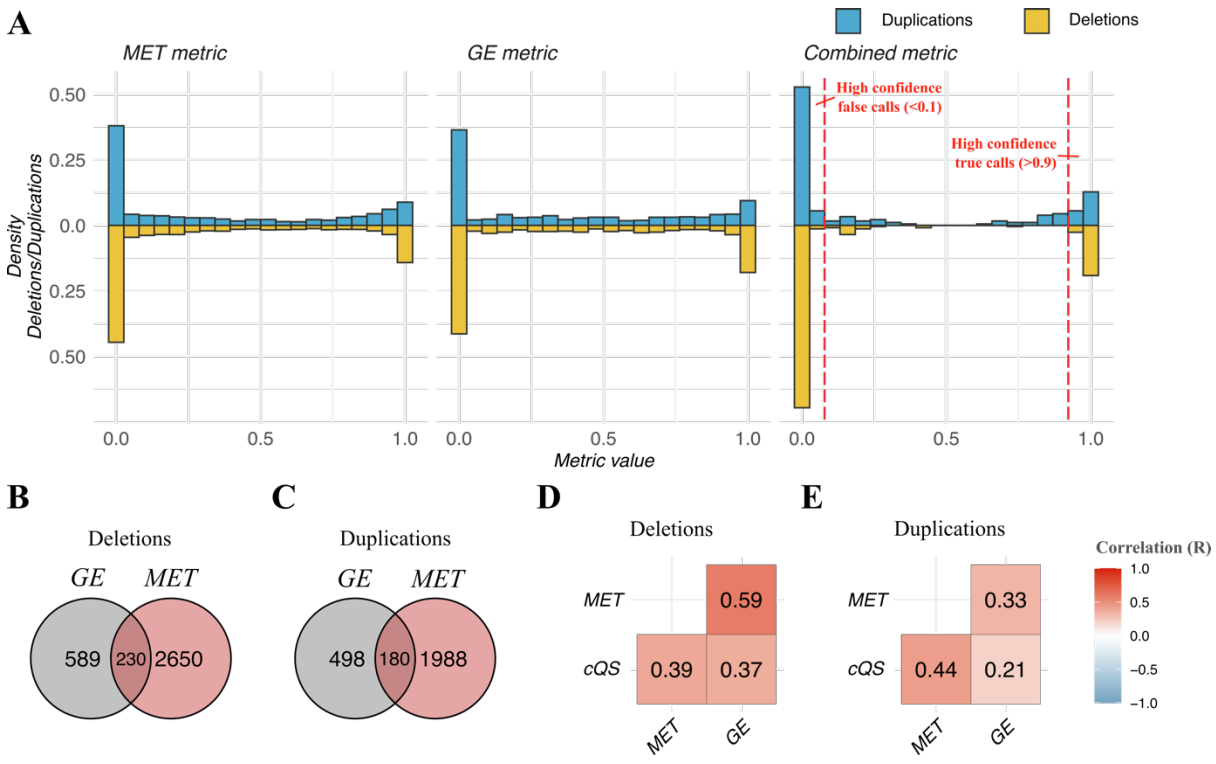

Figure S8. Overview of CNV quality metrics in SkiPOGH

(A) Distributions of CNV quality metrics based on methylation (*MET*) and gene expression (*GE*) as well as their combination metric in SkiPOGH cohort for duplications (blue) and deletions (yellow). Altogether, 6,135 pCNVs have either *MET* or *GE* metric calculated. Combined score was calculated for 410 pCNVs. 70.9% of deletions and 59.4% of duplications are high confidence false calls based on at least one omics metric (value <0.1), only 21.7% of deletions and 20.6% of duplications are high confidence true calls (value >0.9). Since the fraction of false positive calls is much higher and true positive calls much lower in this dataset compared to others, we did not exclude SkiPOGH in CNV quality model building step. Number of deletions (B**)** and duplications (C) in SkiPOGH dataset that could be evaluated with *MET* and/or *GE* metrics. Pearson correlations between different quality metrics for deletions (D) and duplications (E), including with previously published cQS[16].

