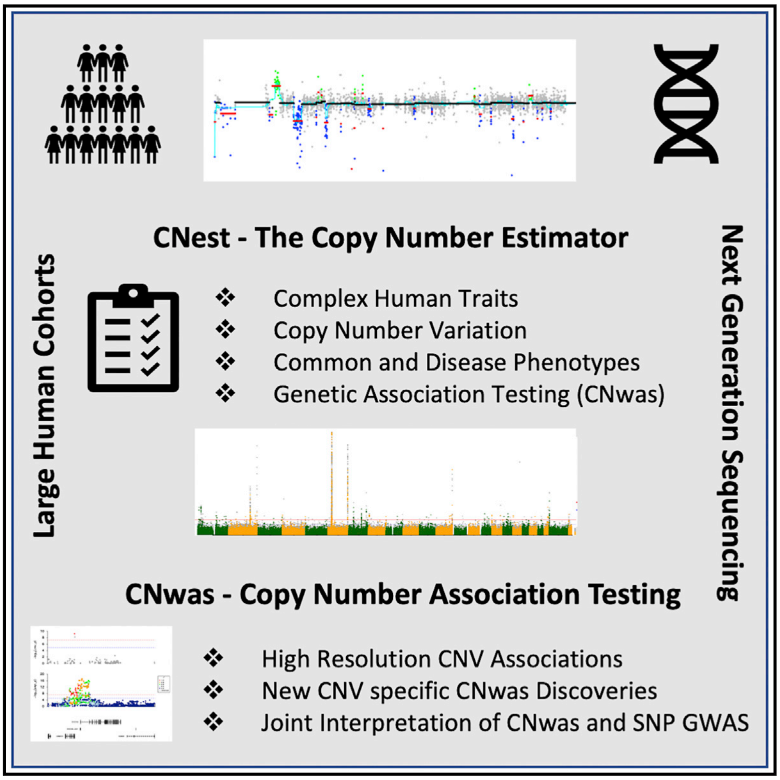
Cell Genomics

Article

CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank

Graphical abstract



Large Human Cohorts

Next Generation Sequencing

CNest - The Copy Number Estimator

- ❖ Complex Human Traits
- ❖ Copy Number Variation
- ❖ Common and Disease Phenotypes
- ❖ Genetic Association Testing (CNwas)

CNwas - Copy Number Association Testing

- ❖ High Resolution CNV Associations
- ❖ New CNV specific CNwas Discoveries
- ❖ Joint Interpretation of CNwas and SNP GWAS

Highlights

- Novel copy number variation (CNV) associations in a large human cohort
- Large-scale CNV association testing from next-generation sequencing data
- Specifically designed GWAS framework for CNV operating under GA4GH standards
- Investigation of SNP and CNV associations and CNV tagging SNPs

Authors

Tomas Fitzgerald, Ewan Birney

Correspondence

tomas@ebi.ac.uk (T.F.),
birney@ebi.ac.uk (E.B.)

In brief

CNest copy number variation (CNV) genome-wide association study (GWAS) testing brings CNVs into a similar framework as those widely used in SNP GWAS. We made over 800 new discoveries in the UK Biobank human cohort at exon-level resolution, paving the way for further CNV GWAS studies using next-generation sequencing data.

Fitzgerald & Birney, 2022, Cell Genomics *2*, 100167
August 10, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.xgen.2022.100167

CellPress

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

Cell Genomics

CellPress
OPEN ACCESS

Article

CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank

Tomas Fitzgerald[1,2,*] and Ewan Birney[1,*]
[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK
[2]Lead contact
*Correspondence: tomas@ebi.ac.uk (T.F.), birney@ebi.ac.uk (E.B.)
https://doi.org/10.1016/j.xgen.2022.100167

SUMMARY

Copy number variation (CNV) is known to influence human traits, having a rich history of research into common and rare genetic disease, and although CNV is accepted as an important class of genomic variation, progress on copy-number-based genome-wide association studies (GWASs) from next-generation sequencing (NGS) data has been limited. Here we present a novel method for large-scale copy number analysis from NGS data generating robust copy number estimates and allowing copy number GWASs (CN-GWASs) to be performed genome-wide in discovery mode. We provide a detailed analysis in the UK Biobank resource and a specifically designed software package. We use these methods to perform CN-GWAS analysis across 78 human traits, discovering over 800 genetic associations that are likely to contribute strongly to trait distributions. Finally, we compare CNV and SNP association signals across the same traits and samples, defining specific CNV association classes.

INTRODUCTION

Genome-wide association studies (GWASs) are a well-established genetic technique, having made thousands of robust associations between traits and sequence-level genetic variation.[1–7] Often these associations can have significant impacts on the understanding and, in some cases, the treatment of human disease.[8–10] However, for most common genetic diseases, these associations only account for part of the heritable disease risk.[11–13] In terms of total base pairs, copy number variation (CNV) accounts for the majority of differences between any two genomes[14–18] and is known to alter human trait distributions,[19–21] often with a strong impact on human health.[22,23] This is highlighted best within the large body of research studying CNV in relation to rare genetic disease.[24–26] Although it is widely accepted that CNV can contribute significantly to differences in human traits, to date, methods for large-scale CNV to phenotype association studies, the equivalent of GWAS for CNVs, have been hampered by a number of factors, including methodological difficulties,[27] the availability of sufficiently large datasets, and the ability to interpret complex rearrangements from sequencing data.[28,29]

CNVs have been a major component of routine clinical medical genetics screening for over a decade; however, the interpretation of individual events remains challenging,[30,31] with most clinical testing laboratories routinely finding potentially pathogenic CNVs in patients with intellectual disability, autism spectrum disorders, and/or multiple congenital anomalies.[32–34] Although CNV detection from sequence data in a clinical setting is in active development, most clinical CNVs are still discovered using specialized microarrays.[35] Most CNVs with strong effects are rare and they are often discovered as *de novo* mutations in patients across a range of genomic disorders.[36,37] Furthermore, it has been observed that the overall burden of CNVs is higher in specific patient groups compared with controls[38,39] indicating potential combinatorial CNV effects.[40] It is conceivable that specific combinations of CNVs, by acting in concert, may have a large potential to cause phenotypic differences due to factors such as dosage compensation, incomplete penetrance, and polygenic effects.[41,42] The impact of CNVs in rare diseases is likely to be large, whereas one might expect weaker effects of all variation in more common diseases, consistent with the polygenic behavior of these diseases.

Several CNV genotype-phenotype correlations have been observed in relatively small-scale studies of specific patient groups[43] or by collaborative efforts to share genetic data for rare disease;[44] however, CNVs have also been associated with a number of complex diseases.[45,46] Recent large-scale CNV association testing using datasets such as the UK Biobank have found some highly significant loci in relation to certain human traits,[47] and previous studies focused on cognitive traits such as schizophrenia[48] and autism[49] have demonstrated the utility of SNP arrays to search for novel CNV associations. Focused studies into specific human traits have used large-scale SNP genotypes to perform association testing with great success;[50–54] however, these studies have often focused on predefined lists of CNV regions known to be important within a clinical setting.[53]

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Another important consideration is that SNP genotyping arrays have a limited resolution to detect small CNVs and a limited sensitivity for CNV discovery genome-wide due to the distribution of SNPs across the genome and a limited dose response.[55] A recent CNV association study showed both the power and limitations of genotype-based CNV association testing in the UK Biobank, finding 131 significant signals across 47 quantitative traits.[56] The smallest CNV association detected was 49 kb at 1p36.11 found to be in association with reticulocyte count, platelet count, and hemoglobin A1c (HbA1c).[56] However, most of the CNV association signals detected involved large recurrent CNVs with a mean size of 817 kb, highlighting the limited resolution when using SNP arrays. Another recent study showed an improvement in resolution from SNP arrays by including information on shared extended SNP haplotypes into their model, detecting 269 independent associations across 56 quantitative traits.[57] Both studies tested quantitative traits only and were limited to the resolution of the SNP array; nevertheless, both found novel discoveries, highlighting a large potential for CNV association testing genome-wide for complex human traits.

It is reasonable to assume that CNVs may account for a substantial portion of the variance observed in common disease risk. Some of these CNVs will be in strong linkage disequilibrium (LD) with SNPs, and so they can be discovered by tagging polymorphisms, but the causal change is impossible to narrow down using SNPs alone. Other CNVs might not have good tagging SNPs and, furthermore, recurrent CNVs are far more common than recurrent SNPs, with the CNV mutation rate currently estimated at 0.2 *de novo* events per individual compared with between 1.8×10^{-8} and 2.5×10^{-8} per base pair per generation for point mutations,[19,58–60] meaning that the aggregate higher-frequency CNVs with the same functional impact are hard to model using the combination of rare haplotypes. With the advance of large data cohorts with datasets that are amenable to copy number estimation,[61–63] the ability to perform high-resolution genome-wide GWAS testing for CNVs has become more feasible. One challenge for large-scale CNV discovery has been variability in raw sequencing depth due to other factors, most likely extraction techniques and immune system state at the time of blood draw. This variation gives rise to complex noise characteristics in raw sequencing read depth between samples, so called genomic waves. To explore this, one needs robust normalization strategies for CNVs, an appropriate discovery method for CNVs, and a way to easily integrate both CNV- and SNP-based associations into one framework.

In this work, we address some of these issues by providing a new discovery method for CNVs from next-generation sequencing (NGS) data, CNest, based on novel normalization techniques for large-scale cohorts. Rather than trying to create individual models of alleles for each CNV locus, we have chosen to use a straightforward linear model for discovery. This linear model is both consistent across all CNV loci and has many similar properties to the linear models used in SNP GWAS. As such, we can use the same covariates, the same diagnostic style QQ plots, and place SNPs and CNVs associations into the same framework. Post discovery, we show we can provide more detailed modeling of at least some loci. We provide a comprehensive CNV analysis using this method on the large UK Biobank cohort with exome sequences. To explore the relationship with established SNP polymorphisms, we also performed both CNV and SNP GWAS within a single framework, applying our methods across the same set of UK Biobank samples and interrogating the resulting associations across a diverse set of traits.

We find many CNV to phenotype associations, although, as expected, many of these associations are also tagged by SNP polymorphisms. However, we have a subset of CNV associations that cannot be discovered via SNPs and another subset that are coincident with strong SNP polymorphisms but not well correlated with any specific SNP, and many where the CNV is taggable but the tagged SNP is at some distance from the CNV locus. Many of these associations recapitulate multiple known associations based on previous studies on both CNV and SNP genome association testing, whereas others discover new CNV-specific findings in relation to the genetics of common human traits. We have made the software, CNest, which performs this discovery open source and provided portable workflows to run CNest, compatible with GA4GH standards.[64]

RESULTS

CNV in 200,629 individuals from the UK Biobank

To identify exon-resolution CNV regions across a large population of individuals from NGS data in the UK Biobank, we developed a suite of flexible, highly scalable CNV analysis tools known as CNest (see STAR Methods). Within this package, we include a robust CNV caller as well as a set of tools and novel approaches to CNV association testing genome-wide in discovery mode. A flow diagram describing the major steps performed for CNest calling and association testing can be found in Figure S1. A central component of these methods is the selection of appropriate reference datasets and normalization procedures by modeling certain noise characteristics of whole-exome sequencing (WES)/whole-genome sequencing (WGS), for example, the presence and scale of genomic waves, to generate optimized copy number measurements across large sample cohorts (see STAR Methods). After calling CNVs in the 200,629 sample cohort with WES data, we applied several quality control (QC) measures to ensure that the copy number measurements and CNV calls were consistent.

A subset of the diagnostic plots of CNest is shown in Figure 1. An obvious but important step in CNV analysis is the classification of sex based on the estimated copy number of the X chromosome. Well-controlled one versus two copy number of the X chromosome indicates that the normalization procedure and relative copy number estimates have worked successfully, at least for the X chromosome (Figure 1A). A side effect of this analysis is the ability to detect sex chromosome aneuploidy. We detected 50 samples showing an unusually high number of copies on chromosome X (Figure 1A). These samples were assumed to be a mixture of data quality issues and real triple X cases. Triple X is a condition caused by random error during reproductive cell division and is found in approximately one in 1,000 women. Although triple X has been associated with several trait differences, it can often go undiagnosed and, depending on other social factors, may never give rise to any noticeable problems.[65] We also detected 51 datasets that show an unusual level of

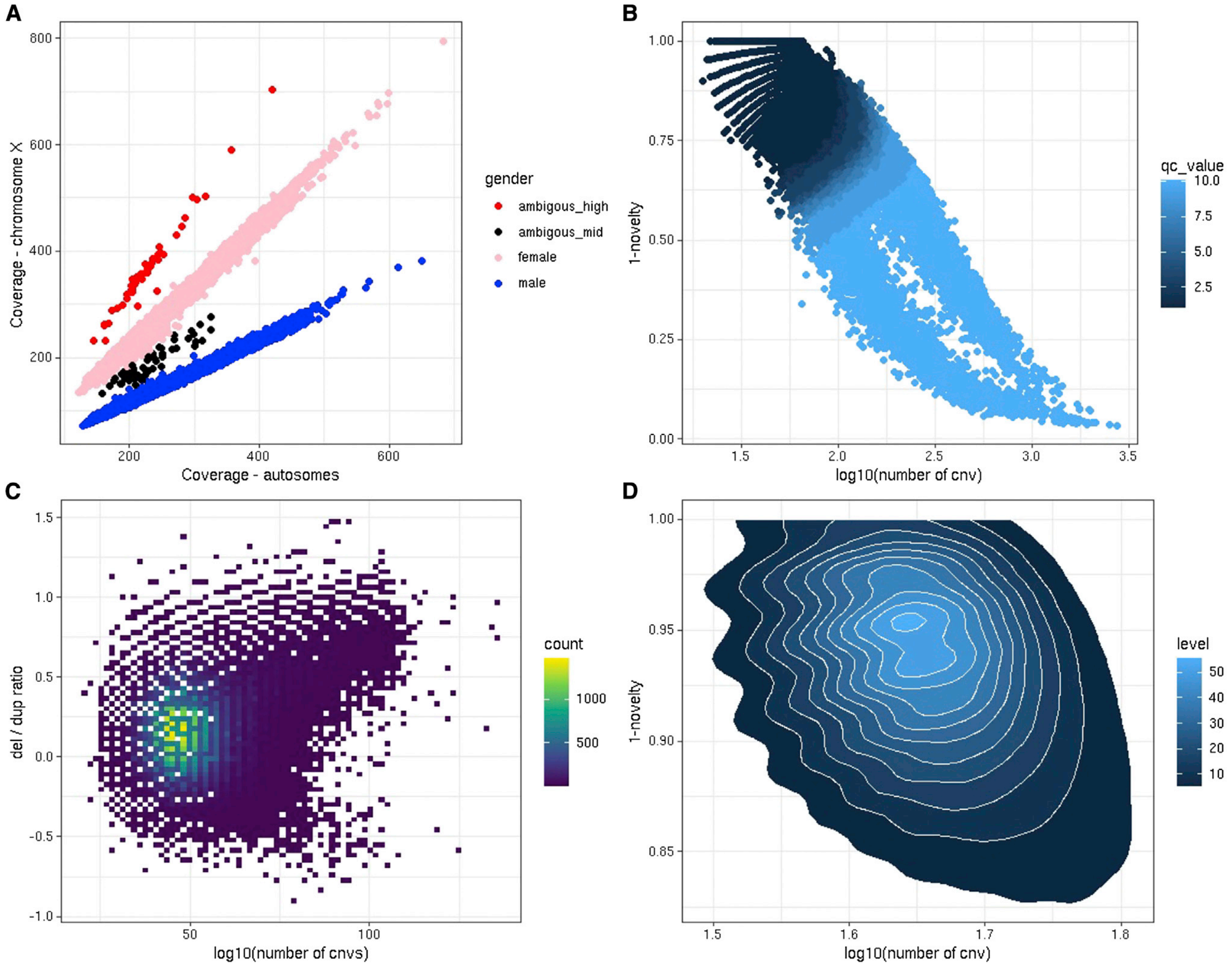


Figure 1. QC of CNV calls in the 200,629 UK Biobank exome sequences

(A) Gender classification, the relative coverage of autosomes compared with chromosome X and the CNest gender classifications shown in different colors across all samples.

(B) The total number of autosomal CNV calls versus a measure of the proportion of rare CNVs per sample using a 1% population frequency.

(C) The log10 of the loss to gain ratio versus log10 of the total number of CNV calls for each sample.

(D) A density plot showing (B) but for QC-passed samples only.

chromosome X coverage and cannot be reliably assigned to either double (female) or single X (male). These we assume to be both inconsistent capture of chromosomal baits and potential mosaic sex chromosome events (i.e., mosaic XXY).

Like previous studies into sex chromosome aneuploidy in the UK Biobank,[66] we classify 50 out of 110,312 women as potential trisomy X (Ambigous_high) giving a prevalence of 45.3 out of 100,000. We also compared the sex classification made by the SNP arrays (f22001), the prediction of sex chromosome aneuploidy (f22019), and the presence of any International Classification of Diseases, Tenth Revision (ICD10) code for sex chromosome abnormality, Q90 to Q99 (Table S1). The majority of sex classifications agree between the exome sequence and SNP array data, and those that are discordant are enriched for the presence of both f22019 and sex chromosome abnormality-related ICD10 codes. Sex chromosome aneuploidy is not a focus of this study, and we simply exclude all samples that could not be reliably assigned to either double or single X based on their coverage profiles (Figure 1A); these copy number sex chromosome calls will be returned to the UK Biobank for further investigation by other investigators.

Some informative CNV quality information is contained within the consistency in the number of CNV calls in all samples against the proportion of those calls that are rare across the entire population (Figure 1B). This is like the genotyping extreme heterozygosity quality parameter used as standard in SNP genotyping QC. Given current estimates on the CNV mutation rate,[67] we would expect very low numbers of *de novo* CNV events (less than one per genome) and rare CNVs to be infrequent in any individual genome, which is supported empirically here with a median of three rare CNVs per UK Biobank exome based on a 1% population frequency for losses and gains separately. For large-scale CNV analysis in assumed healthy individuals, it is sensible to assume that most genomes will on average display a consistent level of rare variation compared with the bulk of the population. Encouragingly, after applying our strictest definition of QC across greater than 200,000 exome sequences for CNV calling, we obtain a greater than 92% pass rate, indicating that, for most samples, our CNV estimation and calling approach is consistent. There is no reason to expect, given known CNV formation mechanisms such as non-allelic homologous recombination (NAHR) and non-homologous end-joining (NHEJ), that there would be any bias between the number of losses and gains when comparing large numbers of genomes in aggregate, and although there are some outliers, we observe a tight loss-to-gain ratio distribution with a median of 1.4 (Figure 1C). When assessing these distributions in samples that passed our QC criteria, the bulk of the data are tightly centered around a mean number of calls of 48 and a mean rarity rate of 0.07 (Figure 1D).

As expected, we observe a bias in loss/gain detection with a median of 28 losses compared with 19 gains per sample (Figure S2A); most CNV callers from both array- and sequence-based data show an increased ability to detect losses compared with gains due to the increased variance for higher copy number signals ("reads") and a lowered dose response.[68–70] This decreased dose response makes it more challenging to detect gains, often requiring a larger number of consistent signals ("responding probes") to be able to distinguish real signals from baseline noise properties.[71,72] When looking at the proportion of CNV calls (deletions and duplications) made by CNest across all 200,000 UK Biobank samples, most calls are small (51% of all calls <100 kb) and the difference in loss-to-gain sensitivity is most evident for smaller events (Figure S2B). Most CNVs detected are small; however, only 1.8% of all calls include only a single exon, and, as size increases, the proportion of losses to gains stabilizes to approximately equal numbers above a size of 500 kb, with calls above this size accounting for 14.1% of the total call set (Figure S2B). Common CNVs are not uniformly distributed throughout the genome, and we find several high-frequency recurrent events in known CNV formation hotspots (often closer to low copy repeats and centromeric regions). We see strong correlation between the number of CNVs called by CNest for each chromosome with the total number of annotated segmental duplications for that chromosome (Figure S2C). There is a stronger correlation for losses compared with gains (Pearson's *R* of 0.82 for losses compared with 0.75 for gains), which is likely due to the dose response difference and decreased sensitivity for smaller gains. To assess the presence of the genomic wave within our final log2 ratio distribution, we calculated a genome wave estimate based on the interquartile range (IQR) of a running median using a 401-data-point span scaled by a scaling factor. Across all 200,000 sample-level normalized log2 ratio distributions, we observe very low levels of extreme wave characteristics, with only 3.2% of all samples having a genomic wave estimate greater than 1 (Figure S2D). Samples with a wave estimate greater than one represent those for which we would expect that the presence of wave-based noise in their log2 ratio distribution may make CNV calling challenging (Figure S3). When looking at the wave estimates in relation to the total number of losses and gains made per sample, we observe very tight distributions across the full range of total CNV calls (Figure S2D). Interestingly, our sample-level CNV calls appear to be largely robust to differences in wave-based noise, with most samples showing higher wave estimates being within the lower ranges of total CNV calls.

