ORIGINAL ARTICLE

Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank

Karen Crawford, Matthew Bracher-Smith, David Owen, Kimberley M Kendall, Elliott Rees, Antonio F Pardiñas, Mark Einon, Valentina Escott-Price, James T R Walters, Michael C O'Donovan, Michael J Owen, George Kirov

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2018-105477).

MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

Correspondence to
Dr. George Kirov, MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Hadyn Ellis Building,Maindy Road, Cardiff CF14 4XN, UK; kirov@cardiff.ac.uk

Received 7 May 2018
Revised 31 August 2018
Accepted 25 September 2018
Published Online First 20 October 2018

To cite: Crawford K, Bracher-Smith M, Owen D, *et al*. *J Med Genet* 2019;**56**:131–138.

ABSTRACT

Background Genomic CNVs increase the risk for early-onset neurodevelopmental disorders, but their impact on medical outcomes in later life is still poorly understood. The UK Biobank allows us to study the medical consequences of CNVs in middle and old age in half a million well-phenotyped adults.

Methods We analysed all Biobank participants for the presence of 54 CNVs associated with genomic disorders or clinical phenotypes, including their reciprocal deletions or duplications. After array quality control and exclusion of first-degree relatives, we compared 381 452 participants of white British or Irish origin who carried no CNVs with carriers of each of the 54 CNVs (ranging from 5 to 2843 persons). We used logistic regression analysis to estimate the risk of developing 58 common medical phenotypes (3132 comparisons).

Results and conclusions Many of the CNVs have profound effects on medical health and mortality, even in people who have largely escaped early neurodevelopmental outcomes. Forty-six CNV–phenotype associations were significant at a false discovery rate threshold of 0.1, all in the direction of increased risk. Known medical consequences of CNVs were confirmed, but most identified associations are novel. Deletions at 16p11.2 and 16p12.1 had the largest numbers of significantly associated phenotypes (seven each). Diabetes, hypertension, obesity and renal failure were affected by the highest numbers of CNVs. Our work should inform clinicians in planning and managing the medical care of CNV carriers.

INTRODUCTION

Genomic CNVs are structural alterations to chromosomes of >1000 bases in length that can intersect multiple genes.[1] Specific CNVs have been shown to increase risk for autism spectrum disorders,[2] developmental delay and other neurodevelopmental disorders,[3] and schizophrenia.[4] Apart from their association with neurodevelopmental and psychiatric outcomes, these CNVs can lead to medical disorders. Several CNVs, for example, deletions at 22q11.2,[5] have been extensively studied on hundreds of carriers and their medical consequences are well established. However, for CNVs with lower penetrance, very rare CNVs or several reciprocal deletions/duplications of known genomic disorders, the associated medical phenotypes have not been identified. Moreover, most research has been performed on children and young people referred to genetic clinics,[3 6] creating a strong referral bias towards recording high rates of developmental delay, early-onset medical conditions and more adverse outcomes. Most CNVs display incomplete penetrance,[7] resulting in apparently unaffected adult carriers in the general population. The rate of medical outcomes in later life of CNV carriers, or in the general population as a whole, has not been addressed in adequately powered studies to date.

The establishment of the UK Biobank presents a unique opportunity to examine the spectrum of medical outcomes of CNVs in middle-aged and old-aged people, as all half a million participants have been assessed with identical methods and blindly to their CNV status. The Biobank collects longitudinal data from hospital admissions, self-report, death certificates, cancer registries and primary care (general practitioners') records. Here, we report on the medical consequences of carrier status for 54 CNVs that are recognised as associated with clinical phenotypes or genomic disorders,[3 6 8] including their reciprocal deletions/duplications.

METHODS

Participants

The UK Biobank recruited just over half a million people from the general population of the UK, using National Health Service patient registers, with no exclusion criteria. Participants have consented to provide personal and health information, urine, saliva and blood samples, and to have their DNA tested. We obtained approval from the UK Biobank to analyse the CNVs in project 14421: 'Identifying the spectrum of biomedical traits in adults with pathogenic copy number variants (CNVs)'.

Participants were between 40 and 69 years of age at the time of recruitment between 2006 and 2010. As the lifetime prevalence of disorders often varies by ancestry, we restricted the analysis to those participants who declared themselves as 'white British or Irish': 421 268 participants who passed our genotyping quality control (QC) filters (CNV calling). After exclusion of first-degree relatives, 396 725 subjects were retained for analysis, 53.8% of whom were female. The mean age at the end of the current follow-up interval for medical outcomes (in 2016) was 64.7 years, SD=8.0 years.

CNV calling

Samples were genotyped at the Affymetrix Research Services Laboratory, Santa Clara, California, USA, on two arrays with 95% common content between them: around 50 000 samples were genotyped on the UK BiLEVE Array (807 411 probes) and the remaining samples on the UK Biobank Axiom Array (820 967 probes).[9] We downloaded the anonymised genotypic data from the UK Biobank as 488 415 raw (CEL) files and analysed them with the methods we reported previously.[10] Briefly, we generated normalised signal intensity data, genotype calls and confidences, using ~750 000 biallelic markers. These were then processed with PennCNV-Affy software.[11] Individual samples were excluded if they had >30 CNVs, a waviness factor >0.03 or <−0.03, a call rate <96% or log R ratio SD >0.35. A total of 25 069 files were excluded after this QC (5.1%). Individual CNVs were excluded if they were covered by <10 probes or had a density coverage of less than one probe per 20 000 base pairs.