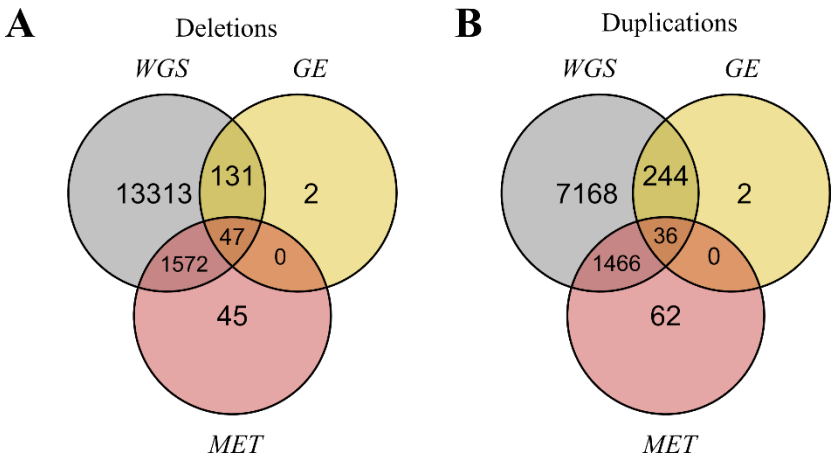

Figure S9. Number of EstBB-MO CNVs per omics layer

Venn diagrams with number of deletions (A) and duplication (B) that can be evaluated using whole-genome sequencing (*WGS*), methylation (*MET*) and/or gene expression (*GE*) metrics in EstBB-MO dataset. The combined metric (*EXTR*) is calculated for pCNVs that have at least two omics metrics available.