To further assess some characteristics of our CNV calls, we looked at how many predicted loss-of-function CNVs (either deletion or truncating duplications) overlapped clinically important genes from the dd gene to phenotype (DDG2P) resource;[24] we expect the common CNVs discovered in UK Biobank to be depleted in overlaps to these genes. Using a 50% reciprocal overlap rule against 218 mono-allelic loss-of-function genes from the DDG2P, we found a total of 342 individuals CNV calls (Figure S4), 40% of which were in the same gene, *GLMN*, which is known to be involved in glomuvenous malformations.[73] Overall, similar to previous work on pathogenic CNVs in the UK Biobank,[52] we detect small numbers of CNVs in clinically important disease genes across the UK Biobank, and rare variant analysis is not a focus of this study; however, we encourage interested researchers to make use of these high-resolution CNV calls (see data and code availability) where it might be possible to look at modifier effects for rare CNV events.

Copy number variation association testing in the UK Biobank

For CNV association testing genome-wide in discovery mode, we made use of both the copy number estimates and CNV calls generated by CNest and applied standard linear and logistic regression models using the copy number estimates as CNV dosage (see STAR Methods), analogous to the common dosage model of alleles from SNPs. Although the choice of linear models restricts our signal to sites displaying a linear relationship between copy number and trait, more sophisticated models that could have non-linear impacts on phenotypes can be complex to select and even more complex to analyze the resulting statistics consistently genome-wide. Furthermore, this simple model is like those most often used in SNP GWAS[74] and so is more easily jointly integrated with SNP discovery. All models were applied to unrelated samples from the principal-component analysis (PCA)-defined European cluster (SNP principal components [PCs] 1 and 2) and include standard covariates with 10 PCs derived from both SNP and CNV estimates independently.

We performed CNV association testing for 46 different main UK Biobank fields, including 30 quantitative and 16 binary traits across a variety of physiological, lifestyle, and health-related categories (Table S2). We used diagnostic QQ plots and the associated genomic inflation statistic to be confident that our model produced a well-behaved statistical test in which the majority of the genome fits the expected null hypothesis (Table S2). In total, after fine mapping to select the most associated probe for each CNV-phenotype association at a locus (see STAR Methods) we discovered 646 significant CNV-specific associations across 34 traits, 24 quantitative and 10 binary (Figure S5). We also selected all instances of the First Occurrences UK Biobank field that had greater than 500 cases mapping to an ICD10 code (UK Biobank field 1712), resulting in 398 different codes

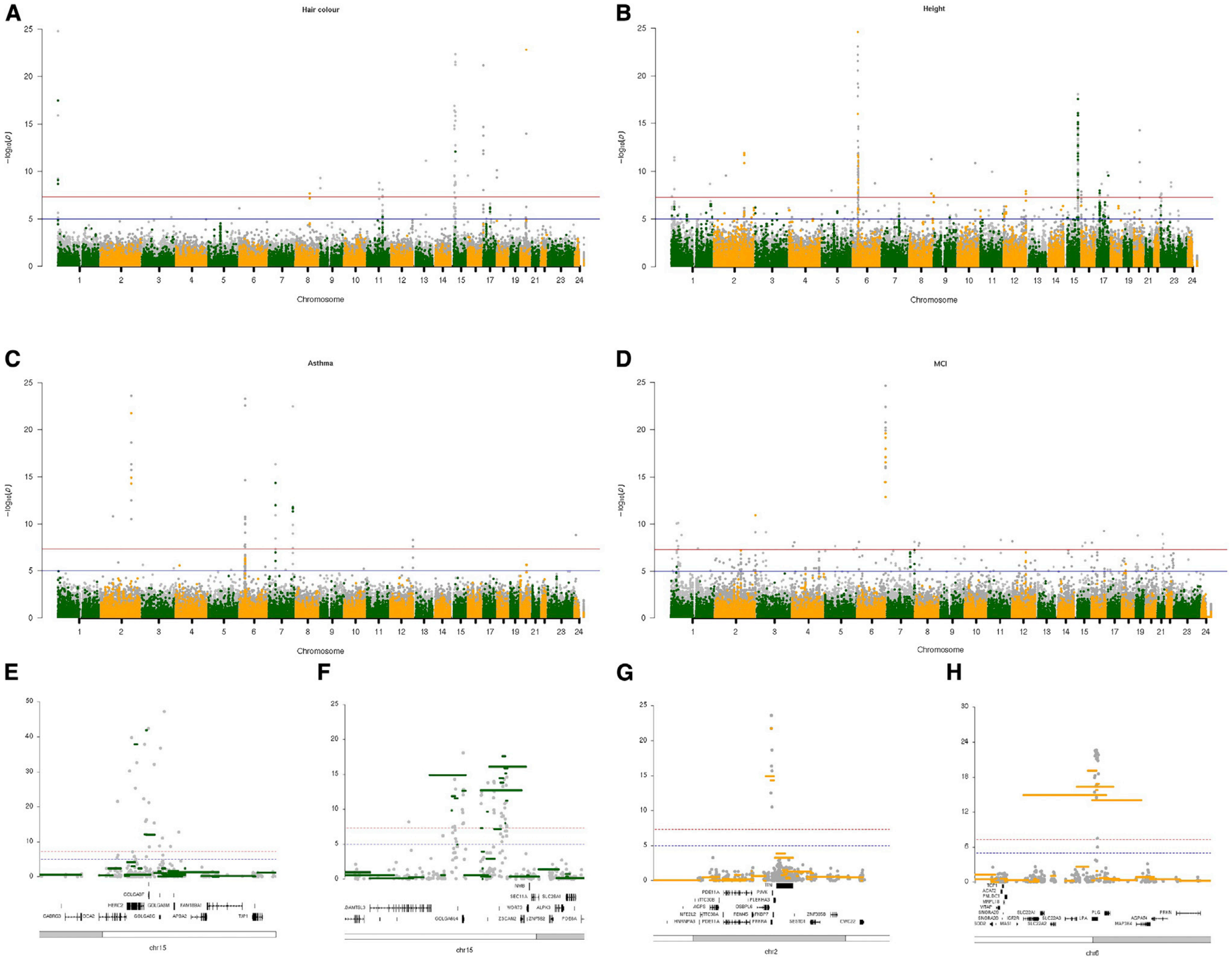


Figure 2. Copy number association Manhattan plots for four different UK Biobank traits

Exon-level signals are shown in different shades of gray and CNV call level signals in orange and green.

(A–D) Associations for (A) hair color using a linear model, (B) associations for standing height using a linear model, (C) associations for disease coding asthma using a logistic model, (D) associations for disease coding myocardial infarction using a logistic model.

(E) Zoom locus plot showing chr15 around the OCA2/HERC2 genes for hair color signal.

(F) Zoom locus plot showing chr15 around the ADAMTSL3/UBE2Q2L/GOLGA6L4 genes for standing height signal.

(G) Zoom locus plot showing chr2 around the genes CHROMR, PRKRA, and PJVK for asthma signal.

(H) Zoom locus plot showing chr6 around the LPA gene for myocardial infarction signal.

that we used as case/control labels for CNV association testing with logistic regression models. These 398 labels covered 15 broader categories (Table S3), and we obtained significant associations for 44 ICD10 codes across 13 broad categories. We show some specific examples (Figure 2) to illustrate these associations and their concordance to previous studies (see Data S1 for a further description).

For most UK Biobank main traits tested, we discovered new CNV-specific associations (Table S8); for example, for the eye-related trait corneal hysteresis, we detect robust CNV associations in exons 15 to 36 of the *ANAPC1* gene (Figure S6), in which sequence variation has been estimated to account for 24% of corneal endothelial cell density variability;[75] and for both corneal hysteresis and intraocular pressure, we discover exon-level associations in the important *TCF4* gene, which is known to be involved in several eye disorders, such as Fuchs corneal dystrophy[76] as well as haploinsufficiency of *TCF4* being strongly associated with Pitt-Hopkins syndrome.[77] Given that the UK Biobank participants were broadly healthy at recruitment, this association with eye phenotypes at the *TCF4* locus deserves further investigation. For red blood cell-related traits we detect a large number of associations that have prior evidence of association from SNP GWAS (Figure S7), such as variation in and around the *ABO* gene;[78] for lifestyle measures such as alcohol consumption, we find associations within known genes[79] such as *NPIPB6*; and for cognitive measures, we also discover CNV association

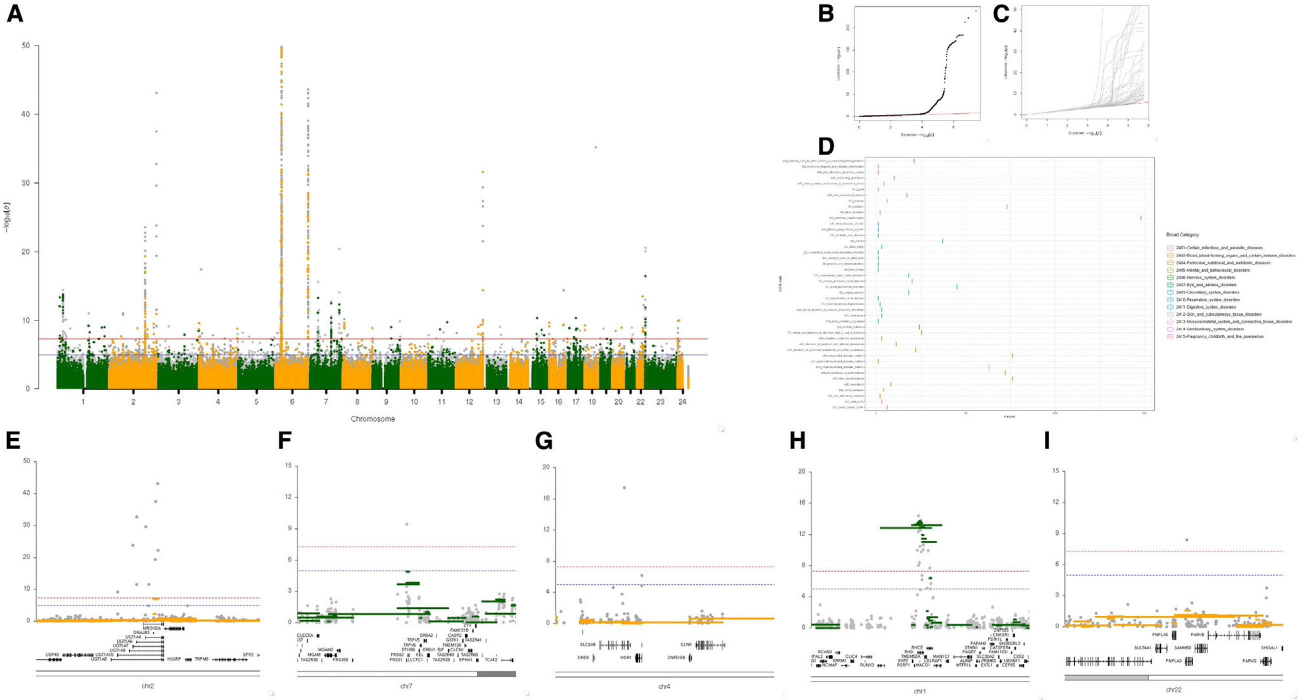


Figure 3. ICD10 code case/control copy number associations

(A) Combined and overlaid Manhattan plot for CNV associations across 44 ICD10 codes.
(B) Combined QQ plot including all p values from association results across all 44 traits.
(C) Overlaid QQ plots showing all individual QQ plots for the 44 traits.
(D) Plot showing the total number of exons for all ICD10 codes that had any significant signal.
(E) Locus zoom plot at UGT1A genes for ICD10 code E80 (disorders of porphyrin and bilirubin metabolism).
(F) Locus zoom plot at the PRSS1 gene for ICD10 code D50 (iron deficiency anemia).
(G) Locus zoom plot at the SLC2A9 gene for ICD10 code M10 (gout).
(H) Locus zoom plot at the RHD and RHCE genes for ICD10 code O 36 (maternal care for known or suspected fetal problems).
(I) Locus zoom plot at the PNPLA3 gene for ICD10 code K74 (fibrosis and cirrhosis of liver).

in genes with previous evidence of association from SNP GWAS testing in the UK Biobank, such as the *ARL17B* gene in association with reaction speed.[80]

All these CNV discoveries deserve integration with SNP polymorphisms and the often well-studied biology around these loci, and, as described in the "data and code availability" section, we have made all these results available to the community in a variety of ways. Here we provide important insight into the type of results possible to achieve for copy number association testing in large NGS cohorts bringing CNV GWAS into a similar framework to SNP-based tests and paving the way for further extensive studies to investigate the relationship between copy number and complex traits in humans.

First occurrences ICD10 code CNV associations

To complement the UK Biobank measured and binary traits we also explore CNV associations to direct healthcare measures, as represented by the Hospital Episode Statistics (HES)-captured data on ICD10 codes in the UK Biobank. We performed CNV association testing using the First Occurrences fields as case control labels for all codes that had greater than 500 cases and did not preselect or filter out any case labels, running CNV association testing across a total of 398 case control labels (Table S3). Across all 398 codes, we discovered 242 CNV-specific associations across 44 codes covering 144 unique genes (Figure 3A). A large fraction (117 out of 242) were located within the human leukocyte antigen (HLA) super locus at 6q21 between chromosome positions chr6:30500001 and 46200000, and there were six traits that had no associations outside of the HLA super locus, 13 traits that had associations both within and outside of the HLA, and 25 traits that had associations exclusively outside of the HLA.

After fine mapping, the majority of ICD10 codes (27 out of 44) had between one and two significantly associating regions, with nine ICD10 codes having between three and 10, and eight ICD10 codes having greater than 10 associations. Almost all the association results were well controlled with inflation factors (lambda) ranging from 0.984 to 1.140, with the exception of ICD10 code F17, mental and behavioral disorders due to use of tobacco, which showed mild inflation with a lambda of 1.382 (Figures 3B and 3C; Table S2). Most fine-mapped regions were small (Table S9), with a median number of significant exons of three per fine-mapped region (Figure 3D), with the largest region involving 52 exons across five different genes in association with ICD10 code K90: intestinal malabsorption. In total, we detected 242 associations ranging from well-known important

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

regions of the genome through to completely novel findings based on CNVs alone. All association results for the 44 significantly associating ICD10 codes are provided in the supplemental information. We provide specific examples (Figure 3) showcasing some of the new CNV associations that we have made by describing in detail some of the associations discovered across 10 different ICD10 code case control sets (see Data S2 for a further description).

In summary, we discovered 862 new fine-mapped CNV associations across 78 different traits (24 quantitative and 54 binary) using the 200,000 whole-exome release from the UK Biobank, the majority of which have either been previously discovered by SNP GWAS tests or have compelling evidence from other research areas such as health care, rare disease, or animal models (supplemental information), but a significant minority are entirely novel. We were able to detect more associations on average from quantitative compared with binary traits, with a median of four associations per quantitative trait compared with two for binary traits. These new association results and genome-wide association testing approach for CNV provides important insights into the contribution of CNV in complex human traits, which in some cases can have a direct relevance to health-related outcomes and genetic risk profiles. We encourage interested readers to pursue the discoveries discussed here and listed in our supplemental information.

Comparison with recent CNV association studies using SNP genotyping arrays

There have been some recent studies leveraging the ability of SNP genotyping arrays to detect CNVs and perform genome-wide association testing for copy number.[56,57] Although the technology is different and both recent studies used all available UK Biobank participants (~450,000), we sought to compare the associations we obtained here against those that could be detected using SNP arrays in the UK Biobank. Overall Auwerx et al.[56] detected 131 new associations across 47 quantitative traits with a mean size of 715 kb, and Hujoei et al.,[57] who additionally utilized identity by decent (IBD) information into the CNV detection, found 269 associations across 56 quantitative traits with a mean size of 467 kb compared with 862 associations across 78 quantitative and binary traits with a mean size of 9,970 bp with CNest using WES (Figures S9A and S9B).

Both recent SNP-based studies only performed GWAS on quantitative traits including many blood-related measurements and metabolic traits and there were nine traits overlapping both SNP array studies and the CNest results, resulting in 57 CNV associations that we could compare (Table S10). To compare these associations, we remapped all SNP array-based associations to the latest genome build (GRCh38) and interrogated the exome association signals across each site from the association tests on the same trait (see STAR Methods). In total, we found that 63% (36 out of 57) of locations could be confirmed (19 genome-wide and 17 suggestive) across all traits (Figure S9C), and certain traits (e.g., height and reticulocyte counts) had higher levels of agreement (Figure S9D). Some signals only reached suggestive significance levels, and we expect these would likely increase to genome-wide levels with larger sample sizes, whereas some regions showed no evidence of association from CNest association testing (Figures S9E–S9G). We also performed 100 rounds of permutation for each association (see STAR Methods) showing that the number of suggestive signals found for each trait was significantly higher ($p = 0.018$) for the SNP array-based CNV association regions compared with random areas of the genome with the same size (Table S11). This is a robust but by no means perfect concordance of CNV associations from these very different datasets (SNP based and WES based), different sample sizes, and different detailed methods (how the association is modeled). As expected, using WES, CNest results achieved higher resolution, detecting many smaller associations than both SNP-based studies using less than half the number of samples, variation exclusively in coding regions (exons) and the standard additive model most often applied to SNP GWASs. As all studies were able to find some unique associations, it may be useful to combine copy number information from whole-exome and SNP genotyping arrays for future association studies. Furthermore, the SNP array associations that show no signal from CNest results deserve further exploration once results are available using the same samples, traits, and model setups.

Combined CNV- and SNP-based associations in the UK Biobank

To investigate the relationship between SNP and CNV associations, we ran SNP-based GWAS tests across six quantitative traits using the same samples we used for CNV association testing (see STAR Methods). These six traits were selected to include a range of signals across different regions of the genome with differential signal strengths. The intention is to allow a direct comparison of SNP with CNV association signals across a variety of human traits. This allowed us to start to explore the underlying genome landscape for associations and to classify individual associations into those that were detectable by SNP and CNV GWAS independently against those that are specific to CNVs (see STAR Methods). We classified CNV signals into CNV only (signals that were detectable by CNV GWAS only), CNV-allele (signals that were present at the same locus by both SNP and CNV GWAS but with very little correlation between them), SNP-CNV near (signals that were detectable by both SNP and CNV GWAS and where those signals were highly likely to be assigned to the same gene), and SNP-CNV far (signals that could be detected by both SNP and CNV GWAS but were highly likely to be assigned to different genes).

Across 133 fine-mapped CNV association regions, we classified 17% (23/113) as CNV only, 44% (59 out of 133) as CNV-allele, 28% (38 out of 133) as SNP-CNV near, and 11% (13 out of 133) as SNP-CNV far (Table S12). We choose to be strict in the definition of novel CNV events (CNV only and CNV-allele) by setting the r^2 cutoff relatively low since very strong SNP-CNV tagging is rare genome-wide. Most exonic signals that could be well tagged by SNPs were found in regions involving recurrent CNVs with 68% found in CNV regions present in greater than 1,854 individuals (1% population frequency). We consider the SNP-CNV-far, CNV-allele, and CNV-only association classes as different types of novel CNV associations, whereas, for SNP-CNV near, we assume that the signals from both variant types are likely to be tagging the same functional

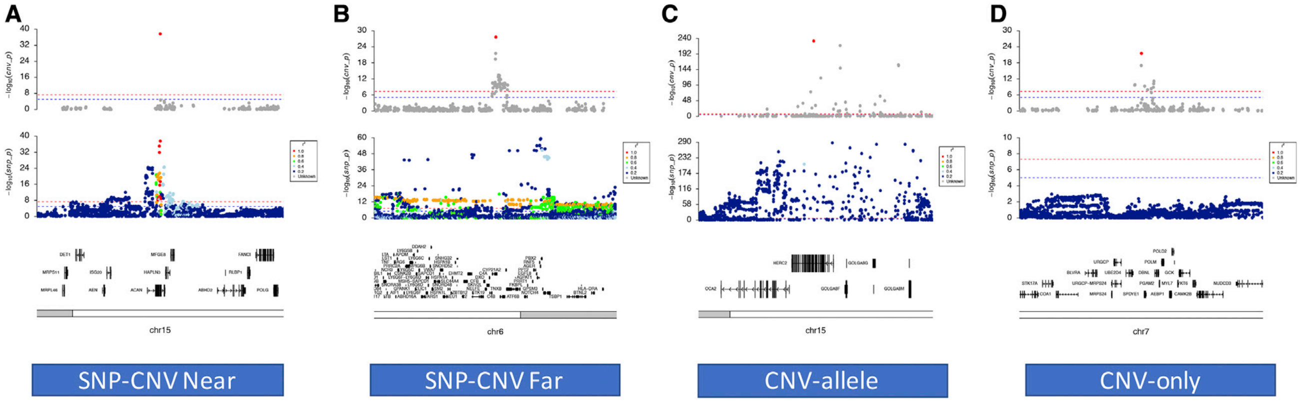


Figure 4. Locus zoom plots showing SNP and CNV association results for the different CNV association type classifications for four different quantitative traits

(A) SNP-CNV near association plot for standing height at ACAN.

