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

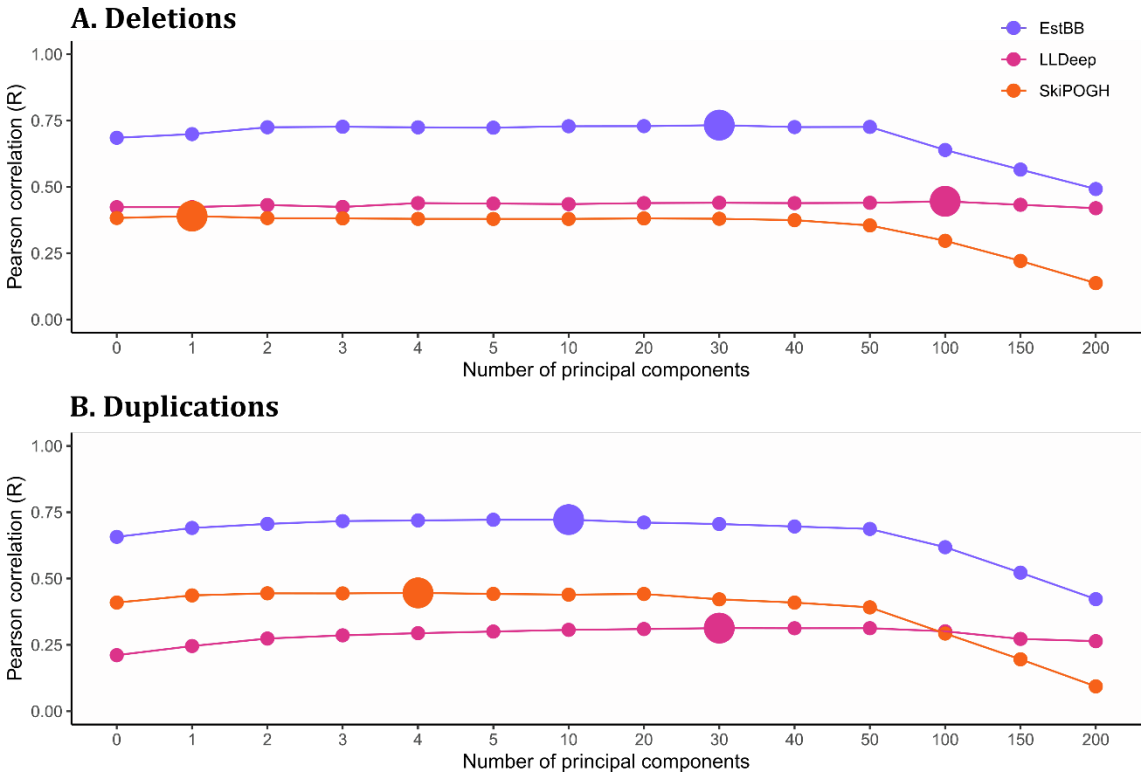

Figure S1. Methylation metric correlations per number of principal components

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.

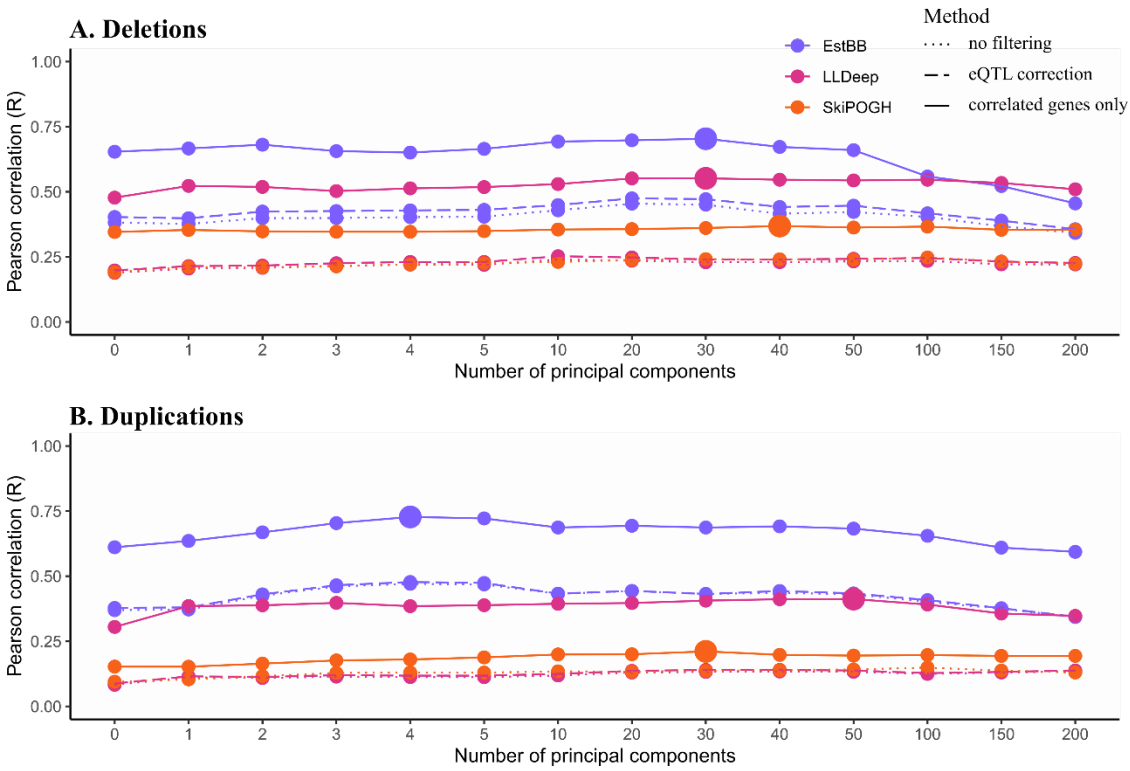

Figure S2. Gene expression metric correlations per number of principal components

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.