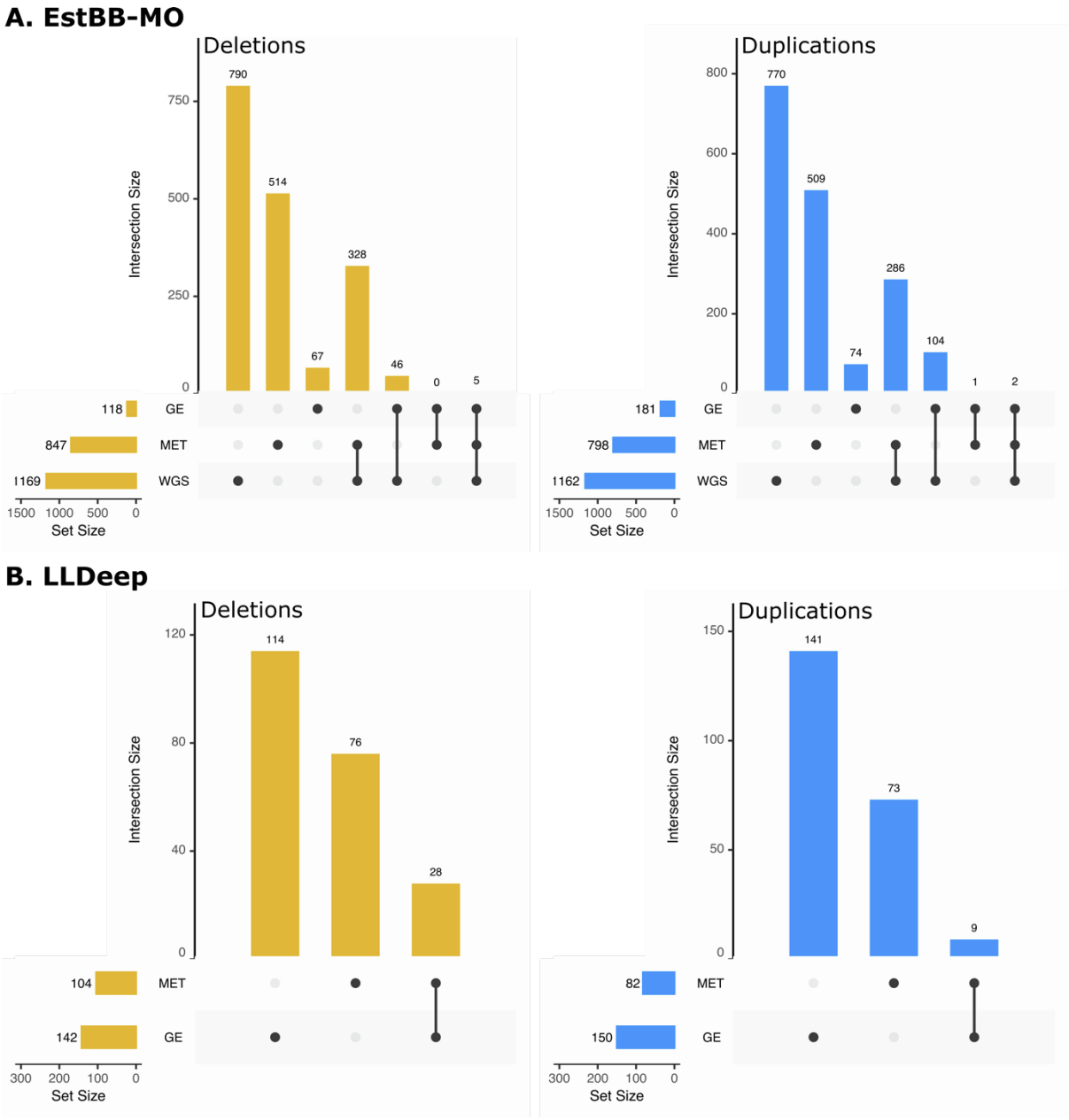

Figure S10. The composition of combined omics-based metric

The composition is shown separately for (A) EstBB-MO and (B) LLDeep datasets and for deletions (yellow) and duplications (blue). The co-occurrence of predictor sources in the figure indicates the cases where two or more omics-based metrics have equal value and either can be chosen for the value of the combined metric.