(B) SNP-CNV far association plot for FEV/FEC ratio at C4A.

(C) CNV-allele association plot for hair color at HERC2.

(D) CNV-only association plot for chronotype at SPDYE1.

variant; nevertheless, the CNV association may well provide functional insight for the locus.

We show one example (Figure 4) and an additional two examples (Figure S10) for each of the CNV association type classifications covering a variety of SNP-CNV correlation patterns and differential signal strengths using locus zoom-style plots but with the focal point being the lead exon from the fine-mapped CNV region. For SNP-CNV-near and SNP-CNV-far classes, we show significant association signals for CNVs and SNPs (Figures 4A and 4B) restricted to a single exon of the *ACAN* gene association to height and a region including *C4A* and *C4B* genes with an association to the FEV/FEC ratio (see Data S3 for a further description).

Next, to investigate which gene-to-trait associations would be detectable by CNV association testing only, we classified novel CNV associations as CNV-allele at the *OCA2/HERC2* locus, which is described in detail above. Interestingly, although there is strong evidence of association, with similar signal strength, from both CNV- and SNP-based tests at *HERC2*, there is very little correlation between the two variant types (Figure 4C), suggesting that these associations are likely to be operating via different functional mutations. For the CNV-only class, we discover a highly specific association to chronotype within the *SPDYE1* gene on chromosome seven where there are no tagging SNPs and no SNPs associations within 1 Mb (Figure 4D). The *SPDYE1* gene has no previous association to measures of sleep patterns and very little description of CNVs at this location; however, a related gene, *SPDYE6*, has been associated with insomnia via SNP GWAS in a much larger sample set (1.3 million samples).[81] We also discovered significant CNV-only associations at *SPDYE6* and the directly adjacent region containing the *POLR* and *SPDYE2B* genes, which have been previously associated with chronotype by SNP GWAS in the UK Biobank.[82] Here we provide strong evidence that CNVs at *SPDYE1*-, *SPDYE6*-, and *POLR*-related genes are associated with a measure of chronotype in the UK Biobank and may act in a dose-dependent manner to influence an individual's sleeping pattern.

We also present two additional examples of each of the different CNV association classes (Figure S10), including SNP-CNV near associations to hair color at the *SPG7* gene[83] and to alcohol consumption at a 0.8-Mb region including four fine-mapped CNV regions involving the *NPIPB6*, *NPIPB7*, *NPIPB9*, and *SH2B1* genes;[79,84–86] SNP-CNV-far classifications for hair color at the *TRIM49C* gene and standing height at a fine-mapped region involving the *EVPLL* and *LGALS9C* genes; CNV-allele association types for heel bone density at the *WNT16* gene[87–90] and the FEV/FEC ratio at the *HTR4* gene[91–95]; and CNV-only classifications for the FEV/FEC ratio at the *ZDHHC11B* gene[91] and for standing height at the *CDK11A* gene.

Competitive SNP-CNV association models

We performed joint CNV-SNP competitive association models for standing height and hair color (see STAR Methods) since these traits include multiple signals genome-wide and have been extensively studied by previous work.[96,97] Across 91 exon-level CNV association signals that had at least one significant SNP within 1 Mb around the lead CNV position, we performed eight different models (see STAR Methods). First, we applied a three-component mixture model to the normalized copy number estimates (log2 ratio) to define copy number genotypes (assuming a simple deletion/gain process). Next, we performed joint SNP and CNV competitive models using both the SNP with the highest signal strength against the same trait, in the same samples, and the SNP showing the highest r^2 against the lead CNV exon within 1 Mb.

After the denoising of copy number estimates into a three-component model, the majority of associations (71 out of 91) showed a lowering of signal with 17 out of 91 sites dropping below genome-wide significance (Figures 5 A and 5B). For most sites, the association signal strength was lowered, with a median −log10 p value reduction of 1.7 for the three-state model; however, there were 20 out of 91 sites that showed a marginal improvement in signal strength with a median increase

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

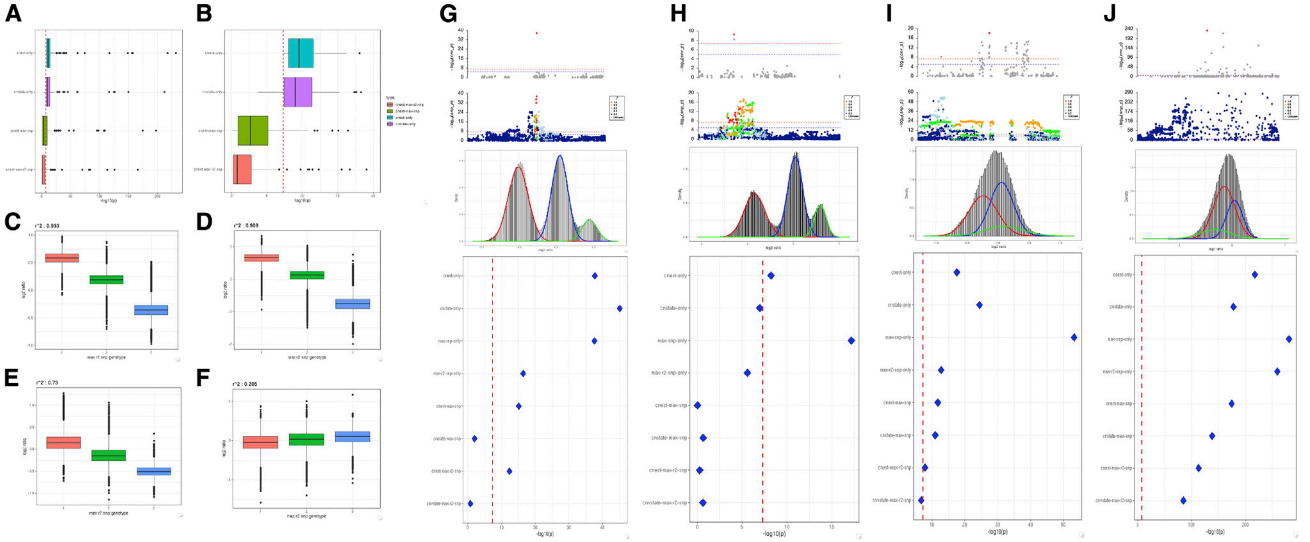


Figure 5. Competitive models for CNV and SNPs using copy number estimates, copy number genotypes, and joint models including SNP genotypes from the most highly correlated SNP or the SNP with the highest association signal for the same trait within 1 Mb

(A) Minus log10 p values for four different models: CNest only, copy number estimates only; cnstate only, copy number genotypes (three-component mixture model) only; CNest-max-snp, joint model with copy number estimates and the SNP with the highest association signal for the same trait within 1 Mb; CNest-max-r2-snp, joint model with copy number estimates and the most highly correlated SNP within 1 Mb.

(B) Zoomed in view of (A) restricting the x axis to a maximum −log10 p value of 20.

(C–F) SNP genotypes from the most highly correlated SNP against the copy number estimate (log2 ratio) for four individual exon-level association signals, further details of which are shown in (G)–(J).

(G–J) Finer-grain details for joint models of four exon-level copy number association signals; top panel shows the copy number estimate association signal with the lead exon highlighted in red, second panel shows the SNP genotypes association signal from SNP GWAS tests in the same samples and trait colored by the r^2 of SNP genotypes against the lead exon signal from the copy number GWAS (CN-GWAS), the third panel shows the copy number estimate (log2 ratio) of the lead exon association from the CN-GWAS fitted using a three-component mixture model to define copy number genotypes, and the fourth panel shows the −log10 p value from eight different types of association model: cnstate-only, cnstate-only, max-snp-only, max-r2-snp-only, CNest-max-snp, cnstate-max-snp, CNest-max-r2-snp, and cnstate-max-r2-snp.

of 0.65, and one site at the *GOLGA6L4* gene showing the relatively large increase of seven when using the CNV state compared with the CNV estimate model. The drop in p values is to be expected due to the selection of the most associated exon (winner's curse phenomenon), but it shows that there are no large gains to be made by a more categorical model for discovery.

We show the relationship between copy number estimates and SNP genotypes (Figures 5C–5F) as well as a finer-grained view of the difference in association signals when performing eight different competitive models (Figures 5G–5J). These four examples range from highly correlated and well-tagged CNVs (Figures 5C, 5D, 5G, and 5H) to moderately tagged CNV (Figures 5E and 5I), and finally a poorly tagged CNV (Figures 5F and 5J). For competitive SNP and CNV modeling, the simplest hypothesis is that sites that are well tagged and show association to the same trait are likely to be able to control each other's signal in a pairwise competitive model. Across the 91 sites, including the SNP with the highest signal strength irrespective of its r^2 can fully control 72% (66 out of 91) of CNV associations (Figures 5A and 5B). When switching the SNP to those with the highest r^2 irrespective of association strength, we observe an increased level of control with 80% (73 out of 91) of CNV signals being reduced below genome-wide significance. Indeed, the most frequent situation is that SNPs that can well tag the CNV in aggregate (across all samples) are best able to control the CNV association signals within a competitive model, with only two cases where highly correlated and significant SNPs are unable to fully control the CNV association (Figures 5G and 5I). As expected for the 18 CNV associations that cannot be fully controlled by either SNP type, the majority show very little tagging, with 80% of these sites having a maximum r^2 below 0.6 for all SNPs within 1 Mb. However, there are two cases where the main assumption that highly correlated SNPs can control the CNV association does not hold true (Figures 5G and 5I).

We show four examples of different types of association control from these competitive models. First, a case where neither SNP type included in the models can fully correct the CNV estimate signal, where we assume the SNPs are tagging the CNV (Figure 5G). Next, an example where both SNPs, either the most significant or the most highly correlated, can fully control the CNV association signal and we assume that the CNV tags the more significant SNP (Figure 5H). It is worth noting that the most significant SNP is found in intron 1 of the *DOCK8* gene, whereas the CNV signal and the tagging SNP are both found in the neighboring *CBWD1* gene, and that both genes have been found to be associated with the trait (hair color) by previous SNP GWAS testing.[83,96] We also show an example of a CNV association where neither SNP can fully control the CNV association, but the highest-correlated SNP is able to push the copy

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

number state joint model down just below genome-wide significance (Figure 5I). The highest signal SNP is found closest to a different gene upstream of the CNV signal and there are multiple tagging SNPs surrounding the CNV locus, which we assume all tag the CNV. Finally, we show an example of a highly significant CNV association signal that has very little SNP tagging around the locus and is not able to be controlled by either SNP in the competitive joint models, supporting our general assumption that SNPs that cannot well tag nearby CNVs are unlikely to be able to control any CNV association even if both variant classes show significant associations to the same trait (Figure 5J).

Here we have shown results from combined SNP and CNV association testing across close to 100 significant exon-level CNV associations. We have shown that the most obvious assumption that nearby tagging SNPs are often able to control CNV associations holds true but that more complex situations exist where aggregate variant correlations are not sufficient to predict the interactions between them in relation to trait association testing. We have also shown that it is possible to use copy number estimates in a dosage-dependent linear model as a reasonable proxy for the underlying copy number state distribution, and, by applying similar methods to the highly successful SNP GWAS approach, it is possible to discover novel CNV associations, adding additional supporting evidence for SNP-based trait association mapping further delineating the underlying genome architecture and variant interactions for trait associations.

DISCUSSION

In this paper, we present a robust CNV-to-phenotype discovery process that uses NGS information that is analogous to the traditional SNP-based GWAS. This paper therefore complements the long-standing use of CNV in rare disease discovery[98] and provides a higher-resolution view of common CNV than established SNP array-based methods.[21] A key foundation is a robust normalization procedure that can handle the diversity of DNA presentation and extraction states in a large cohort. Armed with this normalized copy number level, we decided to model the complexity of CNVs in the genome as a linear dosage variable; this model is obviously an approximation to the reality of structural variation, but it allows consistency of statistical approach and the same degree of freedom across the genome between loci, and means that similar covariates, methods, and diagnostic procedures, with similar expected null model properties to SNP GWAS, can be used.[99] Our resulting linear dosage model produces well-calibrated statistics for both quantitative and qualitative traits, where most associations fit the expected null model. The minority of associations where one can confidently reject the null model at a genome-wide significance level include many well-known individual CNV associations, with a total of 862 associations across 78 different human traits.

We have illustrated the large-scale discovery of CNV associations with 12 examples in the main text and an additional 18 examples in the supplemental information. The examples vary from well-established CNV loci (e.g., *LPA* with heart disease loci) through partially understood CNV loci (e.g., the *UGT1A* gene in porphyrin and bilirubin metabolism) to very credible associations to paralogues with the same phenotype (the RCHE gene in pregnancy complications) or credible novel alleles in a gene with robust association to a phenotype (HERC gene with hair color). Across most UK Biobank traits that we tested with sufficient sample size, we obtained strong discovery signals for CNV associations genome-wide, and, importantly, all signals were within exons providing a direct link to genes. Most (82%) of these fine-mapped association regions contained a single gene and sometimes single exons; however, when there are multiple genes across individual regions it can be hard to determine which gene is most responsible for the association. One such case is shown with the association of ICD10 code E80 (disorders of porphyrin and bilirubin metabolism) to multiple *UGT1A* genes. Although these minority of situations can be challenging to interpret, since we achieve exon resolution within our association tests, it is possible to rank individual exons (or genes) based on association signal strength or effect sizes. There are numerous other examples in the supplemental information, and the full information of the discovery processed here is available both via UK Biobank return of results and via the GWAS catalog. Even so, we have chosen only a subset of phenotypes present in the UK Biobank, itself only one cohort; to enable broader discovery by others, we have released CNest as an open-source package and provided portable workflows consistent with GA4GH standards. One of the central components of the association testing reliability is the generation of accurate copy number estimates. CNest uses a dynamic reference approach similar to previous work on microarrays and sequence datasets,[100–102] selecting an optimized set of internal samples to use as baseline copy number measures. We hope to build out more extensive user-friendly tutorials, including the practical and necessary aspects of QC before discovery. We encourage the community to examine the results we have presented here and use the software to make more discoveries.

When comparing our results with other studies into CNV association in the UK Biobank using WES for CNV detection and association testing focused on asthma,[103] we see good agreement in association signals for many of the asthma-specific associations we have reported in this study. When comparing results with CNV studies using SNP genotyping arrays,[56,57] we can confirm 63% of previous CNV associations made across nine quantitative traits. For those associations that were not confirmed by CNest, it is difficult to know whether these are real false-negatives from the exome association testing or are due to other technical or model choices. Some possibilities are that both SNP array-based studies used more than twice the number of samples, as well as differences in the type of signal that can be derived from the two platforms (WES versus SNP) where, for example, there may be certain exonic baits that are challenging for NGS due to differences in capture efficiency.[104] One other important factor is, unlike the two SNP array studies, we deliberately choose to assume additive effects, allowing us to place our results into a similar framework most often used in SNP GWAS, and did not attempt to model any non-linear effects. Nevertheless, we were able to obtain many more associations and at far higher resolution during this study, providing strong evidence that many copy number variable locations throughout the genome do associate with both continuous and discrete human traits. Further research is needed to fully catalog CNV associations across large cohorts such as the UK Biobank and beyond, and there are also many further interesting methodological questions to address where it may be useful to integrate information across multiple platforms for CNV discovery and association testing.

We have shown that it is possible to bring CNV GWASs into a similar framework as genome-wide SNP tests for trait association mapping in large cohorts, opening several new avenues of investigation into combined modeling of CNV association signals in human traits. We were also able to look at the correlation between SNP and CNV discoveries by performing association testing independently and jointly using the same sample sets and traits. Here it is possible to start to estimate the contribution that both types of variations have on trait associations in a large cohort and to gain some insight into the different types of interactions that can occur. As expected, for the CNV associations with some level of correlation to an SNP, there are complex relationships between SNPs and CNVs, and most CNVs that can be well tagged are relatively common in the population. In some cases, the SNP associations can completely explain the CNV association, whereas, in other less frequent situations, the CNV association cannot be recapitulated by any SNP. This latter case is consistent with multiple CNVs arising on different haplotypes, where the CNV association appropriately aggregates the CNV information in a way that is far harder to achieve via tagging SNPs. Even in the cases where the loci are discoverable by SNP methods, and the SNPs tag the CNV, the large impact of deletion or expansion of an exon makes CNV an interesting potential functional change.

CNV has long been known as an important aspect of germline DNA variation, and has long been a key part of rare genetic disease discovery and diagnosis. The system we have proposed here, CNest, can robustly find associations of CNVs to common phenotypes in large cohorts, but we have only started in providing a full catalog of these results. We encourage the community to explore the discoveries we have made in this paper, to use CNest to make more CNV associations in both UK Biobank and beyond, and to help extend the CNest method further to provide a more comprehensive view of human variation.

Limitations of the study

In this study, we have limited ourselves to associations consistent with a standard additive model that allowed for a more natural comparison with SNP GWAS results; however, there are known examples of CNV loci where both deletion or duplication (and in particular truncating duplications) can have a negative consequence on human phenotypes.[105] In further work, it will be interesting to expand the CNest framework to include additional models that can account for non-linear effects (such as U-shaped distributions or “mirror models”), which is very likely to result in an increase in the number of significant associations that can be made. Unlike most CNV GWAS studies to date, we have included association testing across multiple quantitative and binary traits, discovering many new associations across a large range of phenotypic measures. Similar to SNP GWAS,[106] we were able to discover more associations for quantitative traits compared with disease-related binary codes; however, we did make several discoveries related to human disease code, highlighting that certain genomic regions could contribute to disease progression in a dose-dependent additive manner.

The ability to jointly model SNPs and CNVs in the same framework will more easily allow for integration of these two types of variation. An obvious extension is to polygenic risk scores (PRSs) for traits, where the linear model for CNVs naturally fits with the additive linear SNP loci in a PRS.[107] However, care needs to be taken over ascertainment and modeling for PRSs, in particular for certain traits such as blood-based cancers where the normalization procedures we have employed for CNV association might not be robust enough to distinguish germline from somatic changes in cancer risk. Importantly, this means care needs to be taken about the time of blood sample compared with the onset of diseases in constructing such PRSs. Another extension will be using these linear variables as instrumental variables in Mendelian randomization techniques to understand causality between physiological processes and often disease outcomes.[108] A similar concern on normalization techniques needs to be considered, along with careful consideration of the assumptions behind any instrumental analysis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Sample cohort and phenotypes
  - ○ Genetic data processing and copy number estimation
  - ○ CNV merging, frequency estimation and copy number principal component analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Genetic association testing
  - ○ Definition of the association significance threshold
  - ○ Identifying associated genetic loci and fine mapping
  - ○ Comparison between SNP and CNV association signals
  - ○ Competitive SNP-CNV association models
  - ○ Comparison to previous CNV association studies using SNP genotyping arrays

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100167.

ACKNOWLEDGMENTS

We acknowledge all the UK Biobank participants and the UK Biobank management team without whom this study would not have been possible. T. Fitzgerald and E. Birney were funded by the EMBL European Bioinformatics Institute (EMBL-EBI).