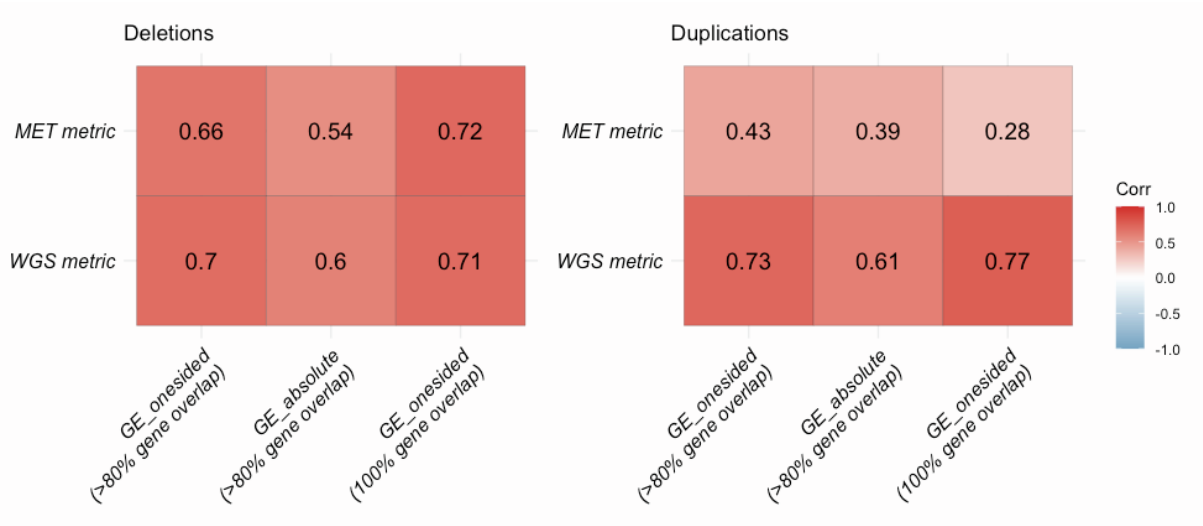

Figure S3. Pearson correlations between WGS/MET metrics and GE metrics calculated using three different approaches

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.

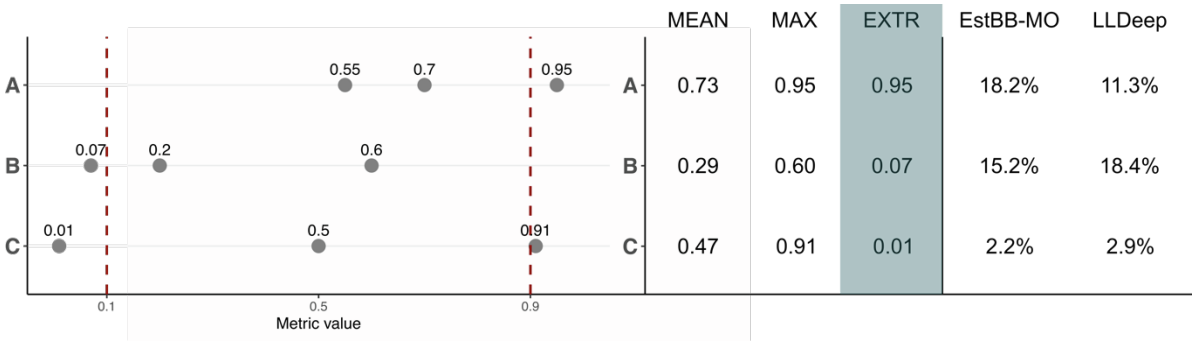

Figure S4. Comparison of omics-based combined metrics

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.

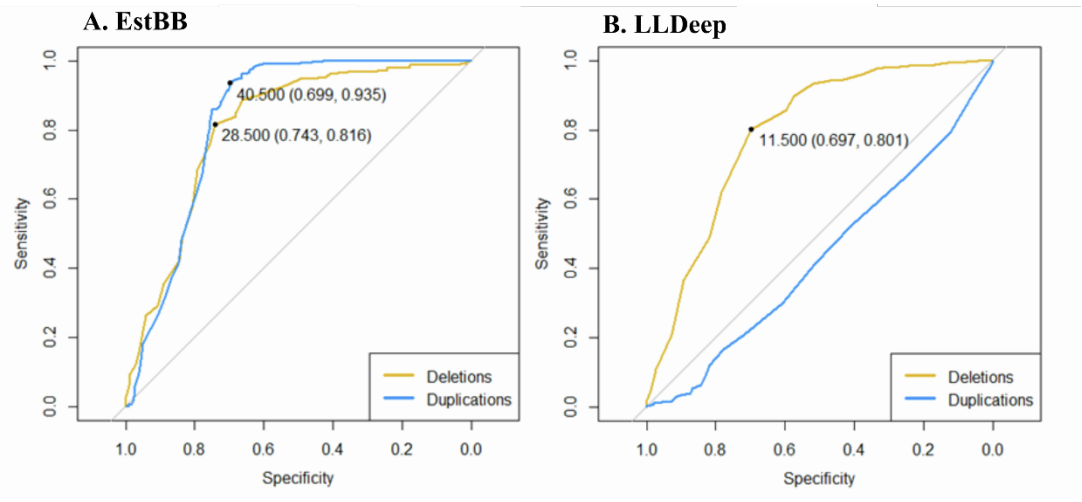

Figure S5. Number of pCNVs per sample as pCNV quality predictor

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.