Choice of CNVs

We compiled a list of 92 CNVs in 47 genomic locations from two widely accepted sources that proposed largely overlapping sets of CNVs (online supplementary table 1 in supplementary material).[3 6] The authors of these studies used information from databases, reviews and publications to produce lists of CNV regions that lead to genomic disorders, congenital malformations, neurodevelopmental or other clinical phenotypes. We refer to this set of 92 CNVs as 'pathogenic', consistent with the criteria proposed by the American College of Medical Genetics standards which describe as pathogenic those CNVs that have been documented as clinically significant in multiple peer-reviewed publications, even if penetrance and expressivity of the CNV are known to be variable.[12] Many (but not all) have been shown to statistically increase the risk for developmental delay.[3] Online supplementary table 1 lists the sources for selection and our criteria for inclusion in analysis. Several overlapping or adjacent CNVs listed as separate loci in the original publications were grouped together (eg, the 'small' and the 'common' 22q11.2 or the 'small' and the 'large' 16p13.11 deletions/duplications). As a rule, the reciprocal deletions/duplications of known genomic disorders were also included by the above authors and by us, in order to examine their medical consequences, even if the evidence for their pathogenicity has not been established.

The criteria for calling CNVs that do not span the full critical region are given in online supplementary table 2. As a rule, a CNV had to intersect at least 50% of the critical region, marked as 'Location (hg19)', and intersect the relevant candidate genes, if known. For single gene CNVs, we required deletions to intersect at least one exon, and duplications to span the whole coding region, as the functional consequences of partial gene duplications can be unpredictable, while deletions of any part of the coding sequence of a gene are likely to act as loss-of-function mutations. We observed several loci, mostly telomeric, where a number of small CNVs were preferentially called on arrays that failed QC (marked 'Unreliable' in online supplementary table 1). We excluded these loci from analysis in order to avoid potential false-positives on this genotyping platform. We also excluded from analysis CNVs with fewer than five observations in the full sample, as being too rare for statistical analysis (marked 'Rare' in online supplementary table 1). The above filtering left 54 CNVs for analysis (table 1).

Choice of medical phenotypes

Data on health outcomes were collected from several sources. Self-declared illnesses were disclosed by participants at their initial assessments and coded into 445 distinct categories. Hospital discharge diagnoses (primary and secondary) and death certificates contain over 11 000 International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) codes assigned to at least one participant. Analysing each individual code separately against 54 CNV loci would result in small numbers of participants with each code and fail to provide the statistical power needed to detect true associations. To reduce the dimensionality of the data and therefore increase power and provide more meaningful results, we grouped together discrete disease entities into broader disease groups. A participant was coded as a 'case' if he/she had a relevant diagnosis on at least one occasion, in any of the above sources of information. We gave preference to common conditions and grouped disorders into recognised categories, based on organ, system or aetiology, while excluding from the current analysis infectious diseases, injuries and neuropsychiatric disorders (the latter being analysed separately). The disease codes used to construct each phenotype group are listed in online supplementary table 3. For myocardial infarction and stroke, we used the 'adjudicated' data provided by the UK Biobank (data fields 42 000 to 42 013). Phenotype groups found in fewer than 2000 participants were not included. The final list of disease groups contains 58 entities, including 'death during follow-up' obtained from the death registries. Data on cancer were taken only from the UK cancer registries, as collected and supplied by the UK Biobank, as this is the most reliable and complete resource for cancers in the UK. For the current work we considered all malignant cancers as a single phenotype. As risk for cancer was not significantly affected by CNVs as a group, and because most individual cancers affected relatively small numbers of patients, we did not analyse the cancers further by subtype.

Statistical analysis

Analyses were performed in the statistical package R (version 3.3.2) using a Linux server. We examined the effect of the presence of a CNV on each medical phenotype with logistic regression analysis. As covariates, we used age, gender, array type (Axiom/BiLEVE), Townsend deprivation index (as a measure of the socioeconomic status) and the first 15 principal components from the genetic analysis, as provided by the UK Biobank. We used Firth's bias-reduced logistic regression method,[13] with the R library 'logistf', as it better handles cells with small numbers. We report the resulting p-values, ORs and 95% CIs for the ORs. We also report the uncorrected relative risk (RR), for having the phenotype in carriers of a specific CNV and non-carriers of any of the 54 CNVs. (RR is used for the additional images on our website (http://kirov.psycm.cf.ac.uk/), as it returns the more intuitive value of zero for associations with zero CNVs in cases.) Conservative Bonferroni correction for the testing of 54 CNVs×58 phenotypes gives a $p<1.6\times10^{-5}$ as a project-wide significance level. As many true-positive associations were expected, it is more appropriate to use the Benjamini-Hochberg false discovery rate (B-H FDR) for correction of p-values.[14] Our preferred B-H FDR is 0.1.

RESULTS AND DISCUSSION

Quality control

The Affymetrix arrays produced reliable calls for the 54 CNVs. This is not surprising, given the large size and good probe coverage of these CNVs. This impression is confirmed by the remarkably similar CNVs frequencies, compared with those reported by us in previous control populations (online supplementary table 4 and supplementary figure 1). There were no