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

AUTHOR CONTRIBUTIONS

T.F. and E.B. conceived and designed the study, performed the analysis, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 21, 2021
Revised: April 11, 2022
Accepted: July 13, 2022
Published: August 10, 2022

REFERENCES

1. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature *447*, 661–678.
2. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.
3. Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. *50*, 1112–1121.
4. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
5. Smith, S.M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K., and Elliott, L.T. (2021). An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. Nat. Neurosci. *24*, 737–745.
6. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease (2011). Nat. Genet. 43, 339–344.
7. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
8. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389.
9. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.
10. Teng, M.W.L., Bowman, E.P., McElwee, J.J., Smyth, M.J., Casanova, J.-L., Cooper, A.M., and Cua, D.J. (2015). IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. Nat. Med. *21*, 719–729.
11. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
12. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.
13. Nolte, I.M., van der Most, P.J., Alizadeh, B.Z., de Bakker, P.I., Boezen, H.M., Bruinenberg, M., Franke, L., van der Harst, P., Navis, G., Postma, D.S., et al. (2017). Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. Eur. J. Hum. Genet. *25*, 877–885.
14. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.
15. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525–528.
16. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.
17. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al. (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.
18. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
19. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.
20. Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., and Szemes, T. (2021). Copy number variation: characteristics, evolutionary and pathological aspects. Biomed. J. *44*, 548–559.
21. 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.
22. Boone, P.M., Bacino, C.A., Shaw, C.A., Eng, P.A., Hixson, P.M., Pursley, A.N., Kang, S.-H.L., Yang, Y., Wiszniewska, J., Nowakowska, B.A., et al. (2010). Detection of clinically relevant exonic copy-number changes by array CGH. Hum. Mutat. *31*, 1326–1342.
23. Grayton, H.M., Fernandes, C., Rujescu, D., and Collier, D.A. (2012). Copy number variations in neurodevelopmental disorders. Prog. Neurobiol. *99*, 81–91.
24. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.
25. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
26. 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.
27. Zhao, L., Liu, H., Yuan, X., Gao, K., and Duan, J. (2020). Comparative study of whole exome sequencing-based copy number variation detection tools. BMC Bioinf. *21*, 97.
28. Moreno-Cabrera, J.M., Del Valle, J., Castellanos, E., Feliubadaló, L., Pineda, M., Brunet, J., Serra, E., Capellà, G., Lázaro, C., and Gel, B. (2020). Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur. J. Hum. Genet. *28*, 1645–1655.
29. Usher, C.L., and McCarroll, S.A. (2015). Complex and multi-allelic copy number variation in human disease. Brief. Funct. Genomics *14*, 329–338.
30. Gross, A.M., Ajay, S.S., Rajan, V., Brown, C., Bluske, K., Burns, N.J., Chawla, A., Coffey, A.J., Malhotra, A., Scocchia, A., et al. (2019).

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet. Med. *21*, 1121–1130.

31. Talseth-Palmer, B.A., Holliday, E.G., Evans, T.-J., McEvoy, M., Attia, J., Grice, D.M., Masson, A.L., Meldrum, C., Spigelman, A., and Scott, R.J. (2013). Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. BMC Med. Genomics *6*, 10.

32. Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W.L., Pereira, S.L., Whitney, J., Chan, A.J.S., Pellecchia, G., et al. (2018). A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. Am. J. Hum. Genet. *102*, 142–155.

33. Minoche, A.E., Lundie, B., Peters, G.B., Ohnesorg, T., Pinese, M., Thomas, D.M., Zankl, A., Roscioli, T., Schonrock, N., Kummerfeld, S., et al. (2021). ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. Genome Med. *13*, 32.

34. Pfundt, R., Del Rosario, M., Vissers, L.E.L.M., Kwint, M.P., Janssen, I.M., de Leeuw, N., Yntema, H.G., Nelen, M.R., Lugtenberg, D., Kamsteeg, E.-J., et al. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. Genet. Med. *19*, 667–675.

35. Rajagopalan, R., Murrell, J.R., Luo, M., and Conlin, L.K. (2020). A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. Genome Med. *12*, 14.

36. Kirov, G., Pocklington, A.J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L., et al. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol. Psychiatry *17*, 142–153.

37. Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M., and Geschwind, D.H. (2016). Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. Am. J. Hum. Genet. *99*, 540–554.

38. Rees, E., Walters, J.T.R., Chambert, K.D., O'Dushlaine, C., Szatkiewicz, J., Richards, A.L., Georgieva, L., Mahoney-Davies, G., Legge, S.E., Moran, J.L., et al. (2014). CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. Hum. Mol. Genet. *23*, 1669–1676.

39. Szatkiewicz, J.P., O'Dushlaine, C., Chen, G., Chambert, K., Moran, J.L., Neale, B.M., Fromer, M., Ruderfer, D., Akterin, S., Bergen, S.E., et al. (2014). Copy number variation in schizophrenia in Sweden. Mol. Psychiatry *19*, 762–773.

40. Krumm, N., O'Roak, B.J., Karakoc, E., Mohajeri, K., Nelson, B., Vives, L., Jacquemont, S., Munson, J., Bernier, R., and Eichler, E.E. (2013). Transmission disequilibrium of small CNVs in simplex autism. Am. J. Hum. Genet. *93*, 595–606.

41. Carvalho, C.M.B., Bartnik, M., Pehlivan, D., Fang, P., Shen, J., and Lupski, J.R. (2012). Evidence for disease penetrance relating to CNV size: pelizaeus-Merzbacher disease and manifesting carriers with a familial 11 Mb duplication at Xq22. Clin. Genet. *81*, 532–541.

42. Zahnleiter, D., Uebe, S., Ekici, A.B., Hoyer, J., Wiesener, A., Wieczorek, D., Kunstmann, E., Reis, A., Doerr, H.-G., Rauch, A., and Thiel, C.T. (2013). Rare copy number variants are a common cause of short stature. PLoS Genet. *9*, e1003365.

43. Kessi, M., Xiong, J., Wu, L., Yang, L., He, F., Chen, C., Pang, N., Duan, H., Zhang, W., Arafat, A., et al. (2018). Rare copy number variations and predictors in children with intellectual disability and epilepsy. Front. Neurol. *9*, 947.

44. Swaminathan, G.J., Bragin, E., Chatzimichali, E.A., Corpas, M., Bevan, A.P., Wright, C.F., Carter, N.P., Hurles, M.E., and Firth, H.V. (2012). DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum. Mol. Genet. *21*, R37–R44.

45. Li, Z., Yu, D., Gan, M., Shan, Q., Yin, X., Tang, S., Zhang, S., Shi, Y., Zhu, Y., Lai, M., and Zhang, D. (2015). A genome-wide assessment of rare copy number variants in colorectal cancer. Oncotarget *6*, 26411–26423.

46. Dauber, A., Golzio, C., Guenot, C., Jodelka, F.M., Kibaek, M., Kjaergaard, S., Leheup, B., Martinet, D., Nowaczyk, M.J.M., Rosenfeld, J.A., et al. (2013). SCRIB and PUF60 are primary drivers of the multisystemic phenotypes of the 8q24.3 copy-number variant. Am. J. Hum. Genet. *93*, 798–811.

47. Aguirre, M., Rivas, M., and Priest, J. (2019). Phenome-wide burden of copy number variation in UK Biobank. Am J. Hum. Genet. *105*, 373–383.

48. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P.H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al. (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232–236.

49. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.

50. Warland, A., Kendall, K.M., Rees, E., Kirov, G., and Caseras, X. (2020). Schizophrenia-associated genomic copy number variants and subcortical brain volumes in the UK Biobank. Mol. Psychiatry *25*, 854–862.

51. Kendall, K.M., Bracher-Smith, M., Fitzpatrick, H., Lynham, A., Rees, E., Escott-Price, V., Owen, M.J., O'Donovan, M.C., Walters, J.T.R., and Kirov, G. (2019). Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. Br. J. Psychiatry *214*, 297–304.

52. Owen, D., Bracher-Smith, M., Kendall, K.M., Rees, E., Einon, M., Escott-Price, V., Owen, M.J., O'Donovan, M.C., and Kirov, G. (2018). Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. BMC Genom. *19*, 867.

53. 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.

54. 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.

55. Wiszniewska, J., Bi, W., Shaw, C., Stankiewicz, P., Kang, S.-H.L., Pursley, A.N., Lalani, S., Hixson, P., Gambin, T., Tsai, C.-H., et al. (2014). Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing. Eur. J. Hum. Genet. *22*, 79–87.

56. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Estonian Biobank Research Team; Porcu, E., Reymond, A., et al. (2022). The individual and global impact of copy-number variants on complex human traits. Am. J. Hum. Genet. *109*, 647–668.

57. Hujoel, M.L.A., Sherman, M.A., Barton, A.R., Mukamel, R.E., Sankaran, V.G., and Loh, P.-R. (2021). Influences of rare copy number variation on human complex traits. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.21.465308.

58. Fu, W., Zhang, F., Wang, Y., Gu, X., and Jin, L. (2010). Identification of copy number variation hotspots in human populations. Am. J. Hum. Genet. *87*, 494–504.

59. Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J., Lin, G.N., Malhotra, D., Watts, A.C., et al. (2016). Frequency and complexity of de novo structural mutation in autism. Am. J. Hum. Genet. *98*, 667–679.

60. Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L., et al. (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2, 396 families. Am. J. Hum. Genet. *108*, 597–607.

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

61. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

62. Torjesen, I. (2013). Genomes of 100, 000 people will be sequenced to create an open access research resource. BMJ *347*, f6690.

63. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e19.

64. Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in Healthcare: GA4GH Looks to 2022. Preprint at bioRxiv. https://doi.org/10.1101/203554.

65. Nielsen, J., and Wohlert, M. (1990). Sex chromosome abnormalities found among 34, 910 newborn children: results from a 13-year incidence study in Arhus, Denmark. Birth Defects Orig. Birth Defects Orig. Artic. Ser. *26*, 209–223.

66. Tuke, M.A., Ruth, K.S., Wood, A.R., Beaumont, R.N., Tyrrell, J., Jones, S.E., Yaghootkar, H., Turner, C.L.S., Donohoe, M.E., Brooke, A.M., et al. (2019). Mosaic Turner syndrome shows reduced penetrance in an adult population study. Genet. Med. *21*, 877–886.

67. Wang, S., Mandell, J.D., Kumar, Y., Sun, N., Morris, M.T., Arbelaez, J., Nasello, C., Dong, S., Duhn, C., Zhao, X., et al. (2018). De novo sequence and copy number variants are strongly associated with tourette disorder and implicate cell polarity in pathogenesis. Cell Rep. *25*, 3544.

68. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

69. Chen, Y., Zhao, L., Wang, Y., Cao, M., Gelowani, V., Xu, M., Agrawal, S.A., Li, Y., Daiger, S.P., Gibbs, R., et al. (2017). SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. BMC Bioinf. *18*, 147.

70. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome Sequencing Project; Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–1532.

71. Tsuang, D.W., Millard, S.P., Ely, B., Chi, P., Wang, K., Raskind, W.H., Kim, S., Brkanac, Z., and Yu, C.-E. (2010). The effect of algorithms on copy number variant detection. PLoS One *5*, e14456.

72. Uddin, M., Thiruvahindrapuram, B., Walker, S., Wang, Z., Hu, P., Lamoureux, S., Wei, J., MacDonald, J.R., Pellecchia, G., Lu, C., et al. (2015). A high-resolution copy-number variation resource for clinical and population genetics. Genet. Med. *17*, 747–752.

73. Brouillard, P., Ghassibé, M., Penington, A., Boon, L.M., Dompmartin, A., Temple, I.K., Cordisco, M., Adams, D., Piette, F., Harper, J.I., et al. (2005). Four common glomulin mutations cause two thirds of glomuvenous malformations ("familial glomangiomas"): evidence for a founder effect. J. Med. Genet. *42*, e13.

74. Cook, J.P., Mahajan, A., and Morris, A.P. (2017). Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. Eur. J. Hum. Genet. *25*, 240–245.

75. Ivarsdottir, E.V., Benonisdottir, S., Thorleifsson, G., Sulem, P., Oddsson, A., Styrkarsdottir, U., Kristmundsdottir, S., Arnadottir, G.A., Thorgeirsson, G., Jonsdottir, I., et al. (2019). Sequence variation at ANAPC1 accounts for 24% of the variability in corneal endothelial cell density. Nat. Commun. *10*, 1284.

76. Afshari, N.A., Igo, R.P., Jr., Morris, N.J., Stambolian, D., Sharma, S., Pulagam, V.L., Dunn, S., Stamler, J.F., Truitt, B.J., Rimmler, J., et al. (2017). Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. Nat. Commun. *8*, 14898.

77. Sweatt, J.D. (2013). Pitt-Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. Exp. Mol. Med. *45*, e21.

78. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746, 667 individuals from 5 global populations. Cell *182*, 1198–1213.e14.

79. Evangelou, E., Gao, H., Chu, C., Ntritsos, G., Blakeley, P., Butts, A.R., Pazoki, R., Suzuki, H., Koskeridis, F., Yiorkas, A.M., et al. (2019). New alcohol-related genes suggest shared genetic mechanisms with neuropsychiatric disorders. Nat. Hum. Behav. *3*, 950–961.

80. de la Fuente, J., Davies, G., Grotzinger, A.D., Tucker-Drob, E.M., and Deary, I.J. (2021). A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. Nat. Hum. Behav. *5*, 49–58.

81. Jansen, P.R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A.R., de Leeuw, C.A., Benjamins, J.S., Muñoz-Manchado, A.B., Nagel, M., et al. (2019). Genome-wide analysis of insomnia in 1, 331, 010 individuals identifies new risk loci and functional pathways. Nat. Genet. *51*, 394–403.

82. Jones, S.E., Lane, J.M., Wood, A.R., van Hees, V.T., Tyrrell, J., Beaumont, R.N., Jeffries, A.R., Dashti, H.S., Hillsdon, M., Ruth, K.S., et al. (2019). Genome-wide association analyses of chronotype in 697, 828 individuals provides insights into circadian rhythms. Nat. Commun. *10*, 343.

83. Landi, M.T., Bishop, D.T., MacGregor, S., Machiela, M.J., Stratigos, A.J., Ghiorzo, P., Brossard, M., Calista, D., Choi, J., Fargnoli, M.C., et al. (2020). Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. Nat. Genet. *52*, 494–504.

84. Zhou, H., Sealock, J.M., Sanchez-Roige, S., Clarke, T.-K., Levey, D., Cheng, Z., Li, B., Polimanti, R., Kember, R.L., Smith, R.V., et al. (2020). Meta-analysis of problematic alcohol use in 435,563 individuals identifies 29 risk variants and yields insights into biology, pleiotropy and causality. Preprint at bioRxiv. https://doi.org/10.1101/738088.

85. Justice, A.E., Winkler, T.W., Feitosa, M.F., Graff, M., Fisher, V.A., Young, K., Barata, L., Deng, X., Czajkowski, J., Hadley, D., et al. (2017). Genome-wide meta-analysis of 241, 258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat. Commun. *8*, 14977.

86. Brazel, D.M., Jiang, Y., Hughey, J.M., Turcot, V., Zhan, X., Gong, J., Batini, C., Weissenkampen, J.D., Liu, M., CHD Exome+ Consortium, et al. (2019). Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. Biol. Psychiatry *85*, 946–955.

87. Medina-Gomez, C., Kemp, J.P., Trajanoska, K., Luan, J., Chesi, A., Ahluwalia, T.S., Mook-Kanamori, D.O., Ham, A., Hartwig, F.P., Evans, D.S., et al. (2018). Life-course genome-wide association study meta-analysis of total body BMD and assessment of age-specific effects. Am. J. Hum. Genet. *102*, 88–102.

88. Surakka, I., Fritsche, L.G., Zhou, W., Backman, J., Kosmicki, J.A., Lu, H., Brumpton, B., Nielsen, J.B., Gabrielsen, M.E., Skogholt, A.H., et al. (2020). MEPE loss-of-function variant associates with decreased bone mineral density and increased fracture risk. Nat. Commun. *11*, 4093.

89. Kim, S.K. (2019). Correction: identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. PLoS One *14*, e0213962.

90. Morris, J.A., Kemp, J.P., Youlten, S.E., Laurent, L., Logan, J.G., Chai, R.C., Vulpescu, N.A., Forgetta, V., Kleinman, A., Mohanty, S.T., et al. (2019). An atlas of genetic influences on osteoporosis in humans and mice. Nat. Genet. *51*, 258–266.

91. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493.

92. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic

functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75.

93. Hancock, D.B., Eijgelsheim, M., Wilk, J.B., Gharib, S.A., Loehr, L.R., Marciante, K.D., Franceschini, N., van Durme, Y.M.T.A., Chen, T.-H., Barr, R.G., et al. (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat. Genet. *42*, 45–52.

94. Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Powell, J.E., Goddard, M.E., Wray, N.R., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Sci. Adv. *5*, eaaw3538.

95. Wain, L.V., Shrine, N., Artigas, M.S., Erzurumluoglu, A.M., Noyvert, B., Bossini-Castillo, L., Obeidat, M., Henry, A.P., Portelli, M.A., Hall, R.J., et al. (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat. Genet. *49*, 416–425.

96. Morgan, M.D., Pairo-Castineira, E., Rawlik, K., Canela-Xandri, O., Rees, J., Sims, D., Tenesa, A., and Jackson, I.J. (2018). Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. Nat. Commun. *9*, 5271.

97. Guo, M.H., Hirschhorn, J.N., and Dauber, A. (2018). Insights and implications of genome-wide association studies of height. J. Clin. Endocrinol. Metab. *103*, 3155–3168.

98. Chaves, T.F., Baretto, N., Oliveira, L.F.d., Ocampos, M., Barbato, I.T., Anselmi, M., De Luca, G.R., Barbato Filho, J.H., Pinto, L.L.d.C., Bernardi, P., and Maris, A.F. (2019). Copy number variations in a cohort of 420 individuals with neurodevelopmental disorders from the south of Brazil. Sci. Rep. *9*, 17776.

99. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nat. Rev. Methods Primers *1*, 21–59.

100. Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.-H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R., et al. (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature *471*, 499–503.

101. Barcelona-Cabeza, R., Sanseverino, W., and Aiese Cigliano, R. (2021). isoCNV: in silico optimization of copy number variant detection from targeted or exome sequencing data. BMC Bioinf. *22*, 530.

102. Fowler, A., Mahamdallie, S., Ruark, E., Seal, S., Ramsay, E., Clarke, M., Uddin, I., Wylie, H., Strydom, A., Lunter, G., and Rahman, N. (2016). Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. *1*, 20.

103. Fawcett, K.A., Demidov, G., Shrine, N., Paynton, M.L., Ossowski, S., Sayers, I., Wain, L.V., and Hollox, E.J. (2021). Exome-wide analysis of copy number variation shows association of the human leukocyte antigen region with asthma in UK Biobank. Preprint at medRxiv. https://doi.org/10.1186/s12920-022-01268-y.

104. Zhou, J., Zhang, M., Li, X., Wang, Z., Pan, D., and Shi, Y. (2021). Performance comparison of four types of target enrichment baits for exome DNA sequencing. Hereditas *158*, 10.

105. Wenger, T.L., Miller, J.S., DePolo, L.M., de Marchena, A.B., Clements, C.C., Emanuel, B.S., Zackai, E.H., McDonald-McGinn, D.M., and Schultz, R.T. (2016). 22q11.2 duplication syndrome: elevated rate of autism spectrum disorder and need for medical screening. Mol. Autism. *7*, 27.

106. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103.

107. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.

108. Smith, G.D., and Ebrahim, S. (2003). Mendelian randomization'': can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22.

109. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49, 960 individuals in the UK Biobank. Nature *586*, 749–756.

110. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat. Genet. *53*, 942–948.

111. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience *10*, giab007.

112. Bellenguez, C., Strange, A., Freeman, C., and Wellcome Trust Case Control Consortium, Wellcome Trust Case Control Consortium; and Spencer, C.C.A. (2012). A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics *28*, 134–135.

113. 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.

114. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.

115. van den Berg, S., Vandenplas, J., van Eeuwijk, F.A., Lopes, M.S., and Veerkamp, R.F. (2019). Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. J. Anim. Breed. Genet. *136*, 418–429.

116. Lee, B.T., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., et al. (2022). The UCSC Genome Browser database: 2022 update. Nucleic Acids Res. *50*, D1115–D1122.

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK BioBank Whole Exome Sequence (200,000 release)	UK BioBank	Data field: 23143
GWAS Catalogue summary statistics ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90103001-GCST90104000/GCST90103348	GWAS Catalogue (master summary stats file)	Project: GCP000324
UK BioBank Data returns (to be submitted on publication)	UK BioBank	Application: 49978
Software and algorithms		
CNest: https://github.com/tf2/CNest	GitHub	3c2b94a
CNest: https://app.terra.bio/#workspaces/ga4gh-cnest-test/cnest-terra	Terra Bio	NA
CNest: https://hub.docker.com/repository/docker/tomas81/cnest	Docker Hub	tomas81/cnest:dev
CNest: https://zenodo.org/record/6770130#.YsLe6-zMJMM	Zenodo	https://doi.org/10.5281/zenodo.6770130
ViteRbi: https://zenodo.org/record/6794409#.YsLliuzMJMM	Zenodo	https://doi.org/10.5281/zenodo.6794409
Examples of association tests and plotting CNest results: https://zenodo.org/record/6806357#.YsavNuzMJMM	Zenodo	https://doi.org/10.5281/zenodo.6806357

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Tomas Fitzgerald (tomas@ebi.ac.uk).

Materials availability

CNV GWAS summary statistics and fine mapped association regions are included in the supplementary material of this paper.

All association results, including all sites tested irrespective of association signal strength, have been submitted to the GWAS catalogue and are available for download under the project ID: GCP000324.