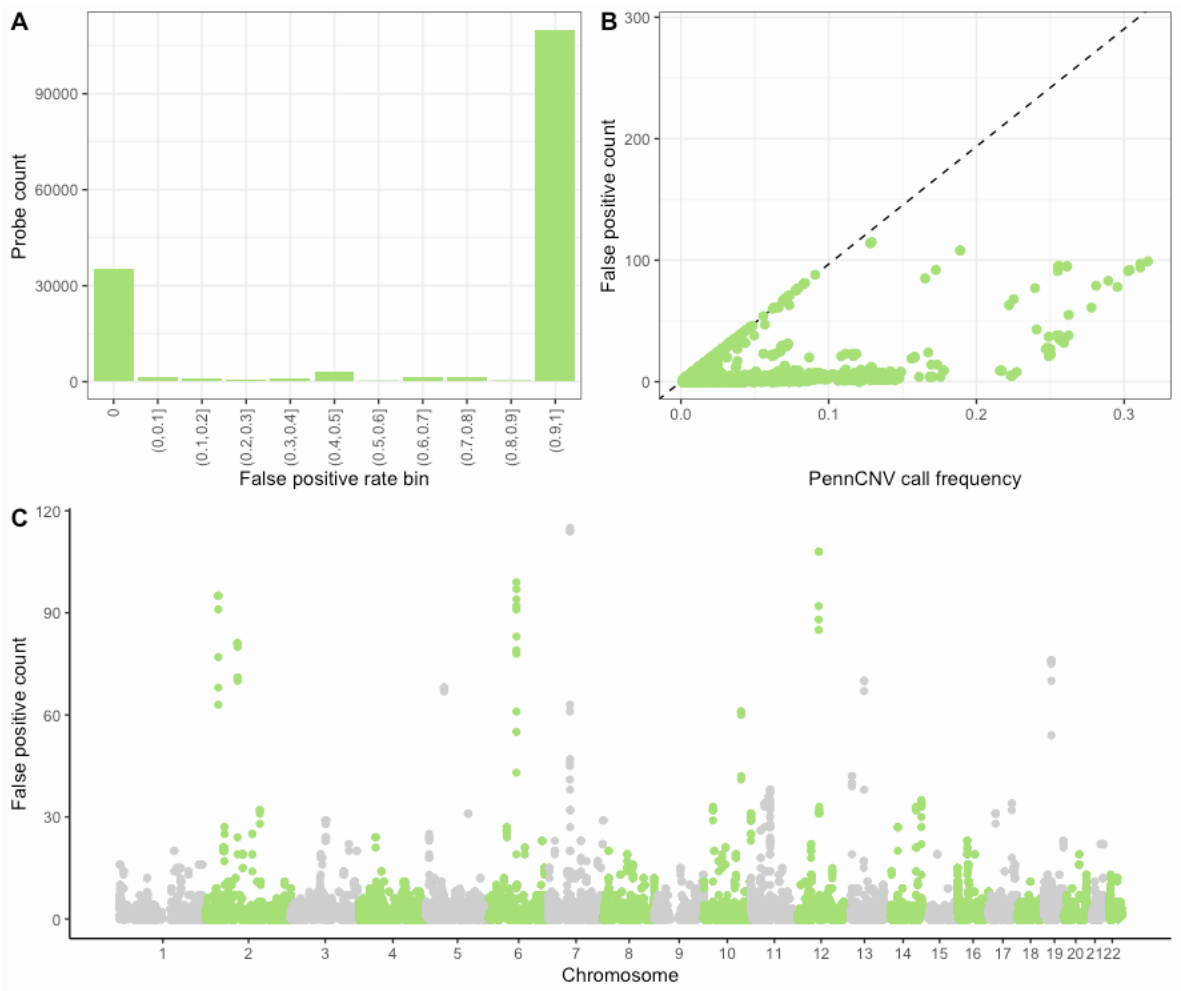

Figure S11. Overview of false positive (FP) CNVs in 979 Estonian samples

FP were defined as PennCNV calls with <10% of its basepairs validated by WGS-CNVs. We calculated FP counts and FP rate (FPR) in a probe-by-probe manner using Illumina OmniExpress array autosomal probe positions (N=709,358). Only probes that overlap with at least one PennCNV call (N=155,448) were included in the figures. We summarised FP CNVs by studying (A) a histogram of FPR bins and (B) FP counts against PennCNV call frequency. Additionally, we illustrated the FP counts across genomic positions.