Table 1 List of 54 CNVs analysed in this study

CNV locus	Location (hg19)	Genes (n)	Carriers, n (%)	N sign. FDR=0.1
TAR_del	chr1:145,39–145,81	17	75 (0.018)	0
TAR_dup	chr1:145,39–145,81	17	436 (0.1)	1
1q21.1del	chr1:146,53–147,39	9	113 (0.027)	2
1q21.1dup	chr1:146,53–147,39	9	177 (0.042)	1
*NRXN1*_del	chr2:50,14–51,26	1	163 (0.039)	1
2q11.2del	chr2:96,74–97,68	22	31 (0.007)	0
2q11.2dup	chr2:96,74–97,68	22	29 (0.007)	0
2q13del(*NPHP1*)	chr2:110,86–110,98	3	2448 (0.58)	0
2q13dup(*NPHP1*)	chr2:110,86–110,98	3	1976 (0.47)	0
2q13del	chr2:111,39–112,01	3	53 (0.013)	0
2q13dup	chr2:111,39–112,01	3	71 (0.017)	1
2q21.1del	chr2:131,48–131,93	5	41 (0.01)	0
2q21.1dup	chr2:131,48–131,93	5	59 (0.014)	0
3q29del	chr3:195,72–197,35	28	9 (0.002)	0
3q29dup	chr3:195,72–197,35	28	5 (0.001)	6
WBS_dup	chr7:72,74–74,14	26	14 (0.003)	0
7q11.23dup_distal	chr7:75,14–76,06	16	24 (0.006)	0
8p23.1dup	chr8:8,10–11,87	35	6 (0.001)	0
10q11.21q11.23del	chr10:49,39–51,06	19	57 (0.014)	0
10q11.21q11.23dup	chr10:49,39–51,06	19	43 (0.01)	0
10q23dup	chr10:82,05–88,93	29	7 (0.002)	0
13q12del(*CRYL1*)	chr13:20,98–21,10	2	379 (0.09)	0
13q12dup(*CRYL1*)	chr13:20,98–21,10	2	10 (0.002)	0
13q12.12del	chr13:23,56–24,88	10	85 (0.02)	0
13q12.12dup	chr13:23,56–24,88	10	236 (0.056)	0
15q11.2del	chr15:22,81–23,09	5	1664 (0.39)	0
15q11.2dup	chr15:22,81–23,09	5	2041 (0.48)	0
PWS_dup	chr15:23,68–28,39	116	19 (0.005)	0
15q11q13del_BP3-BP4(*APBA2, TJP*)	chr15:29,16–30,38	4	16 (0.004)	1
15q11q13dup_BP3-BP4(*APBA2, TJP*)	chr15:29,16–30,38	4	53 (0.013)	0
15q11q13dup_BP3-BP5	chr15:29,16–32,46	17	9 (0.002)	0
15q13.3del	chr15:31,08–32,46	8	42 (0.01)	2
15q13.3dup	chr15:31,08–32,46	8	240 (0.057)	0
15q13.3del(*CHRNA7*)	chr15:32,02–32,46	1	10 (0.002)	0
15q13.3dup(*CHRNA7*)	chr15:32,02–32,46	1	3031 (0.72)	0
15q24dup	chr15:72,90–78,15	77	9 (0.002)	0
16p13.11del	chr16:15,51–16,29	7	131 (0.031)	1
16p13.11dup	chr16:15,51–16,29	7	828 (0.2)	2
16p12.1del	chr16:21,95–22,43	8	246 (0.058)	7
16p12.1dup	chr16:21,95–22,43	8	202 (0.048)	0
16p11.2distal_del	chr16:28,82–29,05	11	58 (0.014)	3
16p11.2distal_dup	chr16:28,82–29,05	11	137 (0.033)	0
16p11.2del	chr16:29,65–30,20	30	110 (0.026)	7
16p11.2dup	chr16:29,65–30,20	30	138 (0.033)	2
17p12del(HNPP)	chr17:14,14–15,43	8	237 (0.056)	1
17p12dup(CMT1A)	chr17:14,14–15,43	8	124 (0.029)	3
Potocki-Lupski syndrome	chr17:16,81–20,21	59	5 (0.001)	0
17q11.2del(*NF1*)	chr17:29,12–30,27	19	9 (0.002)	0
17q12del	chr17:34,81–36,22	17	9 (0.002)	2
17q12dup	chr17:34,81–36,22	17	101 (0.024)	1
22q11.2del	chr22:19,04–21,47	61	10 (0.0024)	0
22q11.2dup	chr22:19,04–21,47	61	280 (0.066)	2
22q11.2distal_del	chr22:21,92–23,65	26	5 (0.001)	1
22q11.2distal_dup	chr22:21,92–23,65	26	13 (0.003)	0

The column 'N sign. FDR=0.1' shows the number of significant associations between the CNV and medical phenotypes at a threshold of FDR=0.1. Further details are given in online supplementary table 1. First-degree relatives are included in the numbers of carriers.
FDR, false discovery rate.

apparent batch effects affecting the calls: the distribution of each CNV in the 106 batches produced no outliers from the expected Poisson distribution, after taking into account the multiple testing for 54 CNVs (online supplementary table 5).

The best confirmation of the data quality would be the identification of well-known phenotypes associated with specific CNVs. This was indeed the case (table 2), as we identified, for example, the known associations of neuropathies and 17p12 deletions/