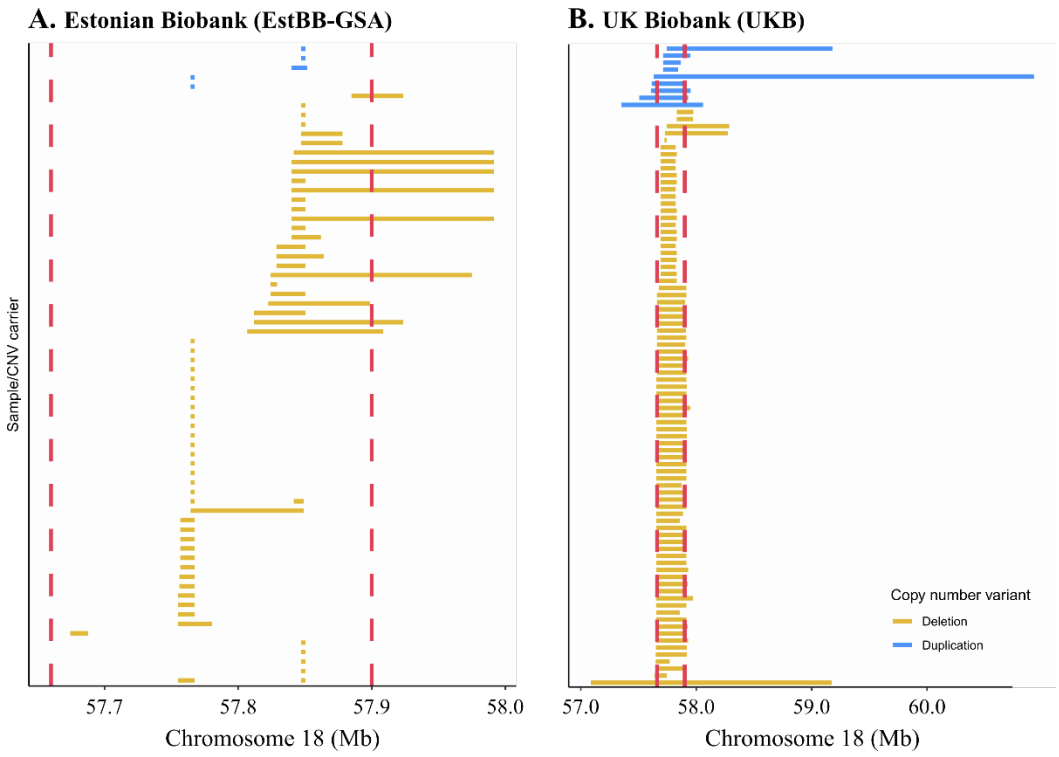

Figure S6. 18q21.32 CNV region in EstBB-GSA and UKB datasets

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.

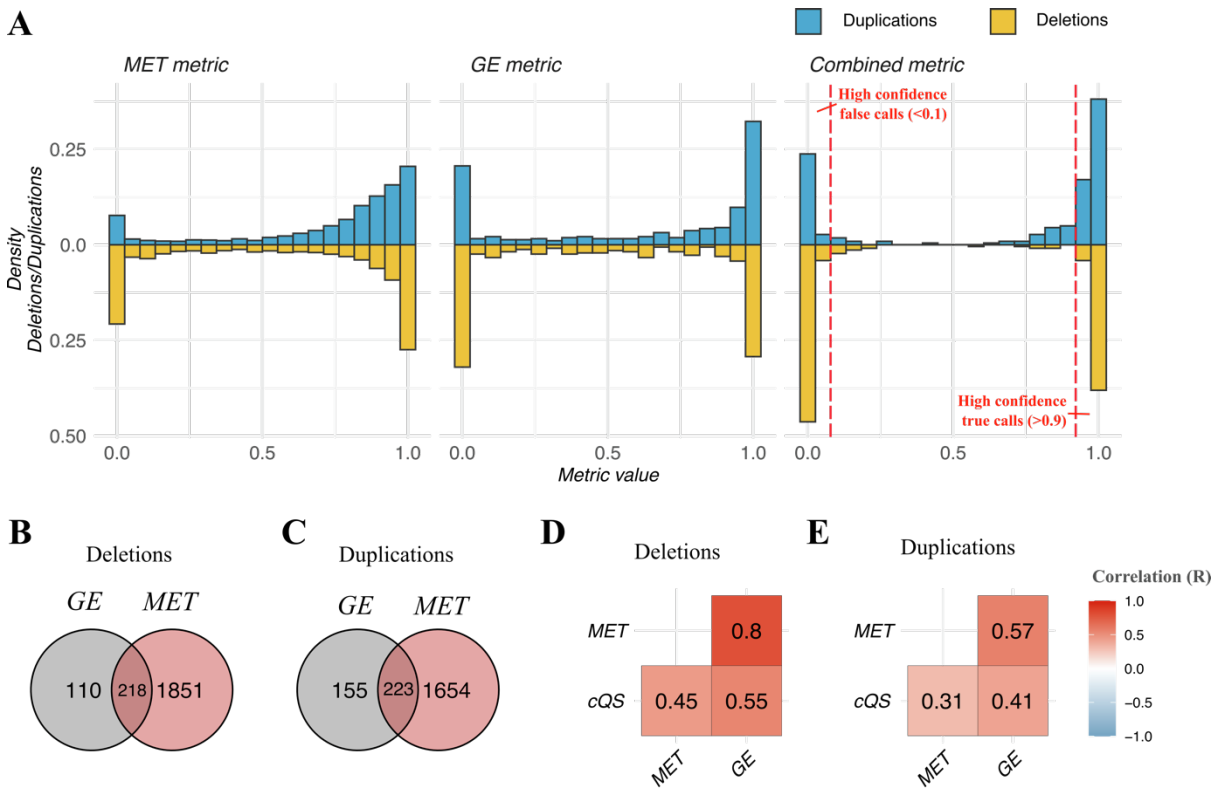

Figure S7. Overview of CNV quality metrics in LLDeep

(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].