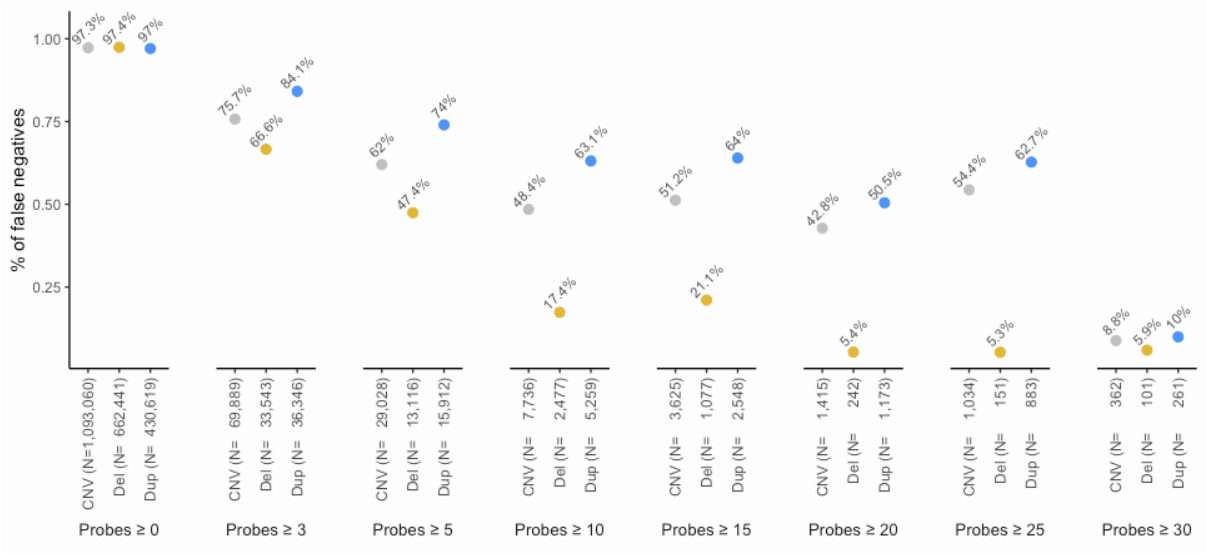

Figure S12. False negative (FN) pCNV across probe bins

Fraction (%) of FN PennCNV calls (y-axis) in function of the number of overlapping genotyping array probes (x-axis), stratified by CNV type (all CNV, deletions only, duplications only). The number of CNVs corresponding to each setting is indicated as N.

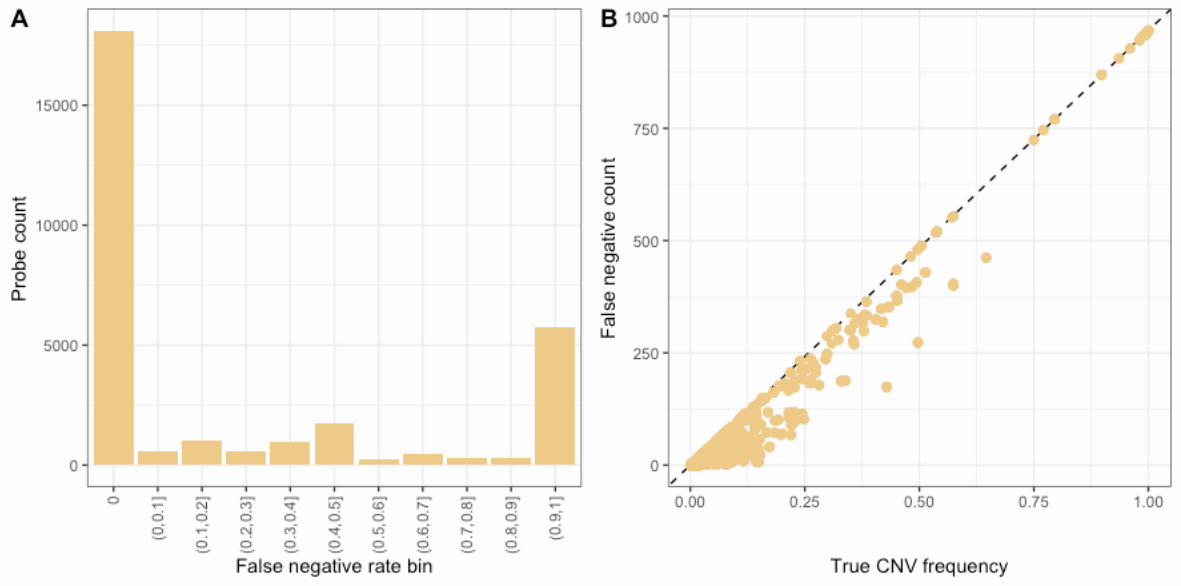

Figure S13. Overview of false negative (FN) CNVs in 979 Estonian samples

In FN calculation, WGS-CNV that overlap with ≥3 probes were defined as true CNVs. FN were defined as a set of true CNVs with <10% basepairs overlapping with a PennCNV call. FN counts and FN rate (FNR) were calculated in a probe-by-probe manner using Illumina OmniExpress array autosomal probe positions (N=709,358). Only probes that overlap with at least one true CNV (N=30,182) were included in the figures. We summarised FN CNVs by studying (A) a histogram of FNR bins and (B) FN counts against true CNV frequency.