Table 2 CNV/Phenotype associations significant at FDR=0.1

CNV	Phenotype	No of controls (no of CNV carriers)	No of cases (no of CNV carriers)	Expected no of CNVs in cases	P-values	P-values B-H FDR	OR (95% CI)	Known finding
TAR dup	Obesity	372 000 (385)	9860 (23)	10.2	0.00054	0.047	2.3 (1.5 to 3.4)	
1q21.1 del	Heart failure	376 477 (100)	5081 (6)	1.3	0.0011	0.083	5.3 (2.1 to 11.2)	Yes
1q21.1 del	Cataract	359 694 (92)	21 864 (14)	5.6	0.00047	0.042	3.2 (1.7 to 5.6)	Yes
1q21.1 dup	Diabetes, type 2	360 864 (146)	20 756 (22)	8.4	0.00017	0.025	2.7 (1.6 to 4.1)	
NRXN1 del	Aneurysm	379 638 (152)	1971 (5)	0.8	0.00042	0.041	7.6 (2.8 to 16.4)	
2q13 dup	Diabetes, type 2	360 778 (60)	20 745 (11)	3.5	0.0012	0.094	3.4 (1.7 to 6.3)	
3q29 dup	Any cancer	331 757 (1)	49 700 (4)	0.15	6.23×10^{-5}	0.011	37.5 (6.5 to 389.1)	
3q29 dup	Diverticular disease intestine	354 321 (2)	27 136 (3)	0.2	0.0001	0.017	41.8 (7.4 to 276.0)	
3q29 dup	Inflammatory bowel disease	362 864 (2)	18 593 (3)	0.1	0.00013	0.021	35.5 (6.7 to 217.8)	
3q29 dup	Renal failure	373 535 (3)	7922 (2)	0.1	0.00022	0.027	58.4 (9.2 to 324.8)	
3q29 dup	Death	370 486 (3)	10 971 (2)	0.1	0.0013	0.093	27.8 (4.5 to 146.0)	
15q11q13 del BP3-BP4	Gastric reflux	347 165 (8)	34 301 (6)	0.8	0.00018	0.025	9.1 (3.1 to 25.4)	
15q13.3 del	Diabetes, type 2	360 746 (28)	20 743 (9)	1.6	0.00038	0.039	4.9 (2.2 to 10.2)	
15q13.3 del	Asthma	332 151 (23)	49 338 (14)	3.4	0.00018	0.026	3.9 (2.0 to 7.4)	
16p13.11 del	Obesity	371 729 (114)	9847 (10)	3.0	0.0013	0.096	3.4 (1.7 to 6.2)	
16p13.11 dup	Hypertension	261 304 (483)	120 931 (300)	223.5	2.05×10^{-5}	0.0043	1.4 (1.2 to 1.6)	
16p13.11 dup	Death	371 226 (743)	11 009 (40)	22.0	0.00097	0.080	1.8 (1.3 to 2.4)	
16p12.1 del	Obesity	371 827 (212)	9860 (23)	5.6	1.11×10^{-7}	4.95×10^{-5}	4.0 (2.5 to 6.0)	
16p12.1 del	Hypertension	260 945 (124)	120 742 (111)	57.4	8.64×10^{-8}	5.41×10^{-5}	2.1 (1.6 to 2.8)	
16p12.1 del	Renal failure	373 750 (218)	7937 (17)	4.6	7.23×10^{-6}	0.0021	3.8 (2.3 to 6.1)	
16p12.1 del	Diabetes, type 2	360 926 (208)	20 761 (27)	12.0	0.00021	0.028	2.3 (1.5 to 3.5)	
16p12.1 del	Heart other	369 848 (217)	11 839 (18)	6.9	0.00029	0.034	2.8 (1.7 to 4.4)	Yes
16p12.1 del	Ureter/bladder	333 111 (186)	48 576 (49)	27.1	0.00033	0.036	1.9 (1.3 to 2.5)	
16p12.1 del	Respiratory	360 911 (209)	20 776 (26)	12.0	0.00069	0.059	2.2 (1.4 to 3.2)	
16p11.2 distal del	Gout	374 410 (48)	7096 (6)	0.9	0.00046	0.042	6.5 (2.5 to 14.4)	
16p11.2 distal del	Obesity	371 660 (45)	9846 (9)	1.2	1.14×10^{-5}	0.0028	7.1 (3.3 to 13.7)	Yes
16p11.2 distal del	Diabetes, type 2	360 757 (39)	20 749 (15)	2.2	8.86×10^{-8}	4.63×10^{-5}	7.0 (3.7 to 12.6)	
16p11.2 del	Diabetes, type 2	360 794 (76)	20 761 (27)	4.4	2.54×10^{-11}	3.98×10^{-8}	6.1 (3.8 to 9.5)	Secondary
16p11.2 del	Obesity	371 699 (84)	9856 (19)	2.2	7.39×10^{-10}	7.71×10^{-7}	6.8 (4.0 to 11.0)	Yes
16p11.2 del	Anaemia	362 396 (84)	19 159 (19)	4.4	2.15×10^{-6}	0.00075	4.0 (2.4 to 6.5)	
16p11.2 del	Asthma	332 199 (71)	49 356 (32)	10.5	1.33×10^{-5}	0.0030	2.7 (1.8 to 4.1)	
16p11.2 del	Renal failure	373 625 (93)	7930 (10)	2.0	6.04×10−5	0.012	5.1 (2.5 to 9.5)	
16p11.2 del	Hypertension	260 873 (52)	120 682 (51)	24.1	9.44×10^{-6}	0.0025	2.6 (1.7 to 3.8)	Secondary
16p11.2 del	Osteoarthritis	312 820 (73)	68 735 (30)	16.0	0.00031	0.035	2.4 (1.5 to 3.6)	Secondary
16p11.2 dup	Irritable bowel syndrome	368 567 (118)	13 016 (13)	4.2	0.00036	0.037	3.3 (1.8 to 5.7)	
16p11.2 dup	Sciatica	338 444 (103)	43 139 (28)	13.1	0.001	0.083	2.1 (1.4 to 3.2)	
17p12 HNPP del	Neuropathies	365 743 (194)	15 928 (25)	8.4	5.44×10^{-6}	0.0017	3.0 (2.0 to 4.5)	Yes
17p12 CMT1A dup	Neuropathies	365 609 (60)	15 959 (56)	2.6	3.9×10^{-124}	1.2×10^{-120}	21.8 (15.0 to 31.5)	Yes
17p12 CMT1A dup	Anaemia	362 411 (99)	19 157 (17)	5.2	6.7×10^{-5}	0.012	3.3 (1.9 to 5.4)	
17p12 CMT1A dup	Stroke	372 741 (107)	8827 (9)	2.5	0.0013	0.094	3.7 (1.8 to 6.9)	
17q12 del	Diabetes insulin dependent	378 861 (3)	2598 (4)	0.0	3.93×10^{-8}	3.08×10^{-5}	135.9 (31.2 to 641.1)	Yes
17q12 del	Digestive	299 609 (1)	81 850 (6)	0.3	0.00044	0.042	15.4 (3.2 to 150.3)	
17q12 dup	Renal failure	373 622 (90)	7929 (9)	1.9	0.00029	0.035	4.6 (2.2 to 8.6)	
22q11.2 dup	Hernia	333 779 (215)	47 939 (51)	30.9	0.0013	0.094	1.7 (1.2 to 2.3)	
22q11.2 dup	Gastric reflux	347 374 (217)	34 344 (49)	21.5	1.82×10^{-6}	0.0007	2.3 (1.7 to 3.1)	
22q11.2 distal del	Aneurysm	379 490 (4)	1967 (1)	0.0	0.0013	0.092	104.9 (9.7 to 673.6)	

The numbers of cases and controls are the numbers of people who have the phenotype. The 'Expected number of CNVs in cases' is extrapolated from their frequencies in the controls. Uncorrected p-values and Benjamini-Hochberg FDR corrected p-values for 3132 tests (p-value B-H FDR) are also shown. OR and 95% CI of the OR are produced by Firth's logistic regression analysis (Methods section). More details are given in online supplementary tables 6 and 7. 'Known finding' refers to known medical consequences, listed in online supplementary table 1, or to phenotypes that appear a consequence of the known ones (marked as 'secondary' in the table, as discussed below).
FDR, false discovery rate.