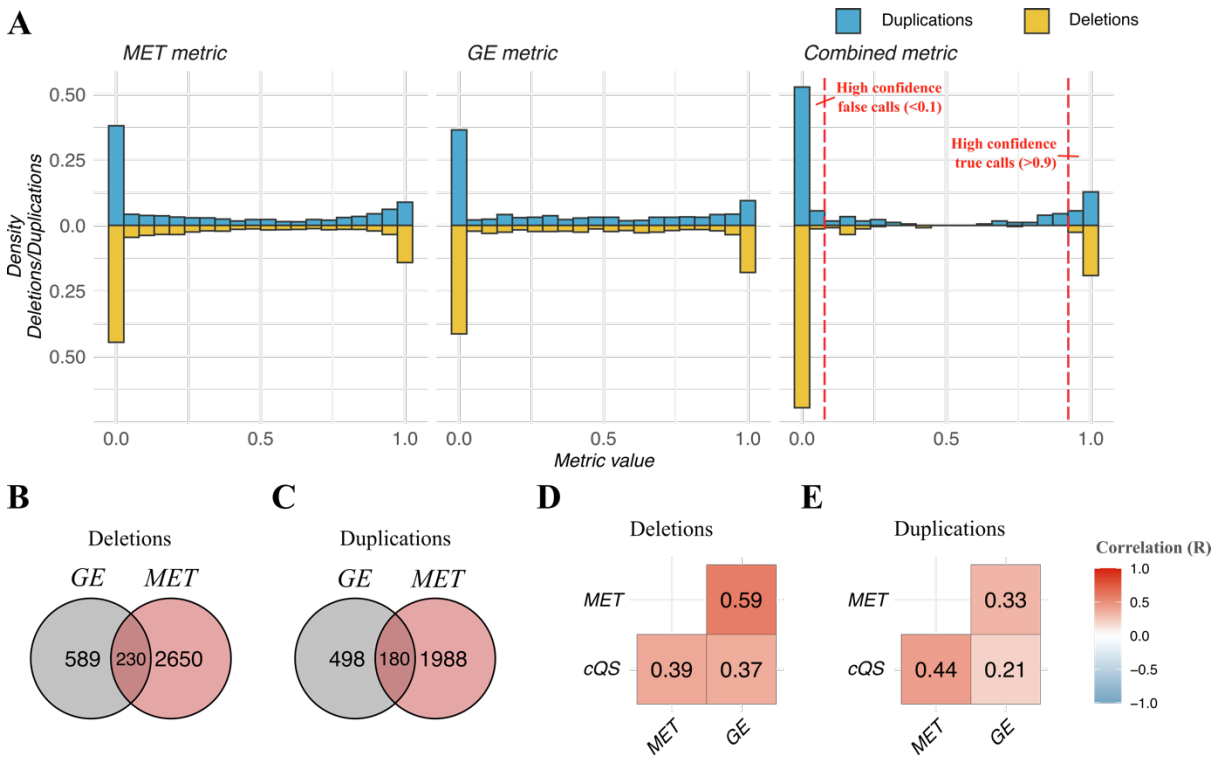

Figure S8. Overview of CNV quality metrics in SkiPOGH

(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].

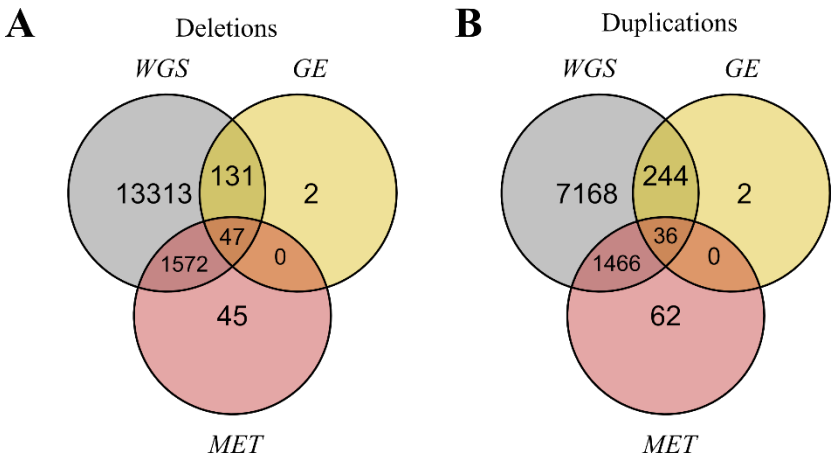

Figure S9. Number of EstBB-MO CNVs per omics layer

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.