CNV calls and copy number estimates will be made available via the UK BioBank data return and linked to UK Biobank application number: 49978.

Data and code availability

The main CNest code base and docker setup can be found here: https://github.com/tf2/CNest

This repository contains all the source code and a docker setup as well as a link to a NextFlow workflow. There is a WDL workflow featured inside the Terra platform along with example datasets and a tutorial for getting CNest up and running across a diverse set of computational infrastructure including cloud based systems (https://app.terra.bio/#workspaces/ga4gh-cnest-test/cnest-terra). This tutorial and the additional workflow implementations is linked to from the CNest main repository. All code has been deposited at Zenodo and can be found at the following links, https://zenodo.org/record/6770130#.YsLe6-zMJMM, https://zenodo.org/record/6794409#.YsLliuzMJMM and https://zenodo.org/record/6806357#.YsavNuzMJMM.

METHOD DETAILS

Sample cohort and phenotypes

For this study, we used 200,624 Whole Exome Sequencing datasets from the UK Biobank 200k release generated using the IDT xGen Exome Research Panel v1.0 including supplemental probes and sequenced with dual-indexed 75 × 75 bp paired-end reads on the Illumina NovaSeq 6000 platform using S2 and S4 flow cells[109]. We used the aligned CRAM files from the OQFE pipeline which aligned and duplicate-marked all raw sequencing data (FASTQs) against the full GRCh38 reference in an alt-aware manner as described in the original FE manuscript[110]. These aligned sequence datasets were used as the primary input in the CNest pipeline (**details below**) for exome-wide copy number estimation and CNV calling. Phenotypes were extracted and linked to the copy number data under UK Biobank application number 49978, resulting in a total of 78 different traits (24 quantitative and 54 binary) that we tested for CNV association (Tables S2).

Genetic data processing and copy number estimation

We used CNest (full source code available: https://github.com/tf2/CNest.git) to carry out large scale copy number estimation in the UK Biobank 200k WES release. This program was designed to provide accurate copy number estimation from very large NGS (WES and WGS) datasets. The first required step is to extract read coverage information for all genomic locations of interest, to do this CNest makes use of the samtools and htslib libraries[111] implementing a custom coverage extraction method that importantly filters reads based on several samtools alignment flags. The main flags of interest are BAM_FPROPER_PAIR, BAM_FDUP and BAM_FSECONDARY where we ensure that aligned reads have a MAPQ greater than 1, are primary alignments with proper pairs and are not PCR duplicates.

After extraction of coverage information, the first important step is to classify the sex of each sample based on the relative coverage on chromosome X. Here CNest implements a simple k-means clustering for the initial classification and quality control steps. This initial step results in the classification of two states relating to 2 or more and 1 or less copies of chromosome X (although less than 1 copy of chromosome X is biologically incompatible there can be data quality issues to account for when processing large volumes of data). CNest also implements a prototype automated classification to detect sex chromosome auniopoly which is based on the ‘abberant’ cluster[112] however we highly recommend that the sex classification in checked by a human before moving onto the next steps. This is because all datasets are different and will contain a variety of sex classification types (Figure 1A), as standard across large numbers of samples we observe 3 types of sex chromosome dosage exclusion types and classify samples into either male or female, or ambiguous low, ambiguous mid, ambiguous high. All samples that are not classified as either male or female are removed from the subsequent steps.

The next step is to derive sample specific dynamic reference datasets, briefly, CNest uses an optimisation process to select groups of appropriate datasets to make up individual references (or baseline) estimates across the genome for each sample individually, similar approaches have been proposed and successfully applied to derive copy number estimates from coverage level data. One such approach can be found within the virtual reference genome (VRG) approach from the C-SCORE method[100]. It is worth noting that by using this type of approach it is impossible to directly obtain the true copy number at any genome location, rather it allows the optimisation of relative copy number estimates that are often (particularly if the reference set size is sufficiently large) likely to be a good reflection of the underlying copy number state distribution.

During its baseline reference selection processes CNest estimates the overall correlation of coverage information between all samples, applies a wavelet model for estimating the scale of genomic waves, and implements a dose response optimisation using sex mismatched samples and the expected single copy dosage change. This process has been designed to be extremely efficient across very large numbers of samples and results in a ranked list of which samples are most correlated in terms of certain coverage patterns and noise characteristics which are assumed to be the ideal set of samples to generate the baseline estimate. The only parameter needed to be decided on at this stage is the total number of datasets to use to generate the dynamic references, for the UK Biobank 200K release we elected to use 2,000 samples within each of the $\sim$200K dynamic reference datasets. Although it may, in some cases, be preferable to allow the dynamic reference sets to be made up of different numbers of individual samples (which is possible using CNest) we decided to fix this number across all datasets as we wanted to minimise any potential biases within the resulting copy number estimates that could be introduced due to using differentially sized reference datasets.

Following the dynamic reference selection process, the median coverage for all genomic locations across all relevant reference datasets is calculated for each sample individually for matched, mismatched, and mixed sex classifications. These data values are stored in the custom CNest binary format to allow fast random access across the genome and across different sample sets. Finally, the coverage information and reference estimates are transformed into the log2 ratio space and median normalised using the median log2 ratio excluding sex chromosomes for sex matched, mismatched, and mixed estimates. These estimates are again saved back into the CNest binary file format to allow efficient extraction during the next steps of CNV detection and CNV GWAS testing.

For CNV calling CNest implements a single custom designed 3 state Hidden Markov Model (HMM) to call losses and gains, the basic implementation of which can be found (https://github.com/tf2/ViteRbi). In our hands this HMM model has been highly reliable across a number of different detection applications, and it is extremely efficient in terms of speed and memory usage. We apply a few important steps during the HMM training (using the EM algorithm) to further improve the reliability of the model. Primarily these include training the model independently for each sample by using all the log2 ratio estimates of all samples that made up the dynamic reference for that particular sample. By doing this we aim to increase the accuracy of the transition probabilities by giving the EM algorithm sight of large numbers of highly similar datasets. Importantly the log2 ratio estimates that we use during this training phase are those generated using the sex mismatched dynamic reference, ensuring that there is always at least one large single copy number loss/gain event present in all training datasets. Having trained the HMM for each sample independently using this approach we apply the forward backward Viterbi algorithm to call the most likely sequence of state paths across each sample independently. The result of this process is the state calls (0,1,2) for every genomic position of interest across all samples, we then apply some merging criteria to obtain both state classification and CNV regions (CNVR) across the genome. In fact, although CNV callers will often impose some complex merging criteria to account for outlier points within each called CNVR (e.g. by allowing a certain number of outlier copy number estimates within a CNVR) we are so confident in the performance of the HMM that we simply merge consecutive state calls without any additional complex merging rules. As illustrated in Figure S4, this process does not result in over fragmentation of CNV regions.

CNV merging, frequency estimation and copy number principal component analysis

Having obtained reliable copy number estimates and CNV calls across all samples we apply some cross sample merging criteria to allow us to generate merged copy number events (CNVEs) with frequency information attached. To do this we merge losses and gains separately across all samples using an iterative 50% reciprocal overlap rule, building up sets of CNVRs across all samples where all member segments (calls) within each set must share at least 50% of its boundaries (start and end positions) with at least one other segment within the set. Once we obtain full closure of the set, when no additional segment can be added to the set, we adjust the final start and end position by 80% of the inner to outer start and end positions. Finally, we calculate loss, gain and overall CNV frequency and standard errors for each CNVE resulting in a set of CNV regions across the genome with frequency estimates that can be used in some of the subsequent analyses (e.g. PCA and Association testing).

On top of the frequency estimate for all merged CNVEs we also assign frequency measures to individual bait regions where we can calculate the frequency that each bait is included within any copy number event. This gives us two different sets of CNV type and frequency datasets that can be used to perform principal component analysis (PCA) across the sample space. PCA is often used in SNP GWAS to control for population structure and other technical (or sample level) variation that is less well understood but that is important to control for during genome wide association scans. Similarly, it is important to correct for larger scale differences between copy number estimates across large datasets for CNV GWAS analysis. Often during SNP based PCA the sites get filtered or subsampled to allow efficient PCA to be performed, for CNV analysis it seems important that we can run PCA analysis using both commonly variable CNV regions and rare regions separately. This is due to an observation that when including all CNV sites in PCA, the commonly variable positions tend to suck out a lot of variation and get overrepresented in the first PCs. We used iterative PCA to perform several different types of PCA using both bait level and CNV call level copy number information stratified by frequency estimates. Overall, we find that PCA based on commonly variable positions are better able to capture sample level information such as population structure, whereas PCA based on rare regions can account for cryptic sample differences which are likely due to certain noise properties of the data that we were unable to accurately model during the previous steps.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genetic association testing

One major point of the CNest methods and approach is that by working with copy numbers in this way we have been able to employ genetic association testing methods like the SNP based GWAS methods that have been applied with great success over many years. For CNV we can use several different estimate types to perform large scale genome wide association tests. Although it would be possible to develop methods for copy number genotyping (i.e. actual copy number states) due to the way we have set up our large scale approach we are not able to accurately determine the real copy number of any individual genome position. Rather we have well calibrated relative copy number estimates across large numbers of samples that can be used to search for associations against any given traits.

We set our models up in a few different ways but always (in this study) by using standard linear and logistic regression techniques, although this choice is potentially suboptimal (Discussion) this was done deliberately to ensure that any CNV association signals follow the general additive model (where the copy number estimate must display a linear relationship against the given trait). All models were applied to unrelated samples from the PCA-defined European cluster (SNP PCs 1 and 2). For quantitative traits we use generalised linear models with covariates and use both the bait level copy number estimate (log2 ratio) and the copy number estimate (mean log2 ratio) for all merged CNVEs across all samples as the test variables. The standard set of covariates we include are sex, age, sequencing batch, the first 10 PC from SNP based PCA and the first 10 PCs from CNV PCA for both rare and common sites separately. Additionally, to ensure that outliers in the phenotype distribution do not impact our association tests, for all quantitative traits we apply an inverse rank normalisation.

For SNP based association tests in exactly the same sample sets we used bgenie[113] for quantitative and regenie for binary case control trait tests[106]. Imputed SNP genotypes from the 500K UK Biobank release[113] were remapped to genome build hg38 using the UCSC liftover tool[114] prior to sample selection based on the 162,633 samples that were used in the CNV association tests. In both cases, for quantitative and qualitative tests, we followed the standard SNP filtering recommendations, including only bi-alleilc SNPs with a minor allele frequency greater than 1%. Association tests were run across the main traits of interest and the genome wide significance cut-off of 5e-08 was used to define associations between SNPs and traits.

Definition of the association significance threshold

To justify the use of the widely accepted genome wide significance threshold of 5e-08 for significance testing in this work we looked at how our results would change when using three different p value correction approaches. We assessed the use of a stringent Bonferroni correction, the Benjamini Hochberg (BH) false discovery rate (FDR) based approach and permutation tests.

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

Overall, applying the stringent Bonferroni correction only slightly lowered the significance threshold to a value of 3.35e-08 and had very little effect on the number of significant exon level associations for most tests (Figure S11) with a median decrease of 0 (mean decrease of 1.19 and maximum decrease of 21) across all traits. However, applying this stringent correction did result in the exclusion of 0.5% (45/862) fine mapped regions across 18 of the 78 traits tested, after correction only 0.5% (4/78) of those traits had zero remaining significant associations at the specific loci since each of these 4 traits only had a single low level exon signal. Next, we applied a 0.01 FDR based BH correction to each association result independently, again we observed highly consistent results for most traits with a median increase of 1 significant exon signal across all traits (Figure S11B). For the majority of traits, the BH correction resulted in less stringent thresholds than both the genome wide and Bonferroni approaches (Figure S11C), and in most cases resulted in either the same (28 traits) or increased (40 traits) numbers of exon level signals that would be defined as significant. For some traits (8/78) the number of additional significant associations for the BH correction was substantially higher (greater than 100 additional exonic signals) suggesting that there could be some value in applying an FDR based correction. Finally, we performed 100 rounds of permutations on 4 main traits (hair colour, height, MI and Asthma) where we randomly ordered the phenotype measurements or case labels and ran our standard linear or logistic regressions models across 100 different random sets for all 4 traits. In all cases after 100 rounds of permutations there were zero signals that passed a genome wide threshold of 5e-08 (Figure S12) indicating that this value is suitable for use in the definition of copy number associations.

Although it may be possible to use a less stringent threshold (such as BH or permutation based) and to obtain a greater number of copy number based trait association, we preferred to remain highly stringent. Copy number associations often display a similar association pattern to SNP GWAS tests genome wide where close by exon signals are highly correlated and associate to the same trait (LD peaks) and the use of the 5e-08 significance value had the additional benefit of allowing us to apply the same definition of significance for both SNP and CNV based association tests.

Identifying associated genetic loci and fine mapping

We have developed a set of tools that build on top of the CNest framework to allow large-scale genome wide association testing for CNVs - CNwas. These tools perform several of the important steps described above - namely CNV merge, PCA and GWAS testing using regression models. Since we have placed CNV GWAS analysis into a similar framework to that often used for SNP GWAS analysis we can make use of standard approaches for genetic loci detection and quality control. Firstly, we use the accepted genome wide significance threshold of 5e-08 to define associations between copy number and traits, although in this case we could theoretically lower this cut-off by using, for example, an FDR or permutation approach we preferred to remain highly stringent for the results we describe in this work (see definition of significance threshold). It also now becomes possible to use standard diagnostic approaches to association results, such as QQ plots and permutation. We apply these standard approaches to the CNV association results described here and see that in general the distributions of p values from our association tests are well controlled. For some tests we do see a degree of inflation and calculate the inflation factor - lambda - for all tests (Table S2), overall for the majority of tests we get inflation factors below 1.13 which is generally considered to be acceptable in GWAS tests[115] and for cases where the inflation factor is above 1.13 we suggest that a level of caution is used when interpreting these results.

We did not perform fine mapping of SNP based association signals as it is not a focus of this work to provide SNP based GWAS results, however we did fine map CNV signals to define fine mapped regions of CNV association that we report (**GWAS catalogue**) and that we could use for investigation into CNV compared to SNP level signals during the next steps. Because our genomic test loci are, by definition, in coding regions of the genome we choose to use a relatively simple approach for fine mapping CNV associations in a gene/exon centric way. First, we merged all directly adjacent significant signals that had no intervening signal below the significance threshold, next we merged all significant signals that were found in the same gene(s). This resulted in a fine mapped list of significant copy number regions that can contain a single exon, multiple exons within a gene or multiple exons across multiple genes, to be clear these regions do not always contain the full coding region of the gene however any intervening not significant signals between two significant signals within the same gene are merged as that intervening region is assumed to be important for the underlying CNVs. For reporting purposes, the -log10 p value is reported for the lead exon signal within each fine mapped region and tests for correlation between SNP and CNV signals in the next section are always performed in relation to the lead CNV signal for each fine mapped region.

Comparison between SNP and CNV association signals

One question that we wanted to address with this work is that of how SNP and CNV associations for the same traits interact with each other and how many of the CNV specific association results would have been detectable using standard SNP GWAS tests. To explore this we defined a set of classification rules to allow us to classify each fine mapped CNV region as either, CNV-only (not detectable using SNP GWAS), CNV-alelle (detectable signal from SNP-GWAS but very hard to discover using tagging variation), SNP-CNV-near (detectable by SNP GWAS and very likely to be fine mapped to the same gene), SNP-CNV-far (detectable by SNP GWAS however likely to be mapped to a different gene).

First, we calculated r^2 between the lead CNV signal and all SNP genotypes within 1 MB. For the CNV-only type, if no significant SNP signal within 1MB was closest to the gene from the lead CNV signal, irrespective of the r^2 value, the fine-mapped region was classified as CNV-only i.e. not detectable by standard SNP GWAS. Regions are classified as CNV-allele if there were significant SNP associations closest to the gene but if none of those SNP had an r^2 greater than 0.6 i.e. association is detectable by both SNP and

Please cite this article in press as: Fitzgerald and Birney, CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100167

CNV GWAS but are not tagging. Next, if any significant SNP within 1MB did have an r^2 greater than 0.6 and if any of those SNPs were closest to the gene(s) inside the fine mapped CNV region than the region was classified as SNP-CNV-near i.e. the association signal was taggable by a SNP that associated with the same trait and was highly likely to be assigned to the same gene. Finally, if there were significant SNPs within 1MB that have an r^2 greater than 0.6 but if all those SNPs were closer to a gene that was not within the fine-mapped CNV region then we classify these as SNP-CNV-far. It is worth noting that we set our SNP to CNV r^2 cut-off quite low at a value of 0.6, this is because genome wide we observe relatively low r^2 between SNPs and CNVs and although there are numerous cases of very well tagged CNVs (r^2 > 0.9) for these results we decided to be strict with the definition of novel CNV associations meaning that when we classify an associated fine-mapped CNV region as CNV-only we can be very confident that it is not well tagged by any associating SNP within 1 MB and this results in an overall decrease in the number of CNV associations that we classify as CNV-only.

Competitive SNP-CNV association models

To look in greater detail at the relationship between SNP and CNV association interactions we performed some joint modelling of two different copy number estimates, log2 ratios and approximated copy number state distributions, by including either the SNP genotypes from the most significant SNP association or the best tagging SNP (highest r^2 values to the lead CNV association) within 1MB around the CNV association location as covariates within a standard linear model. First we selected 91 single exon associations across 2 traits (hair colour and standing height) and extracted the log2 ratio values across all UK Biobank samples included in the association testing and fitted a 3 component mixture model to define the approximate copy number states boundaries, the decision to use a 3 component was twofold, firstly it was to avoid complications in the definition of the number of actual copy number states observed across the 91 sites that would be highly likely to cause problems during mixture model fits and secondly it was to place the copy number state models into a similar categorical distribution to the SNP genotype models. Each sample was assigned a copy number state based on its most likely component from the mixture model and additionally we did not allow any sample to cross the mean of any adjacent component resulting in a 3-state copy number model relating to low, medium, and high copy numbers. Next, we fitted 8 different types of models across all 91 sites where we tested all variant classes independently and additionally included both SNP types (most significant association and best tagging SNP within 1MB around the CNV association) in pairwise competitive linear models for the copy number estimate and copy number state distributions. We extracted both p values and beta effect sizes from each variant type from all the eight different models and carried out a comparison of signal strength across the different models, allowing us to look in more detail at the relationship and interactions between SNP and CNV associations using close to one hundred CNV association discoveries.

Comparison to previous CNV association studies using SNP genotyping arrays

We obtained all reported CNV association signals from two previous studies[56,57] that were performed using CNVs detected in the full UKK SNP genotyping cohort. First, we remapped these association loci to the latest genome build (GCRh38) using the UCSC liftover tool[116] with most regions being successfully lifted over (258 / 269 for *Hujoei* et al.[57] and 114/132 for *Auwerx* et al.[56]). Next, we selected all associations that were made against a trait that overlapped any traits from the CNest results, both studies only tested quantitative traits and we were able to find 9 overlapping traits resulting in 57 CNV associations that we could compare. With these remapped positions and traits, we then interrogated all CNest association signals from the same trait across each region independently and defined the regions as being genome-wide significant if any signal passed the 5e-08 threshold and suggestive if any signal passed the 1e-05 threshold. Although in some cases it may be possible to redefine the critical region within these association loci due to the higher resolution of the exome data, we did not attempt to resolve any region since we did not have access to the raw copy number SNP signals underlying each associated region. Additionally, we performed one hundred rounds of permutation on each of the association regions separately, randomly selecting a region of the genome with the same size and counted the number of times any signal passed either the genomewide or suggestive threshold showing that these regions are strongly enriched for suggestive signals (Table S11).