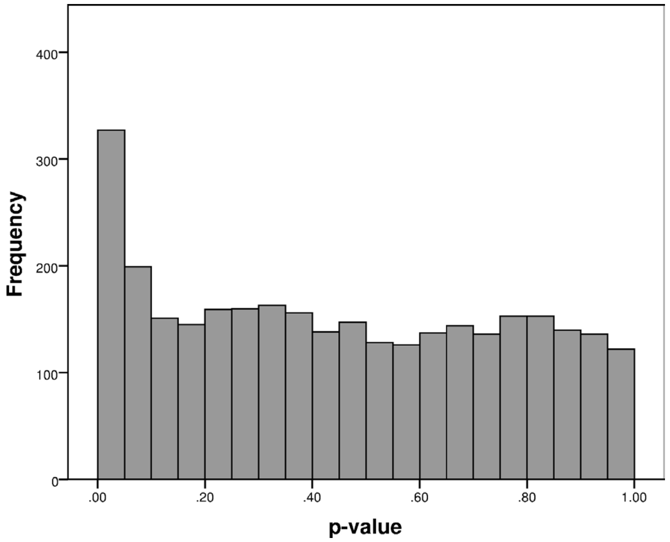

Figure 1 Distribution of all 3132 p-values from CNV/phenotype associations. There are 330 nominally significant CNV/phenotype associations ($p<0.05$), instead of the 157 expected by chance.

duplications,[15] obesity and deletions at 16p11.2 and 16p11.2 distal,[16 17] diabetes and 17q12 deletions (also called 'renal cysts and diabetes syndrome').[18] This increases our confidence that the newly identified associations are also real.

Effects of CNVs on medical phenotypes

Each of the 54 CNVs was tested for association with each of the 58 medical phenotypes (a total of 3132 tests). Results are presented as ORs for risk of developing the phenotype, corrected for age, sex and the other covariates detailed in the Methods section. All results are presented in online supplementary table 6 (grouped by CNV) and in online supplementary table 7 (grouped by phenotype).

The top 14 significant phenotype/CNV associations (table 2) survive a Bonferroni correction for 3132 tests (a project-wide significant p-value threshold of 1.6×10^{-5}). This correction is overconservative, due to medical comorbidities (eg, people with diabetes also have increased rates of heart attacks, stroke and others). A more appropriate correction of statistical significance for this analysis is the B-H FDR.[14] There are 46 CNV/phenotype comparisons that were significant at an FDR=0.1 (table 2). Most of these are novel associations and none are protective for the tested phenotypes (all have OR >1).

A total of 330 tests were nominally significant (at $p<0.05$), instead of the expected 157. Figure 1 shows the distribution of p-values, with a clear trend for over-representation below the $p<0.1$ level. This suggests that there are many more real associations, than those presented in table 2, but they cannot be identified with sufficient statistical significance in a sample of this size. Clinicians might therefore decide to also consider consequences of CNVs that do not survive our corrections.

Deletions at 16p11.2 and 16p12.1 had the largest numbers of significantly associated phenotypes (seven each). Deletions at 16p11.2 are a known risk factor for obesity.[16] We now provide data showing that adult carriers also have a high incidence of diabetes, osteoarthritis and hypertension, possibly as expected consequences/comorbidities of obesity. Other associated phenotypes are not necessarily linked to a high body mass index (BMI), such as asthma, anaemia and renal problems, suggesting that this and other CNVs have pleiotropic effects (see conditional analysis below). This should be expected from CNVs intersecting multiple genes. This has already been shown for some large CNVs, for example, 22q11.2 deletions, where highly variable phenotypic presentations are the norm.[5]

We should point out that CNVs with higher numbers of significant results are not necessarily the most pathogenic ones, as significance depends also on CNV frequency, which is low for the most pathogenic CNVs in this population. Such CNVs are under-represented in the UK Biobank, as the participants are middle-aged and participation is subject to 'healthy volunteer' selection bias.[19] For example, 22q11.2 deletions are highly pathogenic,[5] but there were only 10 such carriers in the Biobank, instead of the expected ~100 (the rate of this deletion among newborns is ~1:4000).[7] These 10 carriers were not sufficient to produce significant results at FDR=0.1, even for ORs>10 (online supplementary table 6). The more informative data from our research is on CNVs with lower penetrance, as they are more common.

The increased risk for medical morbidities or mortality observed in CNV carriers is unlikely to be due to the presence of early neurodevelopmental disorders or schizophrenia in carriers, as the UK Biobank population has largely escaped such conditions: only 34 of the 14791 people who had one of the tested CNVs had schizophrenia, 17 had developmental delay and 4 had autism. Accidental death or death in epilepsy cannot account for the increased death rate in CNV carriers: out of the 504 CNV carriers who had died during follow-up, only 1 had 'sudden unexpected death in epilepsy' and another 4 had accidental deaths (motor/pedal cyclist acidents and falls from a high place). All death causes in CNV carriers, according to the death registries, are listed in online supplementary table 8.

Phenotypes most likely to be affected by CNVs

Diabetes, hypertension, obesity and renal failure were the phenotypes affected by the highest number of CNVs (table 2). The real number of affected phenotypes by the CNVs is probably much higher, as suggested in figure 1. We can provide further evidence for this, by testing the effect on the phenotypes in the group of pathogenic CNV carriers as a whole, thus substantially increasing the statistical power. After excluding the five relatively common CNVs : deletions and duplications at 15q11.2 and 2q13(*NPHP1*) and duplications at 15q13.3(*CHRNA7*) (as they would determine the results due to their high frequencies), the remaining 4782 carriers of 49 rare CNVs had significantly increased risk for developing 26 of the 58 tested phenotypes (figure 2). Hypertension, diabetes, cardiac, respiratory and renal disorders dominate the top results. These are common phenotypes that increase mortality. We do indeed observe an increased death rate among CNV carriers during the follow-up period of Biobank participants (death was the second most-significant phenotype, figure 2). The RR of death from each CNV is presented in figure 3, where the RRs are ordered by the statistical strength of the association (strongest p-value on the left). The vertical line demarcates the 12 CNVs that are nominally significantly associated with increased mortality ($p<0.05$). Not surprisingly, the more pathogenic CNVs were also associated with increased mortality. The top significant CNV was, unexpectedly, the relatively common duplication at 16p13.11, found in ~0.2% of the general population, an association that has not been outlined before.