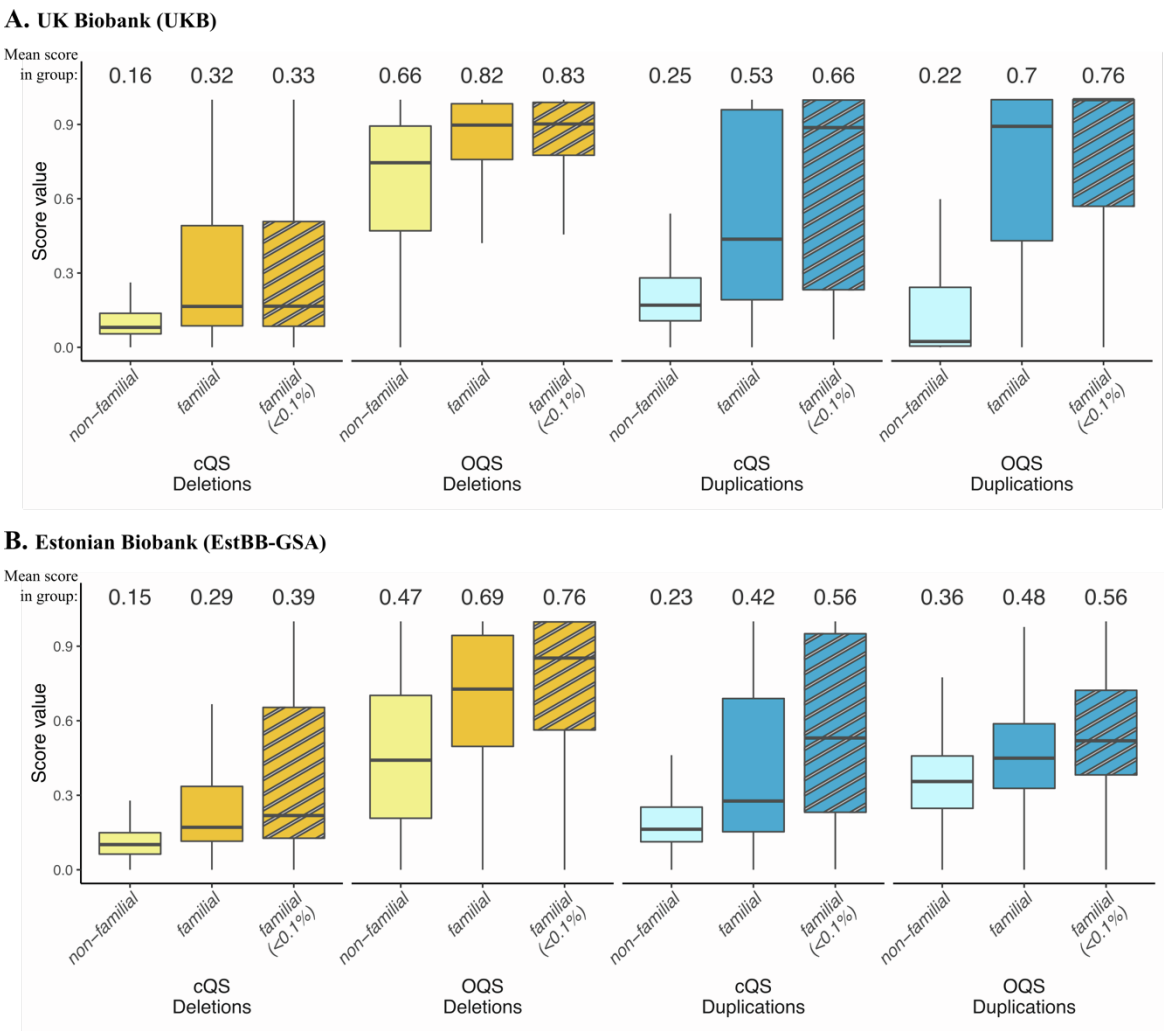

Figure S14. Comparison of quality scores on pCNVs of closely related UKB and EstBB-GSA samples

cQS[16] and omics-informed (OQS) CNV quality score comparisons between *non-familial and familial* pCNVs in (A) the UKB and (B) the EstBB-GSA first-degree relatives. Having a pCNV replicate in a family member is a good proxy for a pCNV to be a true call, while non-familial pCNV sets contain both true and false calls. Additionally, we included rare (frequency <0.1%, striped background) familiar pCNVs as a subset of CNVs unlikely to validate in a relative by chance. The mean score of each pCNV group is shown on top of the figure. For both deletions (yellow) and duplications (blue), OQS shows higher scores for familial pCNVs compared to cQS, and even higher scores for rare familial pCNVs. Outliers are not shown on the figure but are still included in the mean calculations.