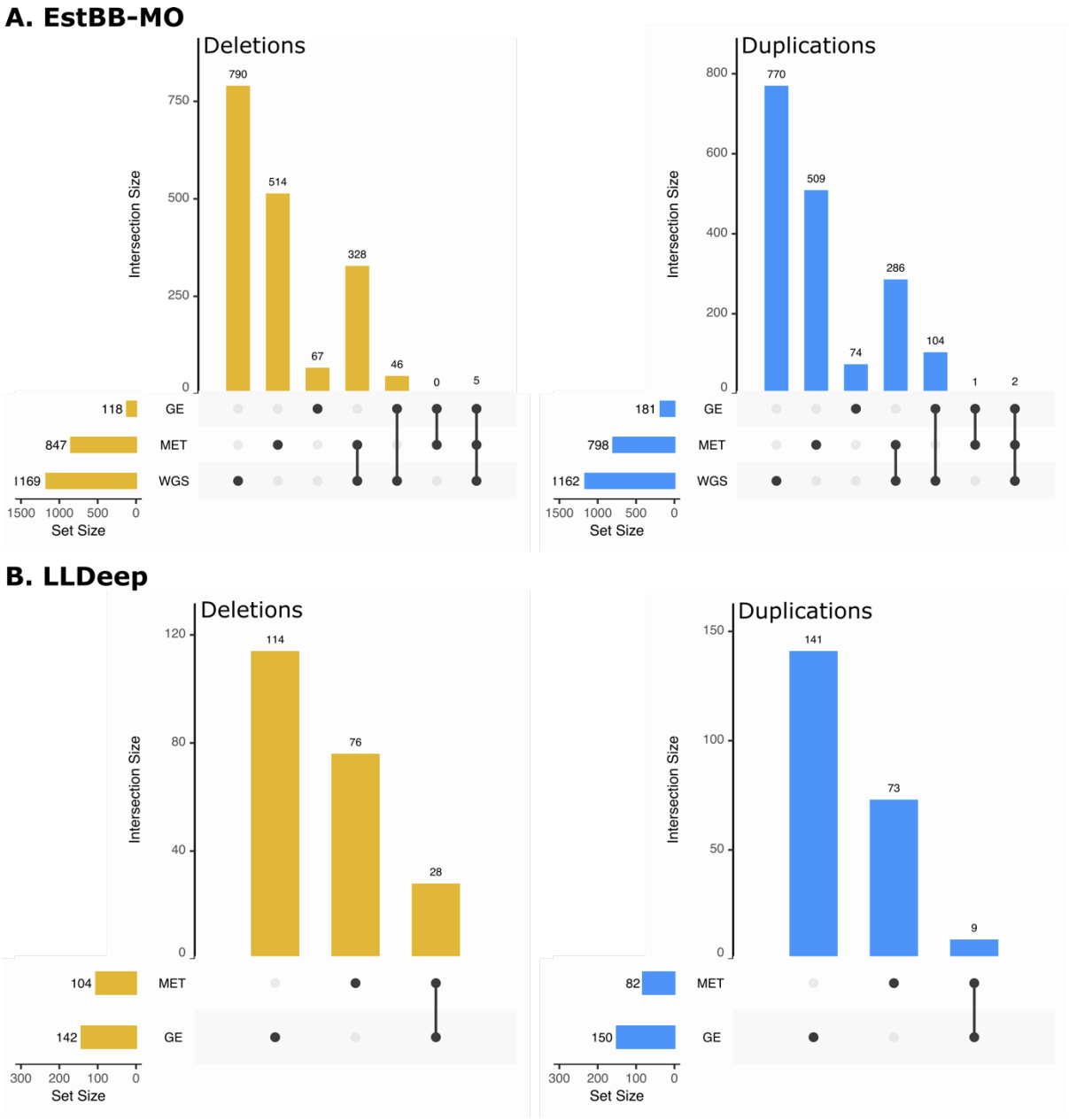

Figure S10. The composition of combined omics-based metric

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.