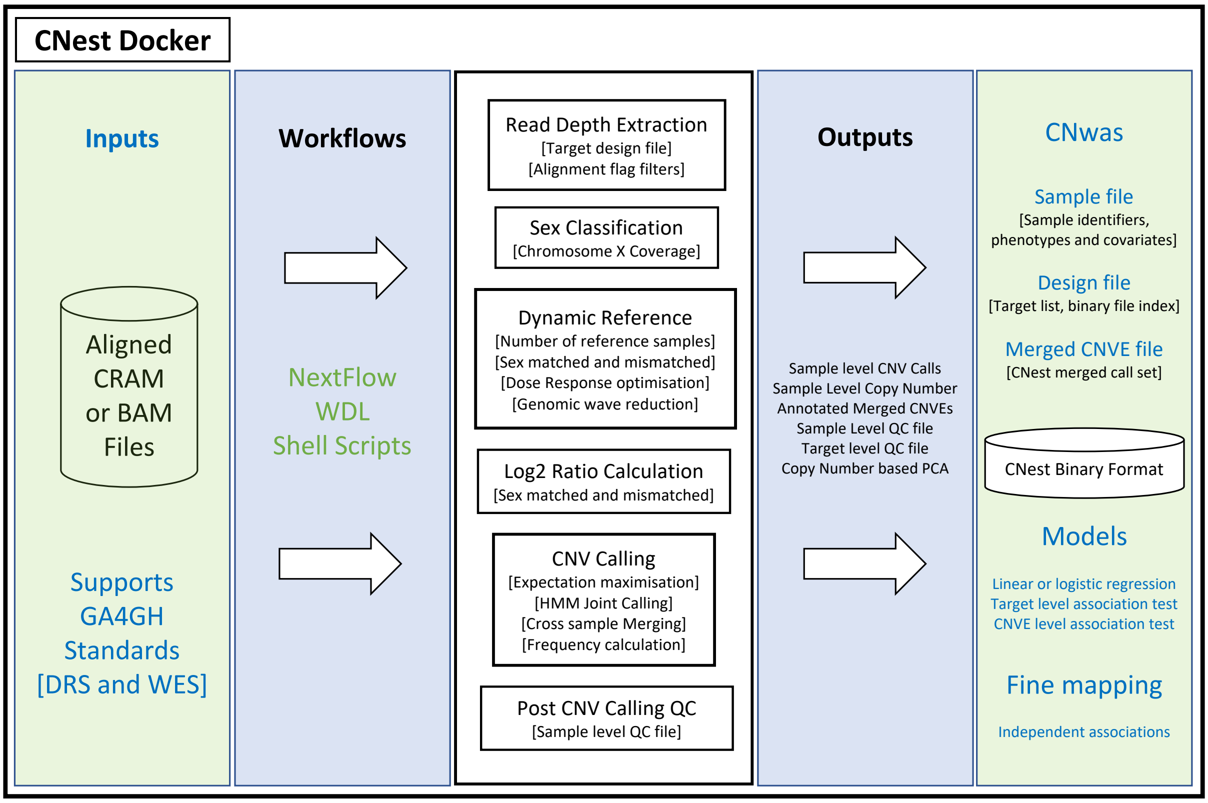
Cell Genomics, Volume *2*

Supplemental information

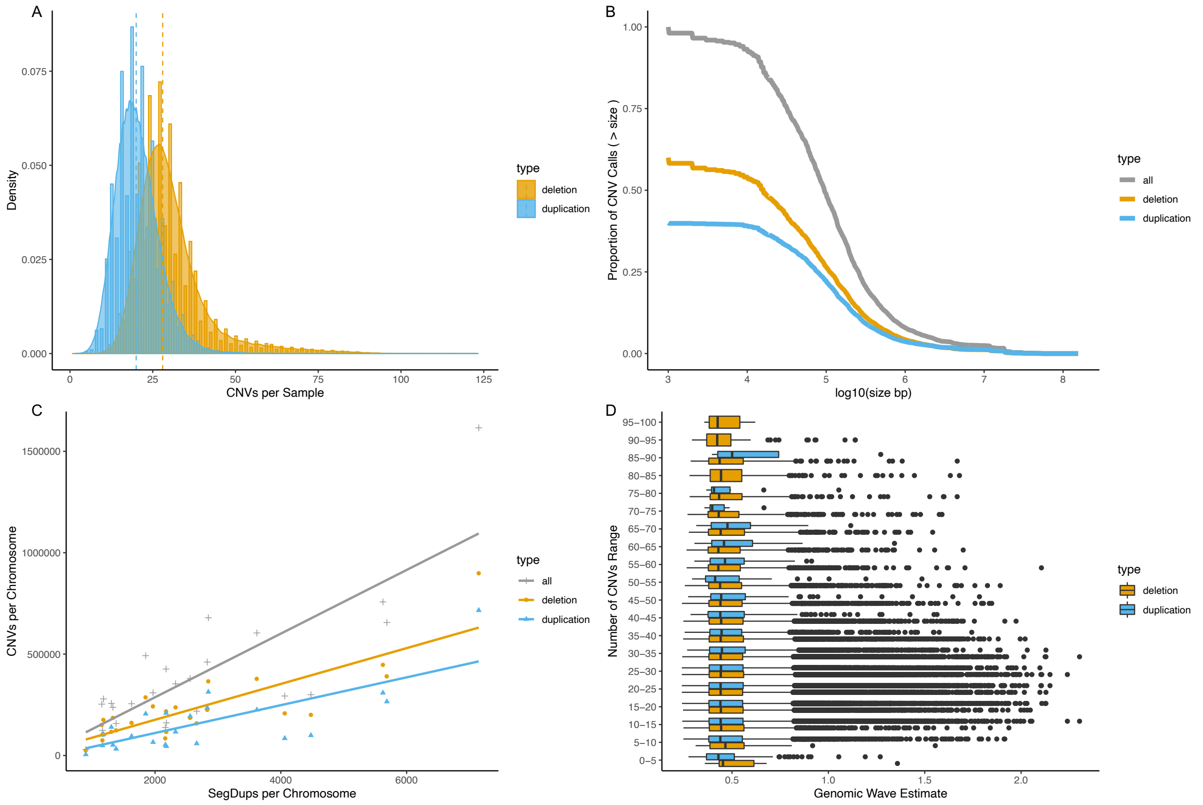
CNest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank

Tomas Fitzgerald and Ewan Birney

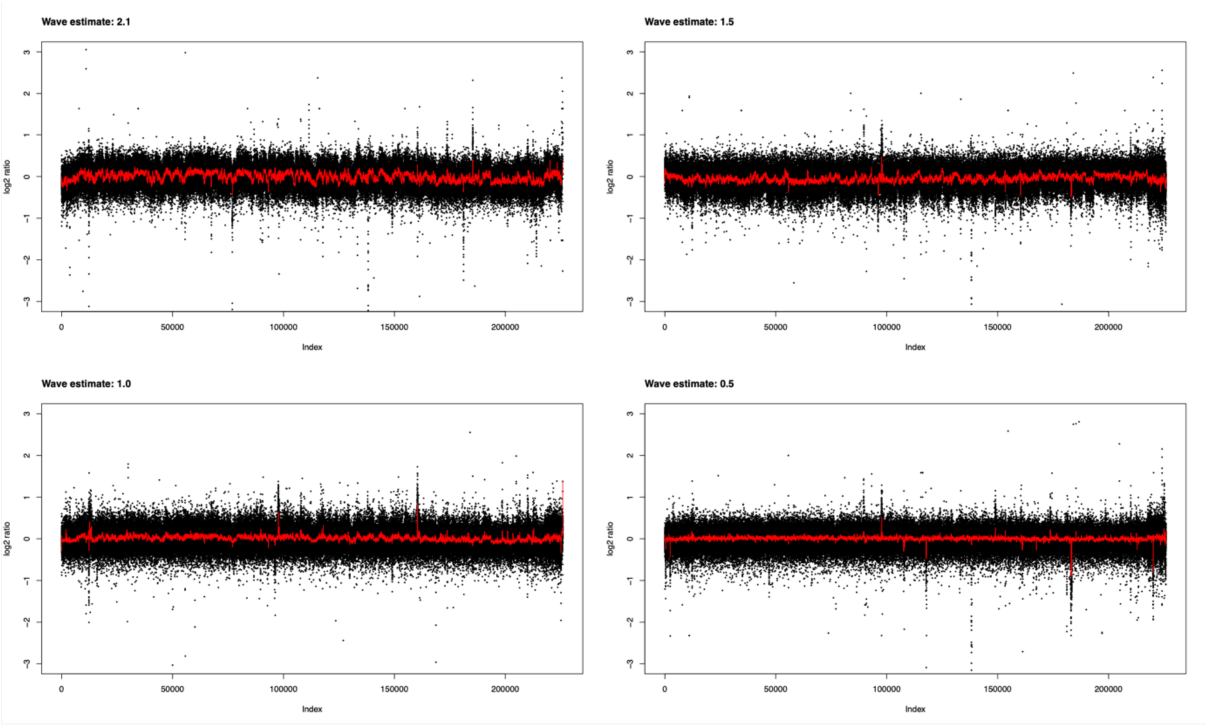
Supplementary Material



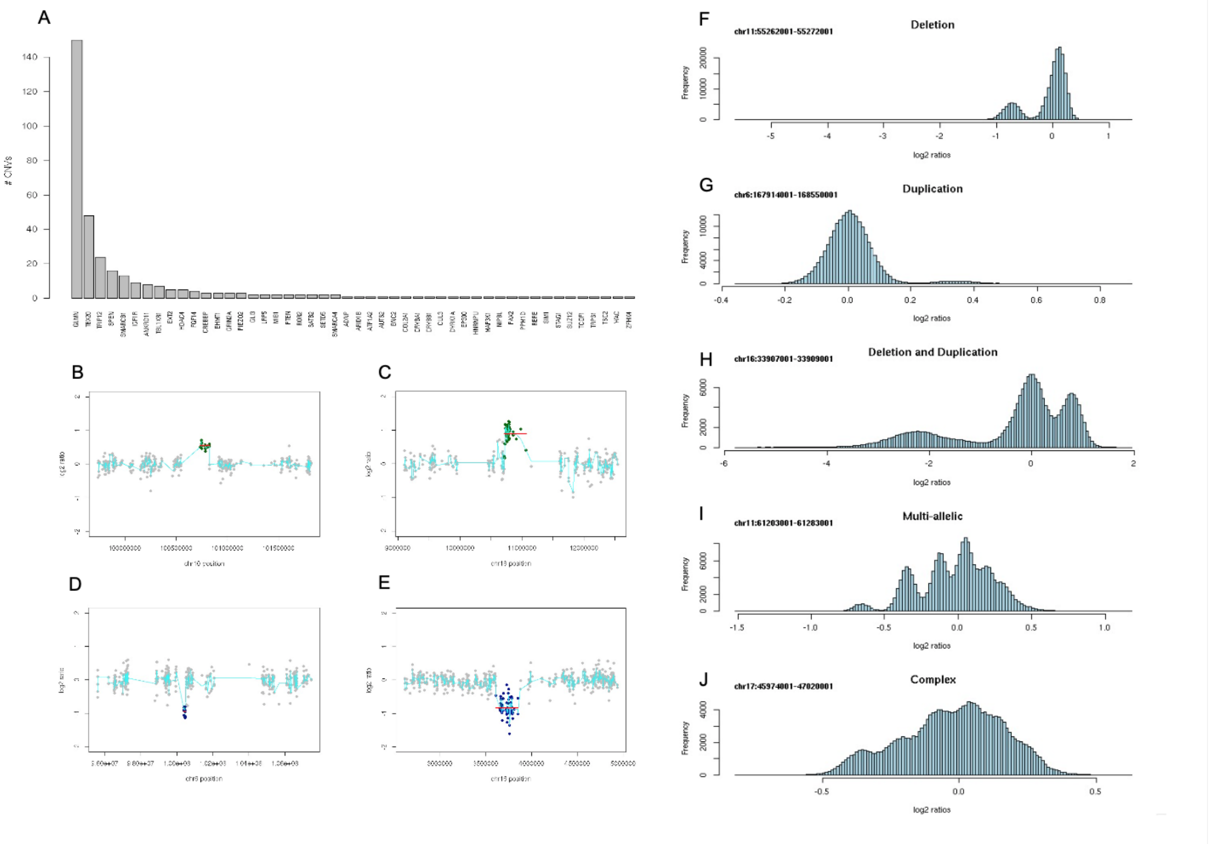
Supplementary Figure 1: *CNest flow diagram showing the main steps performed by CNest for CNV detection and association testing using next generation sequence data – relating to STAR methods (CNest methods). All processes are contained inside the CNest docker and available as workflows in several workflow languages (WDL, Nextflow and shell scripts). The primary input into CNest are aligned sequence files in CRAM or BAM format (and these associated index files). Th workflows run in a number of different ways, where, for example, the WDL workflow runs from start to finish in a single end to end process. Starting from aligned CRAM or BAM files and a target index file specifying the genomic regions to estimate copy number at, CNest via its WDL workflow will run all the individual steps required resulting in, sex classification, sample level CNV calls, merged and annotated CNV events, sample and target level QC files and CNV based PCA results. Next, for association testing via CNwas, a few important decisions need to be made relating to the level of sample and target QC to apply and the covariates to include within the association testing framework (these can be specific to the cohort although we provide some recommended default values). To perform the association testing CNwas needs a few inputs, primarily the output from CNest, copy number estimate in a custom, highly efficient binary format, a sample file including sample identifiers, phenotypes to test and covariates, the target ("bait") design file and the merged CNVEs from CNest. CNwas operates under the same paradigm as tools such as "plink" or "begenie" using 3 primary input file types relating to sample, design and genotype information (".fam", ".bim" and ".bed" style formats). The type of test to perform (linear or logistic) needs to be specified and the process can be split into multiple jobs by specifying a chunking parameter. The result will be target level and CNVE level association results across all traits included in the sample file, which can then be fine mapped using a secondary process.*



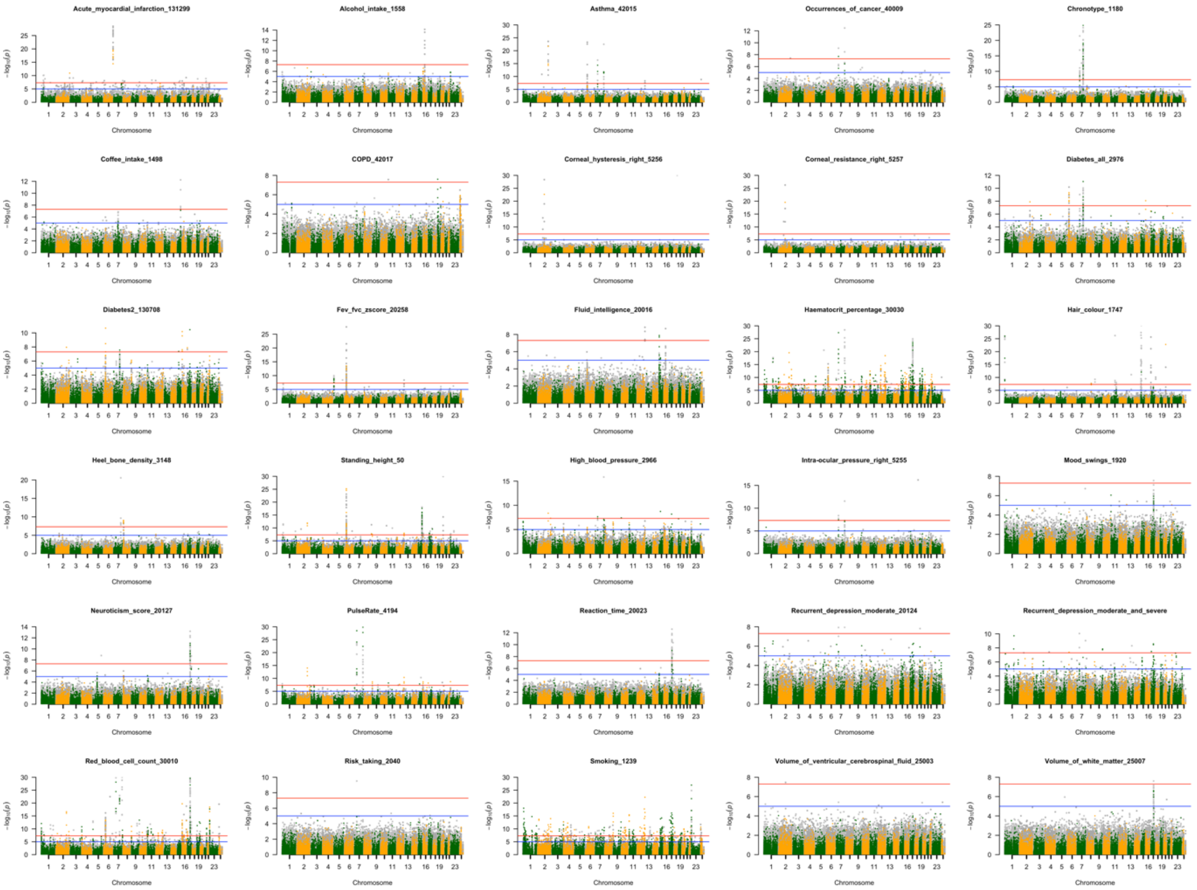
Supplementary Figure 2: *CNV call summary information across ~200K UK Biobank Whole Exome sequences – relating to Figure 1. A: The distribution of the number of CNVs called (deletions in orange and duplications in blue) per sample. B: The log10 of the number of base pairs per CNV call against the total proportion of CNV calls (all calls in grey, deletions in orange and duplications in blue) greater than that size. C: The number of Segmental Duplications per chromosome (GRCh38) against the total number of CNV calls per chromosome (all calls in grey, deletions in orange and duplications in blue). D: The distribution of a genomic wave estimate (IQR of a running median across sample level log2 ratio distributions, using a span of 401 data points) separated across the range of the number of CNV call made per sample between zero and 100 in intervals of 5 CNV calls (deletions in orange and duplications in blue).*



Supplementary Figure 3: *Examples of the log2 ratio values for 4 different samples across the range of genomic wave estimates based on the interquartile range (IQR) of a running median using a 401 data point span scaled by a scaling factor – relating to Figure 1. Upper left: An example of extreme wave characteristics with a wave estimate of 2.1. Upper right: Example of a moderate wave level with a wave estimate of 1.5. Lower left: Example of a mild wave level with a wave estimate of 1.0. Lower right: Example of a sample showing very low level of wave characteristic with a wave estimate of 0.5 Note: wave estimates of 0.5 is where the majority of the CNest normalised sample level log2 ratio copy number estimate in the 200K UK Biobank Whole exome sequence are centred around (**see Supplementary Figure S2D, CNV call summary information, CNV calls**).*



Supplementary Figure 4: *Individual CNV calls and copy number variable locations in the UK Biobank – relating to Figure 1. A: Barplot showing the number of CNV calls overlapping any of 218 monoallelic loss of function genes from the DDG2P (dd gene to phenotype). B: Truncating duplication at the PAX2 gene. C: Truncating duplication at the PIEZO2 gene. D: Deletion at the SIM1 gene. E: Deletion at the CREBBP gene. F: Deletion locus at 11q12.1. G: Duplication locus at 6q27. H: Deletion / duplication locus at 16p11.2. I: Multi-allelic locus at 11q12.2. J: Complex locus at 17q21.31.*