Most of the reported associations are novel, although some of them can be explained as logical adult medical consequences of known, early-onset phenotypes, for example, obesity leading

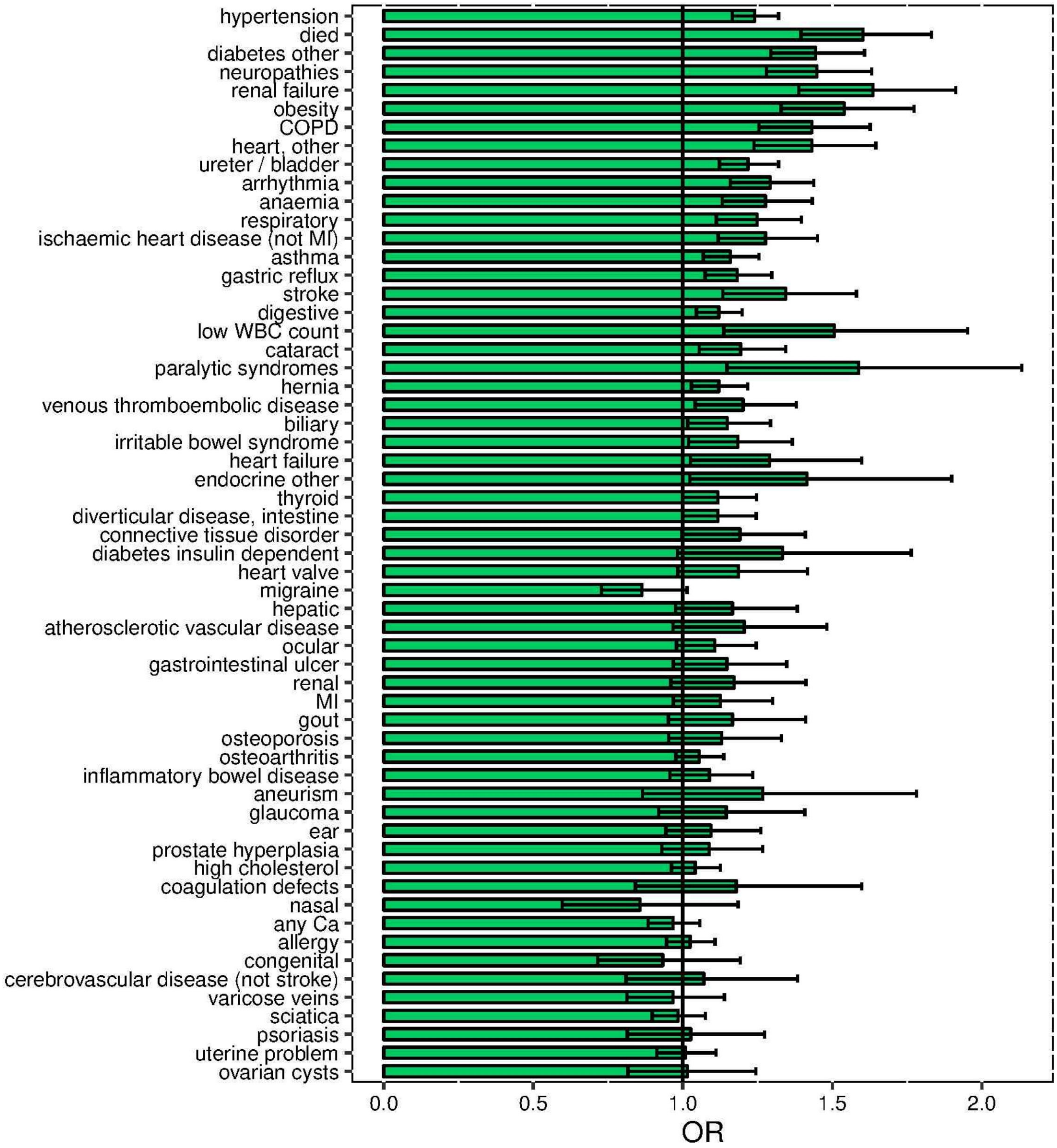

Figure 2 ORs and 95% CI for the ORs for developing the 58 tested phenotypes in carriers of any one of the 49 rare pathogenic CNVs. The phenotypes are ordered by the strength of the p-value. COPD, chronic obstructive pulmonary disease; MI, myocardial infarction, WBC, white blood cell count.

to diabetes, hypertension and increased cardiovascular mortality. In order to test this possibility, we performed a conditional analysis of three CNVs and two phenotypes, where obesity is most likely to account for some or all of the associations, by adding the BMI as a new covariate to the original analysis. This analysis amounted to 276 independent tests, to which we applied again the Benjamini-Hochberg FDR method to establish which associations remained significant at FDR=0.1, after controlling for BMI. Obesity is a well-established phenotype of 16p11.2 classic and distal deletions. The results and comparisons with the original analysis for all phenotypes and these two CNVs are shown in online supplementary tables 9 and 10 and supplementary figures 2 and 3. For 16p11.2 classic deletion, four of the six originally significant associations at FDR=0.1 remained significant (excluding obesity from these numbers). The changes in the ORs give a better global impression of the changes (online supplementary figure 2) and indicate that several associations are much reduced: diabetes type 1 and 2, hypertension, high cholesterol, gout and ostheoarthritis. This indicates that these disorders are, to a large extent, consequences of obesity. However, the ORs for anaemia and asthma did not change substantially. 16p11.2 distal deletions showed smaller reductions in the ORs (online supplementary figure 3) and four phenotypes (excluding obesity) remain significant at FDR=0.1. This pattern suggests that other factors also play a role in the causation of phenotypes in carriers of this CNV. Although deletions at 16p12.1 have not been an established cause for obesity, the pattern of results (table 2) also raised the question as to whether the multiple associated phenotypes could be explained by obesity. Therefore, we included this CNV in the conditional analysis (online supplementary table 11 and supplementary figure 4). Increased BMI appeared to play a smaller role in the causation of disease phenotypes for this CNV, with small changes in the ORs and the number of significant results.