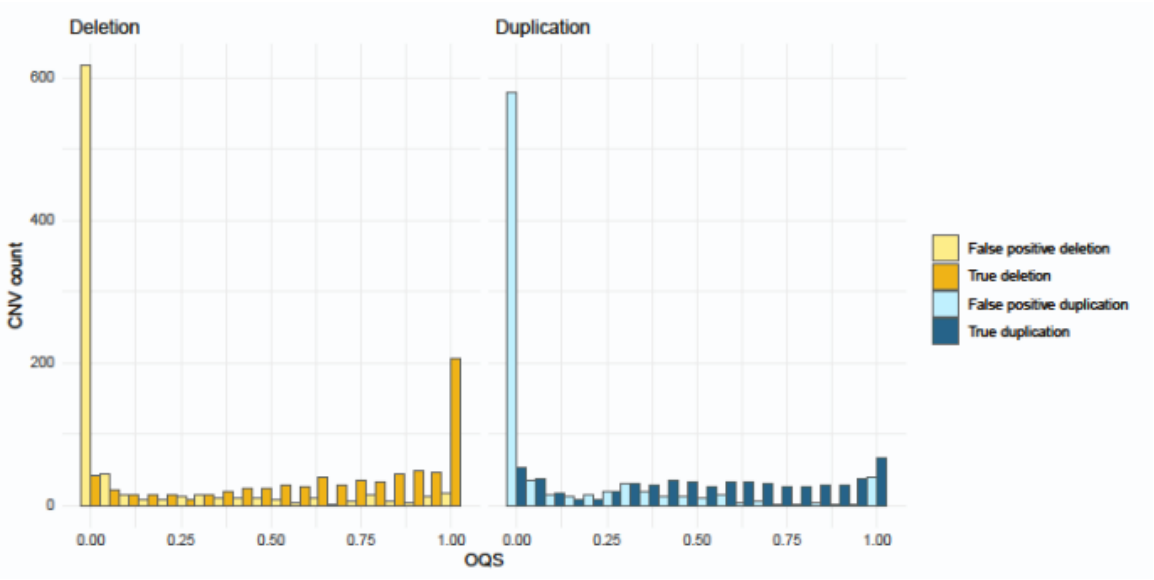

Figure S15. Distribution of predicted OQS in EstBB-MO dataset

The pCNV are divided into false positives (calculated omics-based *EXTR* metric < 0.1) and true CNVs (*EXTR* > 0.9).