Acute_myocardial_infarction_131299

Alcohol_intake_1558

Asthma_42015

Occurrences_of_cancer_40009

Chronotype_1180

Coffee_intake_1498

COPD_42017

Corneal_hysteresis_right_5256

Corneal_resistance_right_5257

Diabetes_all_2976

Diabetes2_130708

Fev_fvc_zscore_20258

Fluid_intelligence_20016

Haematocrit_percentage_30030

Hair_colour_1747

Heel_bone_density_3148

Standing_height_50

High_blood_pressure_2966

Intra-ocular_pressure_right_5255

Mood_swings_1920

Neuroticism_score_20127

PulseRate_4194

Reaction_time_20023

Recurrent_depression_moderate_20124

Recurrent_depression_moderate_and_severe

Red_blood_cell_count_30010

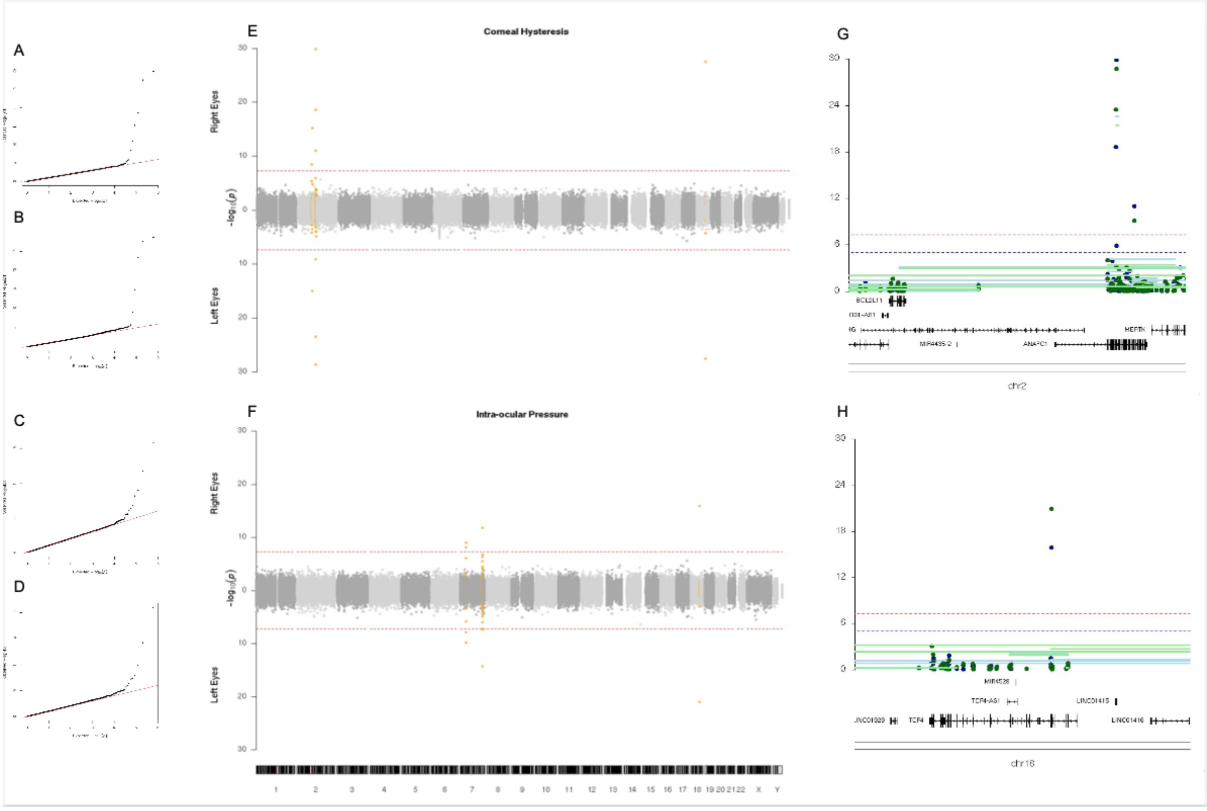
Risk_taking_2040

Smoking_1239

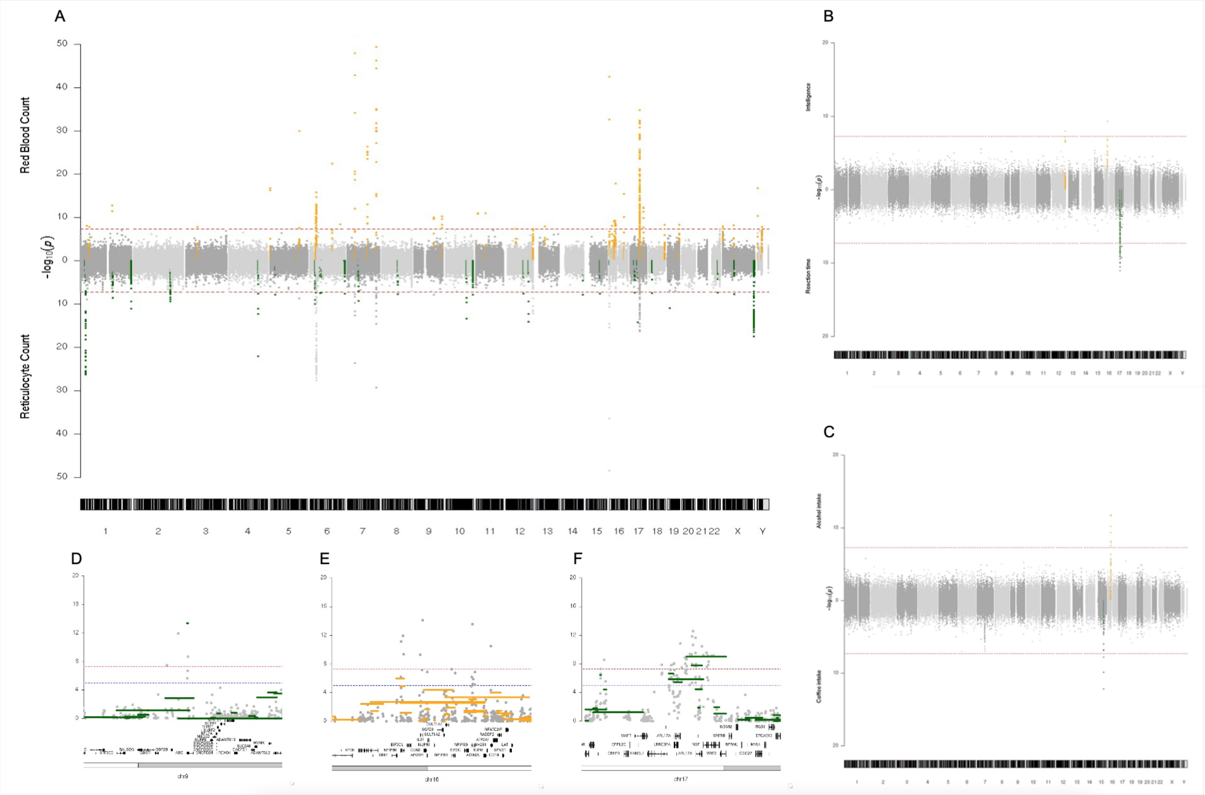
Volume_of_ventricular_cerebrospinal_fluid_25003

Volume_of_white_matter_25007

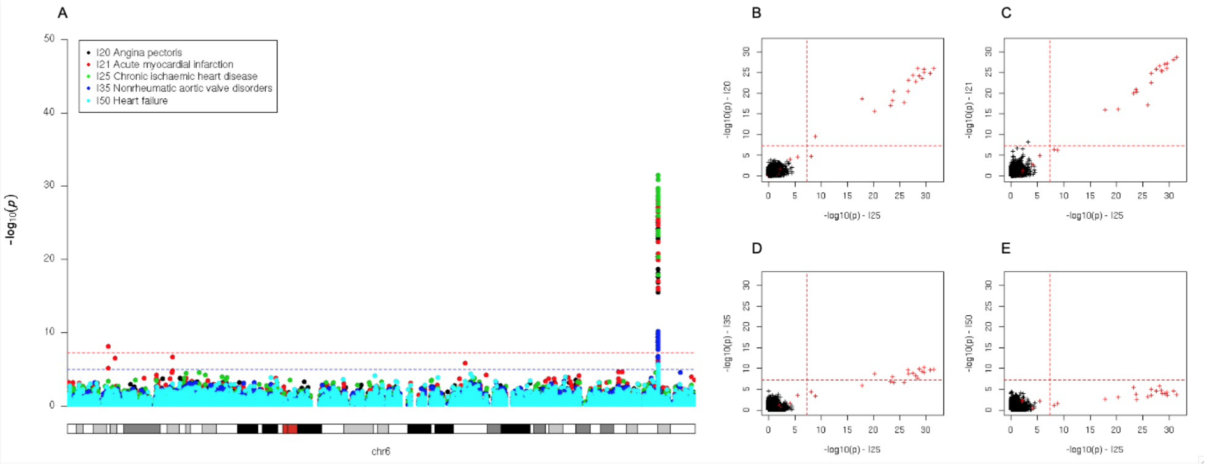
Supplementary Figure 5: *Individual manhattan plots for 30 of the 34 main traits for CNV association in the UK Biobank 200K whole exomes – relating to Figure 2. All plots are pinned to a maximum -log10 p-value of less than 30, meaning that all stronger association signals are not shown but this significantly aids the visualisation across all traits. We exclude 4 traits, showing only right eyes for eye related traits and only red blood cell counts for red blood cell related traits. All the 30 panels have a title showing the trait and all include both p-values from exon level (copy number estimates) trait association testing (grey) and CNV call ("genotype") association testing (green and orange).*



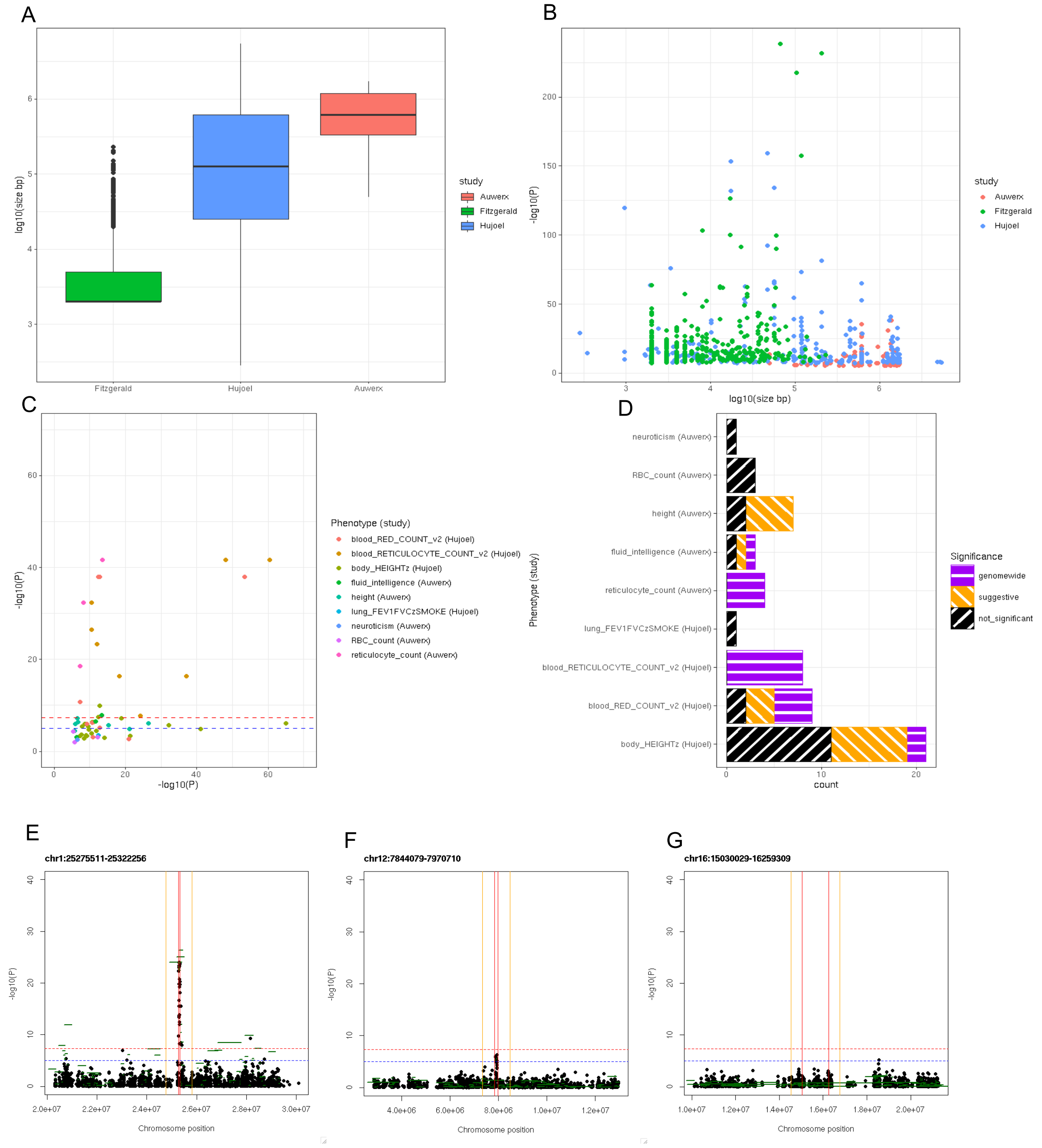
Supplementary Figure 6: *CNV association results for the eye related traits, Corneal hysteresis and Intra-ocular pressure for left and right eyes separately – relating to Figure 2. A: QQ plot for Corneal hysteresis in right eyes. B: QQ plot for Corneal hysteresis in left eyes. C: QQ plot for Intra-ocular pressure in right eyes. D: QQ plot for Intra-ocular pressure in left eyes. E: Bidirectional manhattan plot for Corneal hysteresis in right (top) and left (bottom) eyes. F: Bidirectional manhattan plot for Intra-ocular pressure in right (top) and left (bottom) eyes. G: Locus zoom plot of right eye Corneal hysteresis at the ANAPC1 gene. H: Locus zoom plot of right eye Intra-ocular pressure at the TCF4 gene.*



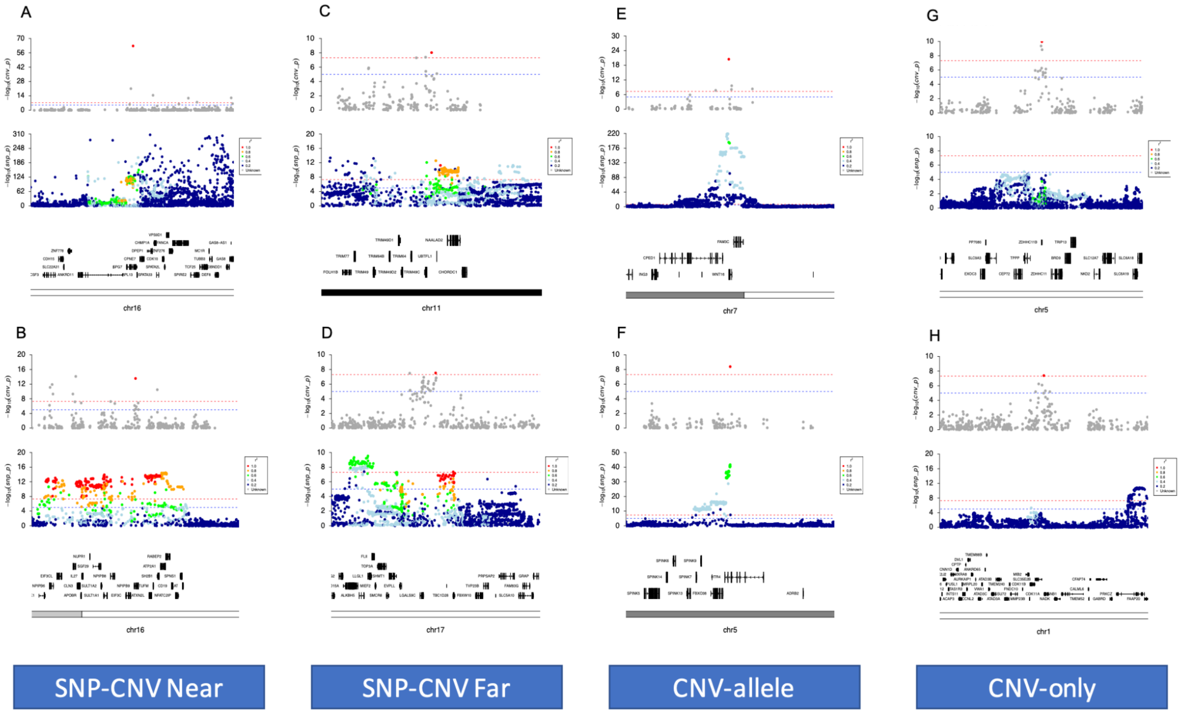
Supplementary Figure 7: *CNV association results for the red blood cell, neurological and behavioural related traits – relating to Figure 2. A: Bidirectional manhattan plots for red blood cell count (top) and reticulocyte count (bottom), fine mapped regions are highlight in orange (red blood cell count) and green (reticulocyte count) with fine mapped regions for reticulocyte count that were also discovered for red blood cell counts not being highlighted. B: Bidirectional manhattan plots for fluid intelligence (top) and reaction time (bottom), fine mapped regions are highlighted in orange and green respectively. C: Bidirectional manhattan plots for alcohol (top) and coffee (bottom) intake, fine mapped regions are highlighted in orange and green respectively. D: Locus zoom plot for red blood cell count at the ABO gene. E: Locus zoom plot for alcohol intake around the NPIPB6 gene. F: Locus zoom plot for reaction time around the ARL17B gene.*



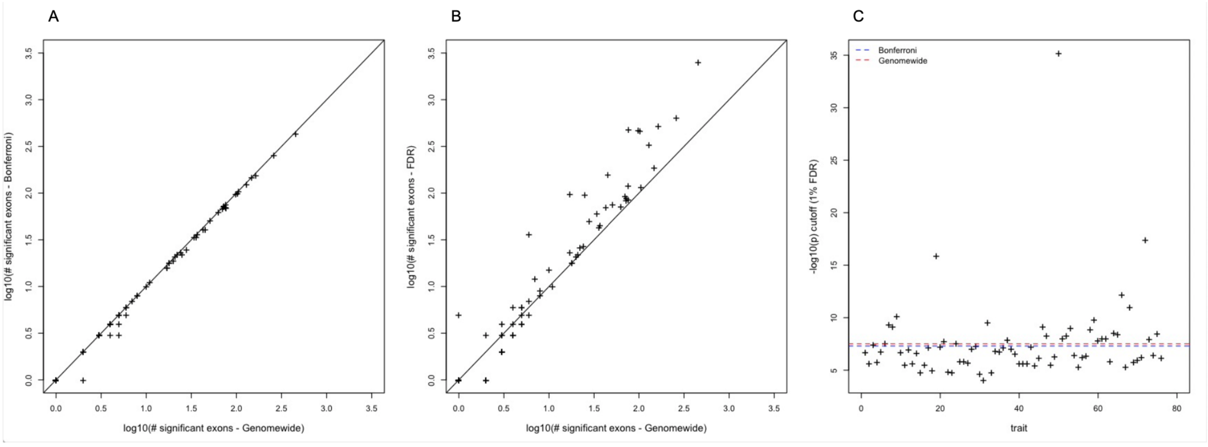
Supplementary Figure 8: *Comparison of association signal strength for heart related ICD10 codes at the LPA gene – relating to Figure 3. A: Overlaid manhattan plot from chromosome 6 including 5 heart related ICD10 code based case/control tests, the colour of points and legend indicate the ICD10 code. B: Minus log10 p values for ICD10 code I25 against code I20. C: Minus log10 p values for ICD10 code I25 against code I21. D: Minus log10 p values for ICD10 code I25 against code I35. E: Minus log10 p values for ICD10 code I25 against code I50.*



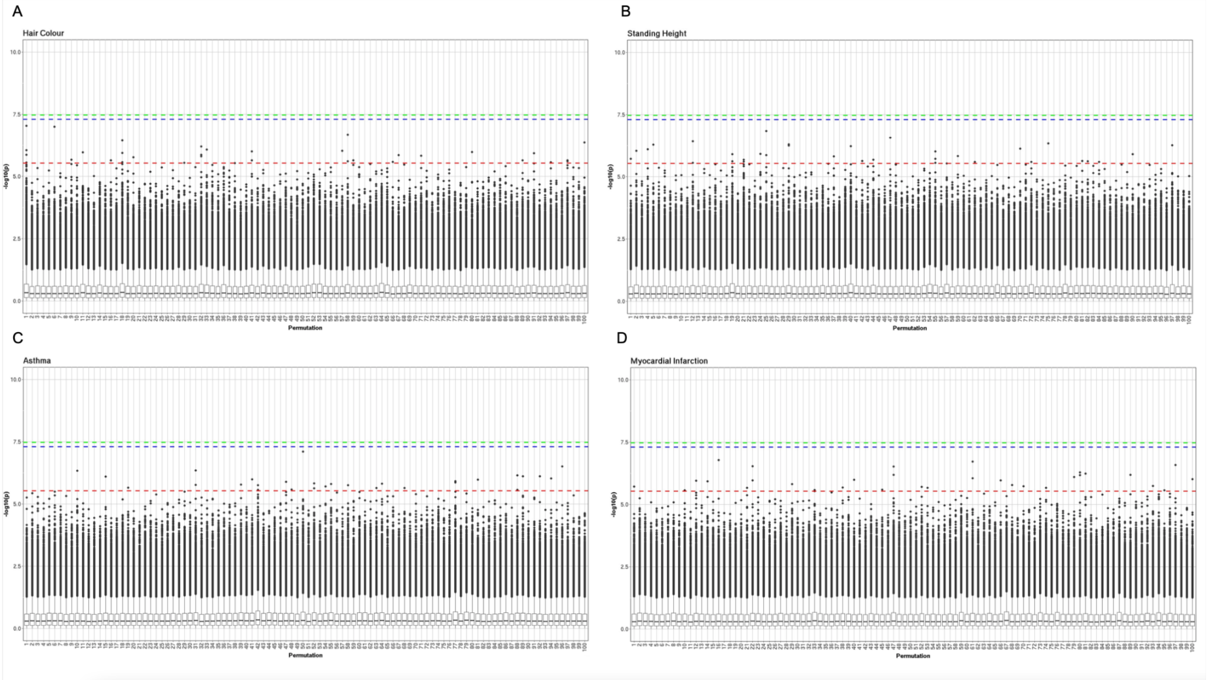
Supplementary Figure 9: *Summary of comparison into CNV association results between 2 SNP based CNV association studies (Auwerx et al*[56] *and Hujoel et al*[57]*) and CNest (Fitzgerald – this study) – relating to STAR Methods (comparison to previous association studies). A: Boxplot plots showing the log10 of the size (number of bases) for all fine mapped association regions with Auwerx in red, Hujoel in blue and Fitzgerald in green. B: The log10 of the size of all association regions against the -log10 p value for each of the 3 studies (Auwerx in red, Hujoel in blue and Fitzgerald in green). C: The -log10 p-value for the most significant exon level signal from CNest (Fitzgerald) against the -log10 p-value from each of the 2 SNP based CNV association studies across the 9 different traits, the color of point and legend indicate the trait and study. D: Stacked barplots showing the number of associations from the 2 SNP base studies and traits that show genomewide (purple), suggestive (orange) or no significant (black) association from exon level CNest association, the y-axis labels indicate the study and trait. E: Association region for reticulocyte count that shows genome wide significance in CNest results. F: Association region for height that shows suggestive significance in CNest results. G: Association region for the FEV/FEC score that shows no significance in CNest results.*