Somewhat counterintuitively, the association with obesity does not get fully abolished when the analysis is corrected for BMI. There are, however, several factors that can explain this apparent anomaly. Most relevantly, the phenotype 'obesity' is

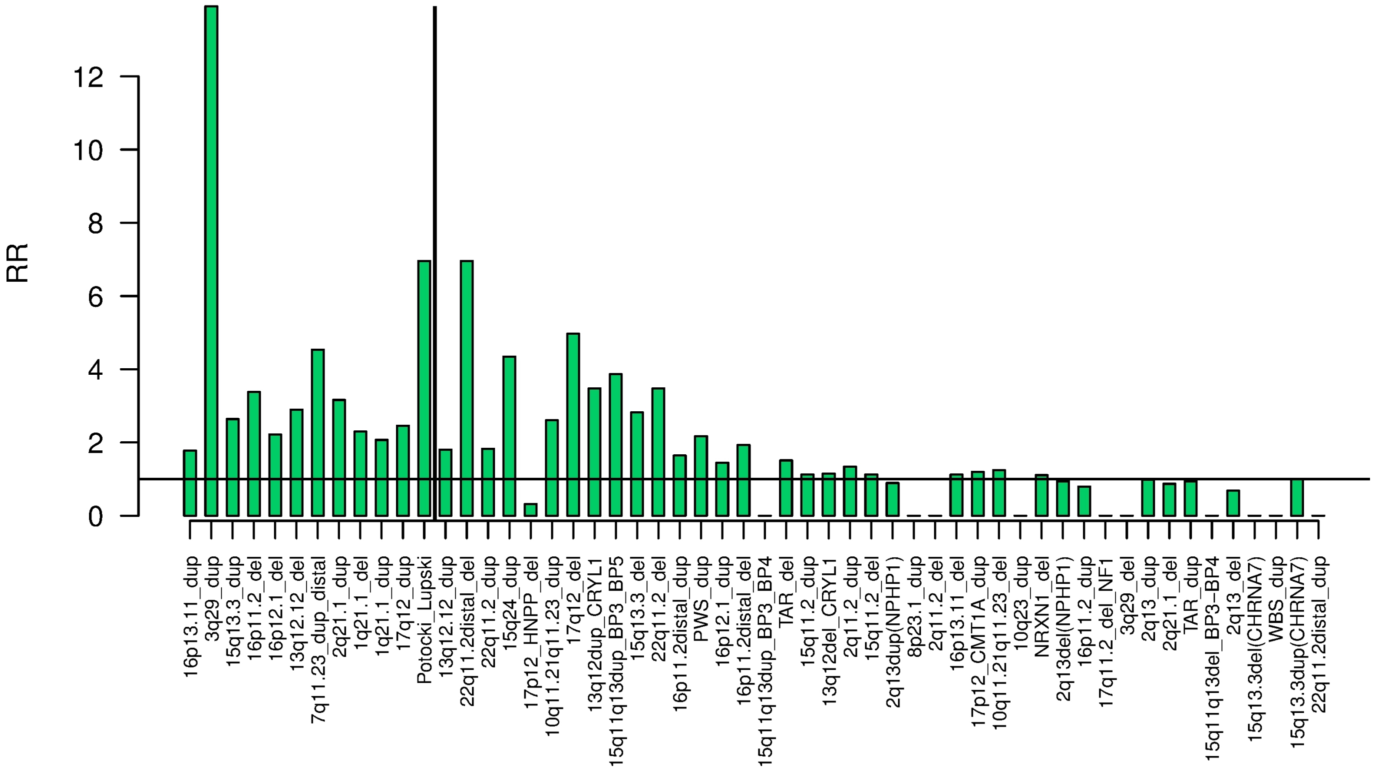

Figure 3 Relative risk (RR) for dying during the follow-up to 2016 for carriers of the 54 CNVs. The CNVs are ordered by the strength of the significance (strongest result on the left, for 16p13.11 duplications). The vertical line demarcates the nominally significant results (p<0.05). Due to zero observations in cases for some CNVs, RRs are shown, instead of ORs.

not equivalent to high BMI. It is a hospital ICD-10 diagnosis, made on a small proportion of people who have a BMI>30. In fact, 24.3% of the Biobank population has a BMI>30, qualifying them for a diagnosis of obesity, but only 9.2% of them received this diagnosis. Furthermore, obesity is a categorical variable, while BMI is a continuous one, making them not equivalent from a statistical point of view, and therefore adjusting an analysis of one for another does not necessarily remove all evidence for association. The distribution of BMI values is very different in the three CNVs tested: 71.6% of 16p11.2 deletion carriers had a BMI>30, compared with 55.6% of 16p11.2 distal deletion carriers and 37% of 16p12.1 deletion carriers (online supplementary figure 5a–c). ICD-10 diagnosis of 'obesity' was given to correspondingly smaller proportions of carriers: 18.6%, 16.7% and 9.8%. These differences could explain why correcting for BMI does not lead to identical changes to the associations of the three CNVs.