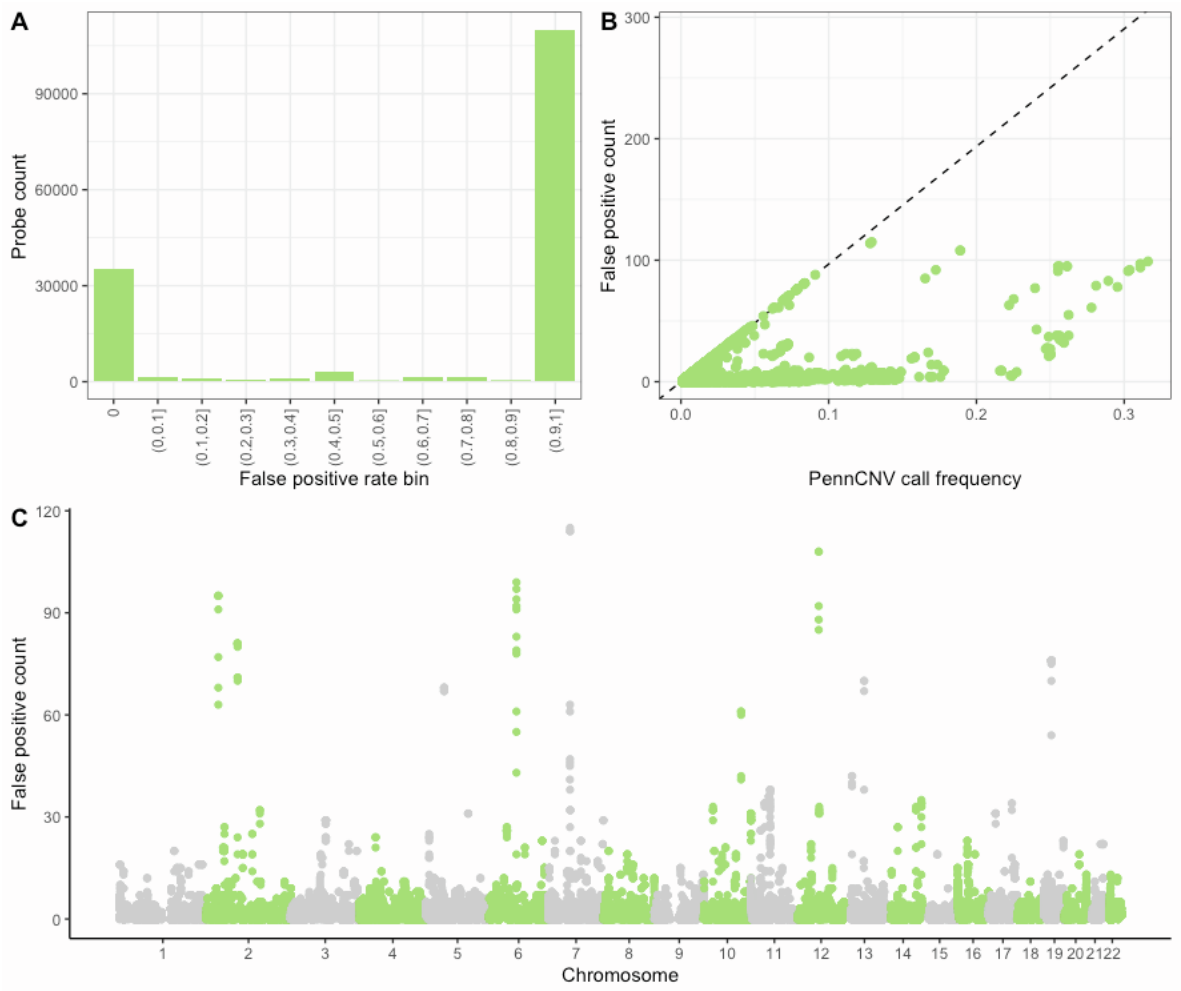

Figure S11. Overview of false positive (FP) CNVs in 979 Estonian samples

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.

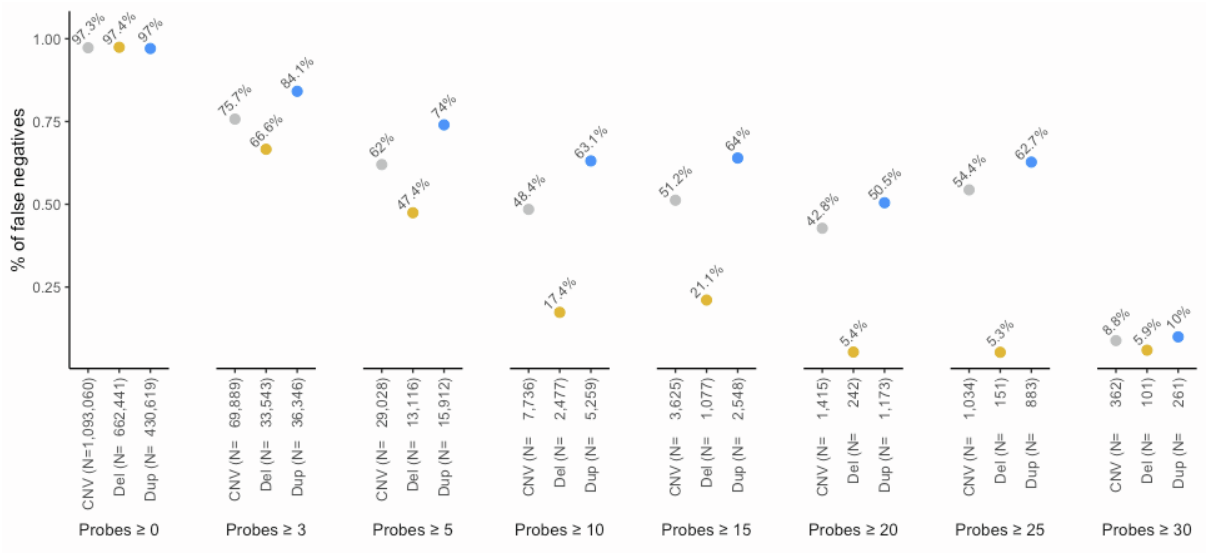

Figure S12. False negative (FN) pCNV across probe bins

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.

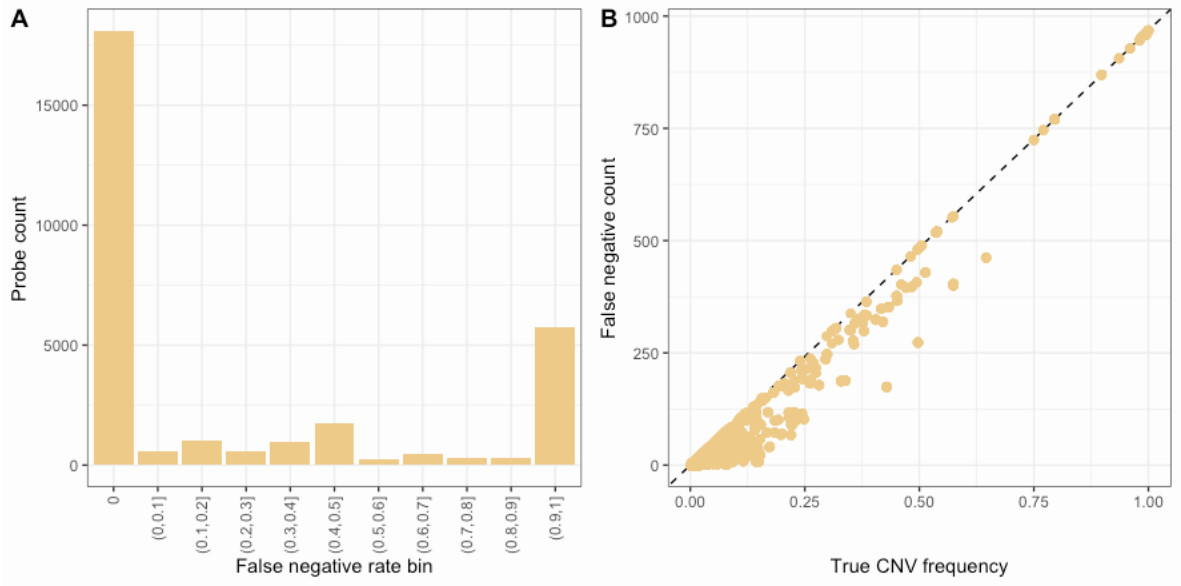

Figure S13. Overview of false negative (FN) CNVs in 979 Estonian samples

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.