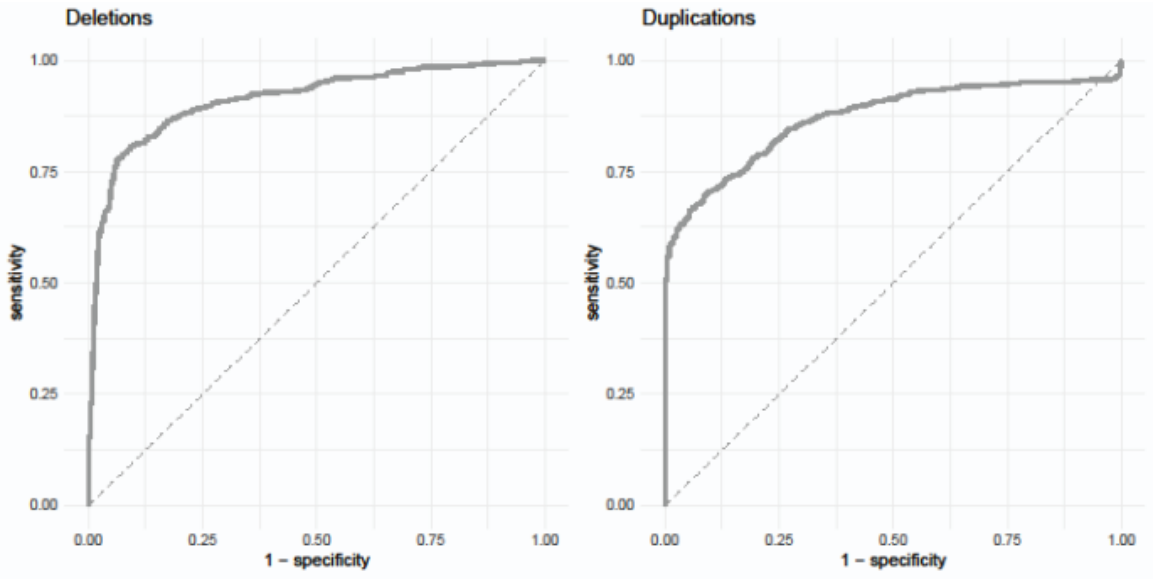

Figure S16. ROC curves of OQS in EstBB-MO dataset

The pCNV are divided into false positives (calculated omics-based *EXTR* metric < 0.1) and true CNV (*EXTR* > 0.9). For deletions AUC=0.91, for duplications AUC=0.87.

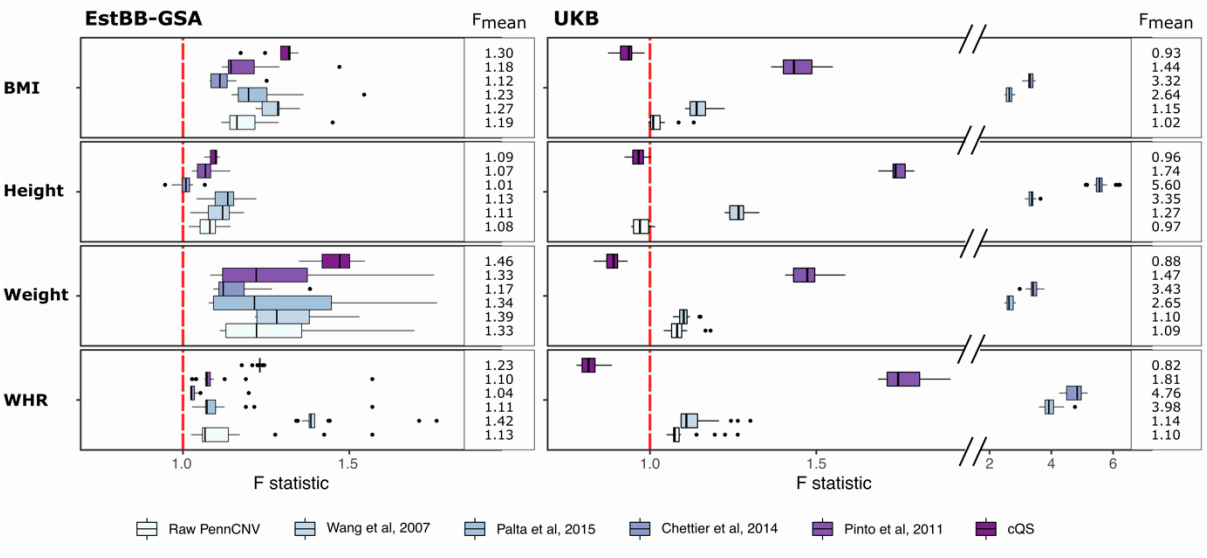

Figure S17. **Impact of OQS on CNV-trait deletion-only associations**

The change of variance explained in deletion-only model when using OQS over raw PennCNV, four published quality filtering approaches[4,19–21] or cQS[16] in both EstBB-GSA and UKB depicted as distribution of F statistics calculated by randomising the probe pruning priority order 20 times (see **Methods**). Explained variance is increased when F>1 and decreased when F<1. Larger F values indicate greater improvement in statistical power when using OQS over the given reference approach.