***Supplementary Figure 10**: Locus zoom plots showing some CNV association categories – relating to Figure 4. A: SNP-CNV near discovery for hair colour involving exons 7-10 of the SPG7 gene. B: SNP-CNV near discovery for alcohol consumption at a 0.8MB region containing 4 fine mapped CNV regions involving the NPIPB6, NPIPB7, NPIPB9 and SH2B1 genes. C: SNP-CNV far discovery for hair colour at the TRIM49C with tagging SNPs downstream at UBTFL1 or NAALAD2 genes. D: SNP-CNV far discovery for standing height at a 12.6KB region including the EVPLL and LGALS9C genes. E: CNV-allele discovery for heel bone density at the WNT16 gene. F: CNV-allele discovery for the FEV/FEC ratio involving exon 1 of the HTR4 gene. G: CNV-only association for the FEV/FEC ratio on chromosome 5 at the ZDHHC11B gene. H: CNV-only discovery for standing height including several genes that are pulled up towards suggestive genome wide significance with a single exon signal that passes genome wide significance within the CDK11A gene.*



Supplementary Figure 11: *Comparison of significance level approaches – relating to STAR Methods (definition of the association significance threshold). A: log10 of the number of significant exon level signals per trait using the genome wide 5e-08 cut-off vs. a strict Bonferroni cut-off. B: log10 of the number of significant exon level signals per trait using the genome wide 5e-08 cut-off vs. a 1% FDR cut-off. C: The -log10 significance level at a 1% FDR cut-off for all traits tested. Red dashed line indicates the Genomewide 5e-08 cut-off and the blue dashed line shows the Bonferroni cut-off.*



Supplementary Figure 12: *Permutation tests for the 4 main traits – relating to STAR Methods (definition of the association significance threshold). All tests were performed using a 100 different random ordering of the trait or case labels followed by association testing genomewide. A: Genomewide tests across 100 differently permuted traits for hair colour. B: Genomewide tests across 100 differently permuted traits for standing height. C: Genomewide tests across 100 differently permuted case labels for asthma. D: Genomewide tests across 100 differently permuted case labels for myocardial infarction.*

Supplementary data 1 – relating to Figure 2 (CNwas results on UK Biobank main traits)

Example - Hair Colour:
For hair colour (**Figure 2A**) after fine mapping we detect 30 copy number variable regions that pass genome wide significance, the majority of which (20/30) have been found to be associated with pigmentation by previous SNP GWAS (**Supplementary Table S4**). A particularly strong signal is found in a region containing the *OCA2* and *HERC2* genes with a -log10 P > 230. Both OCA2 and HERC2 are well known genes involved in pigmentation in both humans and mice and have been shown to be involved in iris, skin and hair pigmentation in human[117–120]. Pathogenic variation in *OCA2* including CNVs are known to be a leading cause of the rare genetic disorder oculocutaneous albinism and an increased susceptibility to melanoma[121–123]. The strongest signal and lead exon (**Figure 2E**) around the fine mapped CNV region is found in the hect domain and RCC1-like domain 2 (*HERC2*) which has been well described by multiple studies and has a strong association to pigmentation of the eyes, hair and skin[96,117,124]. Some of the novel associations we discover in relation to hair colour (**Supplementary Table S4**) include signals around 15q11.2, including several of the *GOLGA* genes which have functions relating to membrane traffic and Golgi structure; however, the precise function is unclear. Human chromosome 15 contains multiple copies of the GOLGA core elements close to the evolutionary conserved chromosome 15 low copy repeat (LCR15) duplicons[125] in primates at which several structural rearrangements break points have been described and linked to disorders and structural abnormalities such as Prader-Willi and Angelman syndromes[126]. Additionally, downstream *GOLGA* genes also detected here (*GOLGA8F*, *GOLGA8G* and *GOLGA8M*) have been linked to hair colour in previous SNP GWAS studies and from analysis of the UK Biobank[127].

Example - Standing Height:
After fine mapping we discovered 45 distinct regions associated with standing height encompassing between 1 and 2 genes (**Figure 2B**). The majority of these regions (27/45) contained at least one gene that had previously been associated with height from SNP GWAS (**Supplementary Table S5**) however with little to no evidence of CNV associations. The strongest signal outside of the HLA was seen in a region downstream of *ADAMTSL3* at 15q25.2, including the *UBE2Q2L* and *GOLGA6L4* genes and which is enriched for segmental duplications[128]. There is a single exon signal at exon 15 of the *ADAMTSL3* gene; *ADAMTSL3* has previous evidence from multiple SNP GWAS studies of being associated with height[129–132] and with certain neurological disorders such as schizophrenia[133] and bipolar disorders[134] acting under a proposed alternative splicing mechanism[135]. CNV at *ADAMTSL3* has yet to be described in relation to human height and interestingly this region contains multiple different CNV events varying in size all with strong association signals to height (**Figure 2F**). One for these CNVs overlaps exons 28-30 of *ADAMTSL3* which would likely result in truncation of the PLAC (protease and lacunin) domain[136]. Strong human height CNV association signals are observed in the segmental duplication rich region downstream of *ADAMTSL3* including exons within the *UBE2Q2L* and *GOLGA6L4* genes both of which have been linked to neurological disorders by previous work[137,138] with *UBE2Q2L* also having been specifically linked with human height[92,131]. Recurrent deletions and duplications at 15q25.2 have been described in relation to rare disease including neurological traits[139] however have not yet been described as a hotspot for structural rearrangements associated with common human traits such as height. We also found additional novel regions with no evidence of prior association to height (**Supplementary Table S5**), including one region at the Neuroblastoma BreakPoint Family *NBPF1* gene involving the

highly copy number variable *DUF1220* domain[140] which have be previously associated in a dose-dependent manner with important human traits such as microcephaly and macrocephaly, brain size and neurological disorders[19,141,142].

Example - Asthma:
For Asthma we discover 18 fine mapped CNV regions (**Figure 2C**). Strong CNV association signal was found around a region containing the 3 genes *CHROMR, PRKRA* and *PJVK* and upstream of *TTN* (**Figure 2G**). None of the 3 genes have previous evidence of a link to asthma however both *PRKRA* and *PJVK* have been found to be associated with lung functions such as vital capacity and forced expiratory volume (FEV) in a recent study[91]. In contrast the cholesterol induced regulator of metabolism RNA *CHROMR* has no previous association to asthma and its precise function is poorly understood. Unsurprisingly the strongest signal for asthma is found in the HLA with the lead signal specifically restricted to the *HLA-DQA2* gene which has strong prior associations to asthma and hay fever all based on intergenic SNPs from multiple SNP based GWAS investigations[143–147]. Some of our novel CNV associations for asthma (**Supplementary Table S6**) include signals in genes, *TAP2* and *STARD3NL*, which have not been linked with asthma by previous SNP GWAS studies but have evidence of association to certain other respiratory diseases[148,149]. A recent study using a different approach for CNV association testing from WES in the UKBB and specifically focussed on Asthma have rediscovered many of the associations we have made here[103].

Example - myocardial infarction:
For acute myocardial infarction (MI) we discover 26 fine mapped associations (**Figure 2D**). The strongest signal for MI is found within the *LPA* gene and this signal is found consistently across most heart related traits that we have tested in the UK Biobank. The *LPA* gene encodes a substantial portion of lipoprotein(a) and has been linked to numerous heart related diseases including coronary artery disease (CAD), aortic atherosclerosis and MI[150–152]. Changes in dosage of the *LPA* gene, specifically the KIV-2 copy number alteration, has been previously linked to changes in lipoprotein(a) levels and a modified risk of heart disease (CAD)[153–156]. This analysis of CNV association across a large cohort provides important additional information and allows a detailed estimate of the effect size for differences in *LPA* copy number in relation to the risk of MI. Interestingly we detect multiple different sized CNV events that hit the *LPA* gene (**Figure 2G**) but also include other coding regions with the lead exonic signal always restricted solely to the *LPA* gene. To our knowledge, only 2/26 fine mapped CNV associations (*LPA* and *BMP1*) have a direct association to MI from previous SNP GWAS testing[151,157] however a large fraction of the remaining regions have prior associations to other important heart related traits or cardiac disease risk factors (**Supplementary Table S7**). For example, the *TM2D1* gene that has prior association to electrocardiography[158] and the structure of the left cardiac ventricle[159]; the *DPP6* gene that has been associated with multiple heart related phenotypes including sudden cardiac arrest[160]; and genes associated with blood lipid level measurements such as *LCAT* and *RCAN1*[161,162].

Supplementary data 2 – relating to Figure 3 (CNwas results on UK Biobank ICD10 first occurrences fields)

E*xample - E80:* For disorders of porphyrin and bilirubin metabolism, we found multiple strong signals involving specific exons across UDP-glucuronosyltransferase genes (*UGT1A10*, 9, 8, 7, 6 and 4) (**Figure 3E**). Genetic variation of *UGT1A* genes has been associated with disorders of bilirubin metabolism including Gilbert's syndrome by multiple previous SNP GWAS studies[163,164] with, for example, very strong association signal at *UGT1A10* for the intron variant rs6742078 (2_233763993_G_T)[165]. This specific SNP has also been linked to other lipid metabolism disorders such as Gallstones Disease (GSD)[166] and although studies looking at CNV burden analysis of lipid metabolism genes have shown a significant enrichment in GSD cases none of those associations could be attributed to any single gene[167]. Here we provide novel CNV associations at *UGT1A* genes with a direct link to bilirubin metabolism that could be an important risk factor for several lipid metabolism related disorders.

Example - D50: For iron deficiency anaemia we discovered two significantly associated loci on chromosome 7 (**Figure 3F**) one of which covers exons 4-6 of the cationic trypsinogene gene *PRSS1* that has been linked to chronic pancreatitis by multiple studies[168,169]. Autosomal dominant mutations in *PRSS1* are thought to be a leading cause of hereditary pancreatitis, a rare condition that results in recurrent inflammation of the pancreas, and an increased risk of pancreatic cancer[170]. As such *PRSS1* is regularly tested in patients with suspected hereditary pancreatitis[171] however the *PRSS1* gene contains multiple known variants, including copy number changes, often with unknown clinical importance[172]. Iron metabolism and pancreatic function are closely related processes[173] with evidence that pancreatic enzyme levels influence the efficiency of iron absorption[174]. Here we provide a link between the copy number at exons 4-6 of the *PRSS1* gene with the ICD10 code D50 relating to iron deficiency anaemia that may be a result of pancreatic dysfunction.

Example - M10: For Musculoskeletal disorders we discovered 17 fine mapped association loci across 4 different traits including one location at 4p16.1 at exon 3 of the *SLC2A9* gene that was associated with ICD10 code M10: gout (**Figure 3G**). Gout is a swelling of joints, normally in the feet, that is caused by hyperuricemia (an excess of uric acid in the blood) with mutations at *SLC2A9* having been found to be associated with serum urate concentrations and the onset of gout[175,176]. A non coding CNV near *SLC2A9* (integenic and approximately 200 kb upstream of the *SLC2A9* gene) has been described in association with serum uric acid levels[177] however CNVs in coding regions of the *SLC2A9* have not yet been discovered in relation to uric acid level or with a direct association to gout. Here we provide a novel CNV association result at exon 3 of the *SLC2A9* gene with a direct association to gout from the UK Biobank.

Example - O36: For Pregnancy childbirth and the puerperium we discovered fine mapped CNV associations against code O36: maternal care for known or suspected foetal problems at 1p36.11 including the *RHD* and *RHCE* genes (**Figure 3H**). Variation and *RHD* gene deletion in the human Rh blood group system has been extensively studied in relation to pregnancy risk[178] where prior to the development of medical treatments, Rh-negative (D-negative) mothers were at significant risk of haemolytic disease of the newborn (HDN). It is still unclear what potential benefit the *RHD* gene deletion may have that merits its relatively high frequency in the human population[179]. Blood tests are normally carried out in D-negative expectant mothers to determine the Rh factor status of the child and direct treatment if using anti-D injection is required[180]. However variation in the less well understood Rh C and E alleles of *RHCE* is clinically relevant, influences the risk of HDN[181] and this association discovered in this study merits further investigation for this well understood risk factor for pregnancy.

Example - K74: For fibrosis and cirrhosis of liver we discovered a single CNV association at exon 3 of the *PNPLA3* gene (**Figure 3I**). Cirrhosis of the liver is a disorder in which the liver parenchyma is replaced with fibrous tissue and is often caused by alcoholism as well as hepatitis B and C infection[182,183]. The *PNPLA3* gene has been found to be associated with liver cirrhosis by multiple SNP GWAS studies[184,185] and although CNVs at *PNPLA3* has not been well described or linked to Cirrhosis in the past it has been shown that transcriptional regulation of *PNPLA3* has an impact on liver disease with higher levels of *PNPLA3* mRNA in the cytoplasm being negatively associated with the severity of alcoholic fatty liver disease (NAFLD) in humans[186].

Example - heart related ICD10 codes: Across 5 heart related ICD10 codes (I20, I21, I25, I35 and I50) we found strong CNV association signals at the *LPA* gene with the exception of I50: heart failure (**Supplementary Figure S8).** When comparing signal strength between the 5 heart related ICD10 codes we observe a clear sample size effect with the tests showing the stronger signals tending to have larger number of cases (**Supplementary Figure S8 B-E**). The five ICD10 codes included were I25: chronic ischaemic heart disease (20,503 cases), I20: angina pectoris (10,117 cases), I21: acute myocardial infarction (3,698 cases), I35: nonheumatic aortic valve disorders (1,692 cases) and I50: heart failure (3,557 cases). Although there is heterogeneity between the effect sizes for these different ICD codes, ranging from -0.39 for I25 to -1.41 for I35, they are similar for the 3 codes I20, I25 and I50 (-0.42, -0.39 and -0.38 respectively), suggesting that this association may become significant for I50 with increased sample sizes. CNV association at *LPA* is a major feature in all heart related phenotypes we have tested in the UK Biobank providing further evidence that changes in dosage of *LPA* is a significant risk factor for heart disease in humans.

Supplementary data 3 – relating to Figure 4 (combined CNV and SNP based associations)

For standing height we show clear complementary signals at the ACAN gene for SNP based and CNV based association results (**Figure 4A**), an example of the SNP-CNV near class with the CNV signal being well tagged by 30 SNPs within the gene body between exons 6 to 12. The CNV signal is restricted to exon 12 that encodes the chondroitin sulfate attachment (CS) domain[187] which is important for aggregation with hyaluronan resulting in a strong negative charge that gives rise to load-bearing properties of cartilage[188]. Mutations in *ACAN* have been studied in relation to both syndromic and nonsyndromic human traits with a number of different impactful variants having been discovered[130,189,190]. Earlier studies found *ACAN* to be a strong candidate for autosomal dominant disorders such as spondyloepiphyseal dysplasia Kimberley type (SEDK) and early-onset osteoarthritis (OA) from genetic linkage analysis and mouse models of chondrodysplasia[187] however, heterozygous mutations in *ACAN* display highly variable nonsyndromic phenotypes including short stature, early onset osteoarthritis and mild dysmorphic features[191]. In this case, although the CNV is both well tagged and close to the SNP associations, the CNV directly suggests the functional variant of these SNPs is the deletion of this exon implying that haploinsufficiency of the *ACAN* gene is the main mechanism underlying these associations.

A *SNP-CNV far* example is shown in **Figure 4B**, being a 60KB region on chromosome 6 that is associated with the lung function measure FEV/FEC ratio including the *STK19, C4A, C4B* and *CYP21A2* genes with the lead exon CNV signal encompassing exons 26-30 of the *C4A* gene (**Figure 4B**). Interestingly, there are 208 tagging SNPs that pass genome wide significance for FEV/FEC ratio however none of these SNPs are located closest to either of the *C4* genes (*C4A* or *C4B).* The C4 genes encode an important part of the immune complement system and deficiencies (including CNV) at *C4* genes have been strongly associated with immune disorders such as Systemic Lupus Erythematosus[192,193]. The *C4A* and *C4B* genes encode different components of the highly polymorphic C4 complement protein and can be distinguished from each other by four specific amino acids at positions 1101–1106[194]. Due to high sequence similarity the total copy number of *C4* can be defined as the sum between *C4A* and *C4B*[195], however both *C4A* and *C4B* are multiallelic CNV locations displaying common differences in copy number with *C4A* ranging between 0 to 5 and *C4B* between 0 to 4 copies[196]. The CNVs at *C4* has not been linked previously with lung function; however, a recent study into chronic obstructive pulmonary disease (COPD) in the Korea Associated Resource cohort has investigated genome wide SNP interactions mapped to *C4B* in relation COPD and the FEV/FEC ratio measure[197]. We provide new evidence for the role of *C4* CNV in lung function as measured by the FEV/FEC ratio in the UK Biobank.

SUPPLEMENTAL REFERENCES

117. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., et al. (2008). Two newly identified genetic determinants of pigmentation in Europeans. Nat. Genet. *40*, 835-837.

118. Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., Wang, L.-E., Wei, Q., Lee, J.E., Amos, C.I., et al. (2013). Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum. Mol. Genet. *22*, 2948-2959.

119. Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., Zhao, Z.Z., et al. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet. *4*, e1000074.

120. Lin, B.D., Willemsen, G., Abdellaoui, A., Bartels, M., Ehli, E.A., Davies, G.E., Boomsma, D.I., and Hottenga, J.J. (2016). The genetic overlap between hair and eye color. Twin Res. Hum. Genet. *19*, 595-599.

121. Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe’er, I., and Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. *6*, e1000993.

122. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393.

123. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173-1186.

124. He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J.-Y., Johnson, T.A., Li, H., Yang, X., Dai, J., et al. (2015). Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Hum. Mol. Genet. *24*, 1791-1800.

125. Tachmazidou, I., Süveges D., Min, J.L., Ritchie, G.R.S., Steinberg, J., Walter, K., Iotchkova, V., Schwartzentruber, J., Huang, J., Memari, Y., et al. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. Am. J. Hum. Genet. *100*, 865-884.

126. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

127. Gudbjartsson, D.F., Walters, G.B., Thorleifsson, G., Stefansson, H., Halldorsson, B.V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., et al. (2008). Many sequence variants affecting diversity of adult human height. Nat. Genet. *40*, 609-615.

128. Rüeger S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. *14,* e1007371,

129. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570,* 514-518.

130. Nagy, R., Boutin, T.S., Marten, J., Huffman, J.E., Kerr, S.M., Campbell, A., Evenden, L., Gibson, J., Amador, C., Howard, D.M., et al. (2017). Exploration of haplotype research consortium imputation for genome-wide association studies in 20, 032 Generation Scotland participants. Genome Med. *9*, 23.

131. Ramasamy, A., Kuokkanen, M., Vedantam, S., Gajdos, Z.K., Couto Alves, A., Lyon, H.N., Ferreira, M.A.R., Strachan, D.P., Zhao, J.H., Abramson, M.J., et al. (2012). Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA. PLoS One *7,* e44008.

132. Shrine, N., Portelli, M.A., John, C., Soler Artigas, M., Bennett, N., Hall, R., Lewis, J., Henry, A.P., Billington, C.K., Ahmad, A., et al. (2019). Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. Lancet Respir. Med. *7,* 20-34.

133. Pividori, M., Schoettler, N., Nicolae, D.L., Ober, C., and Im, H.K. (2019). Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. Lancet Respir. Med. *7,* 509-522.

134. Zhu, Z., Zhu, X., Liu, C.-L., Shi, H., Shen, S., Yang, Y., Hasegawa, K., Camargo, C.A., Jr, and Liang, L. (2019). Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. Eur. Respir. J. *54,* 1901507.

135. Klimentidis, Y.C., Raichlen, D.A., Bea, J., Garcia, D.O., Wineinger, N.E., Mandarino, L.J., Alexander, G.E., Chen, Z., and Going, S.B. (2018). Genome-wide association study of habitual physical activity in over 377, 000 UK Biobank participants identifies multiple variants including CADM2 and APOE. Int. J. Obes. *42,* 1161-1176.

136. Hartiala, J.A., Han, Y., Jia, Q., Hilser, J.R., Huang, P., Gukasyan, J., Schwartzman, W.S., Cai, Z., Biswas, S., Trégouët D.A., et al. (2021). Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. Eur. Heart J. *42,* 919-933.

137. Richardson, T.G., Sanderson, E., Elsworth, B., Tilling, K., and Davey Smith, G. (2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. BMJ *369,* m1203.