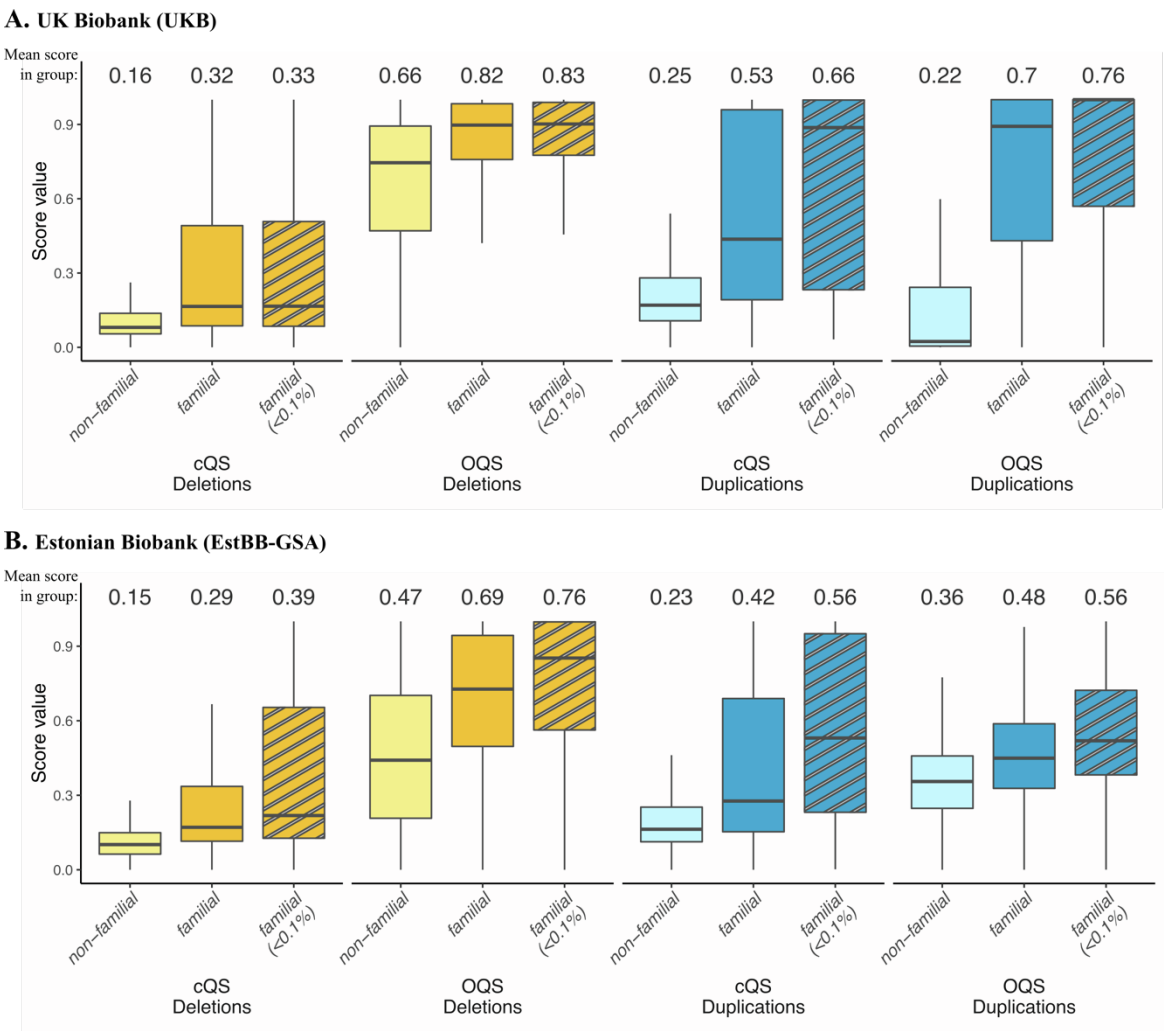

Figure S14. Comparison of quality scores on pCNVs of closely related UKB and EstBB-GSA samples

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.

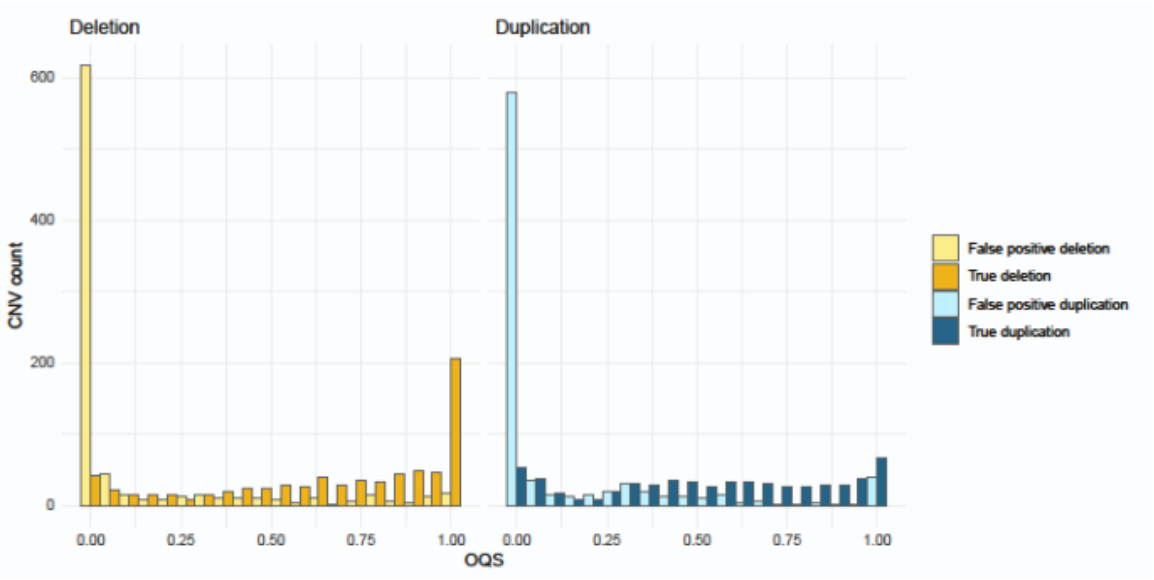

Figure S15. Distribution of predicted OQS in EstBB-MO dataset

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

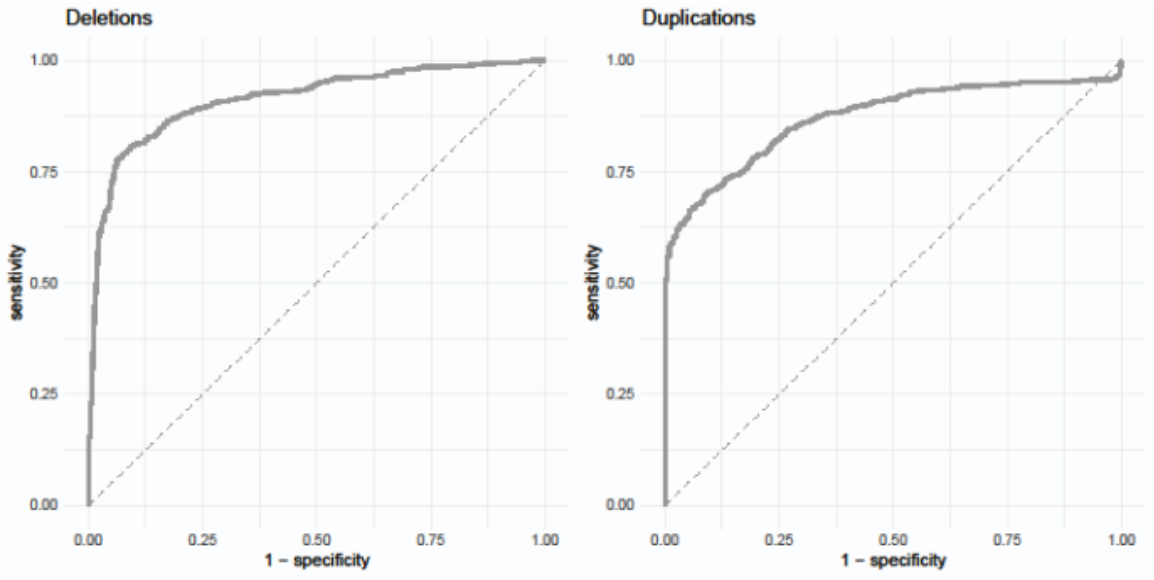

Figure S16. ROC curves of OQS in EstBB-MO dataset

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.

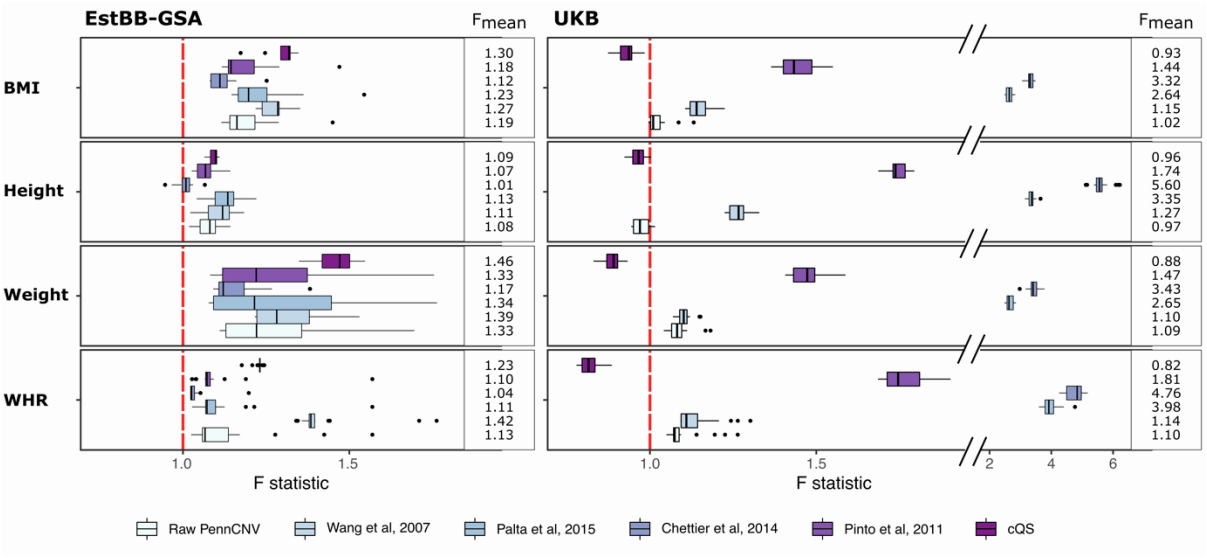

Figure S17. **Impact of OQS on CNV-trait deletion-only associations**

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.