We also tested whether increased BMI accounted for associations of diabetes type 2 or mortality with any of the 54 CNVs (online supplementary tables 12 and 13 and supplementary figures 6 and 7). As already reported above, this was the case for diabetes and the 'classic' and 'distal' 16p11.2 deletions. However, for 1q21.1and 2q13 duplications, 22q11.2 distal deletions and 17q12 deletions (also known as 'renal cysts and diabetes syndrome'), the ORs for diabetes increased, suggesting that these CNVs have a more direct effect on the development of diabetes. In total, six CNVs were significantly associated with diabetes, after controlling for BMI (online supplementary table 12). The associations with mortality remained essentially unchanged after correction with BMI, with four significantly associated CNVs (online supplementary table 13) and very similar ORs (online supplementary figure 7), indicating that obesity is only one of many consequences that shortens the lives of CNV carriers.

Homozygous deletions and more than one CNV per person

Only four carriers of homozygous deletions were found, perhaps not surprisingly for this relatively healthy population. Three of these clustered in a single locus, 2q13 (11 086–11 098 kb), affecting the gene *NPHP1*. Homozygous deletions at this locus are known to cause the kidney disorder juvenile nephronophthisis. All three Biobank individuals with homozygous deletions at *NPHP1* had renal failure (Fisher's exact test $p=9\times10^{-6}$). We also examined the data for the occurrence of two CNVs in the same person. 264 people carried two of these CNVs, not significantly different from the 249 expected by chance. All combinations of two CNVs observed in the same person are presented in online supplementary table 14.

Monitoring of CNV carriers

Our results indicate a need for regular medical monitoring of apparently healthy carriers of specific pathogenic CNVs. Examples include monitoring for blood pressure, kidney function and glucose levels for carriers of 16p12.1 and 16p11.2 deletions, and for cancer in 3q29 duplication carriers. Apart from specific medical phenotypes, it appears that such carriers require enhanced medical monitoring in general, as their health can be affected in multiple ways. Our results should enable clinicians to better plan the medical management of CNV carriers.

Finally, the reported CNV morbidity map can provide researchers with another avenue for the elucidation of pathophysiological disease mechanisms.

Acknowledgements This research has been conducted using the UK Biobank Resource under Application no: 14421.

Contributors KC, MB-S and DO analysed the data; KMK, ER and MB-S called the CNVs; AFP and ME contributed to the bioinformatics and website design; VE-P, JTRW, MCO'D and MJO contributed to the statistical analysis; JTRW, MCO'D and MJO edited the paper; GK conceived the project, drafted the paper and took part in all analysis steps.

Funding The work at Cardiff University was funded by the Medical Research Council (MRC) Centre Grant (MR/L010305/1) and Programme Grant (G0800509).

Competing interests None declared.

Patient consent Not required.

Ethics approval Ethical approval for the study was granted by the North West multi-centre ethics committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All CNV calls will be made available to the UK Biobank, in accordance with their requirements, within 6 months of the first publication of results.

REFERENCES

1. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 2010;12:e8.
2. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney JF, Klei L, Mandell JD, Moreno-De-Luca D, Poultney CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, Lowe JK, Mane SM, Martin DM, Morrow EM, Talkowski ME, Sutcliffe JS, Walsh CA, Yu TW, Ledbetter DH, Martin CL, Cook EH, Buxbaum JD, Daly MJ, Devlin B, Roeder K, State MW. Autism Sequencing Consortium. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 2015;87:1215–33.
3. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LE, Schuurs-Hoeijmakers JH, Hoischen A, Pfundt R, Krumm N, Carvill GL, Li D, Amaral D, Brown N, Lockhart PJ, Scheffer IE, Alberti A, Shaw M, Pettinato R, Tervo R, de Leeuw N, Reijnders MR, Torchia BS, Peeters H, O'Roak BJ, Fichera M, Hehir-Kwa JY, Shendure J, Mefford HC, Haan E, Gécz J, de Vries BB, Romano C, Eichler EE. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 2014;46:1063–71.
4. Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, Mahoney-Davies G, Legge SE, Moran JL, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G. Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* 2014;204:108–14.
5. McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JA, Zackai EH, Emanuel BS, Vermeesch JR, Morrow BE, Scambler PJ, Bassett AS. 22q11.2 deletion syndrome. *Nat Rev Dis Primers* 2015;1:15071.
6. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodríguez Rojas LX, Elton LE, Scott DA, Schaaf CP, Torres-Martinez W, Stevens AK, Rosenfeld JA, Agadi S, Francis D, Kang SH, Breman A, Lalani SR, Bacino CA, Bi W, Milosavljevic A, Beaudet AL, Patel A, Shaw CA, Lupski JR, Gambin A, Cheung SW, Stankiewicz P. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* 2013;23:1395–409.
7. Kirov G, Rees E, Walters JT, Escott-Price V, Georgieva L, Richards AL, Chambert KD, Davies G, Legge SE, Moran JL, McCarroll SA, O'Donovan MC, Owen MJ. The penetrance of copy number variations for schizophrenia and developmental delay. *Biol Psychiatry* 2014;75:378–85.
8. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE. A copy number variation morbidity map of developmental delay. *Nat Genet* 2011;43:838–46.
9. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M, Billington CK, Kheirallah AK, Allen R, Cook JP, Probert K, Obeidat M, Bossé Y, Hao K, Postma DS, Paré PD, Ramasamy A, Mägi R, Mihailov E, Reinmaa E, Melén E, O'Connell J, Frangou E, Delaneau O, Freeman C, Petkova D, McCarthy M, Sayers I, Deloukas P, Hubbard R, Pavord I, Hansell AL, Thomson NC, Zeggini E, Morris AP, Marchini J, Strachan DP, Tobin MD, Hall IP. UK Brain Expression Consortium (UKBEC) OxGSK Consortium. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 2015;3:769–81.
10. Kendall KM, Rees E, Escott-Price V, Einon M, Thomas R, Hewitt J, O'Donovan MC, Owen MJ, Walters JTR, Kirov G. Cognitive performance among carriers of pathogenic copy number variants: Analysis of 152,000 UK Biobank subjects. *Biol Psychiatry* 2017;82:103–10.
11. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007;17:1665–74.
12. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST. Working group of the American college of medical genetics laboratory quality assurance committee. American college of medical genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med* 2011;13:680–5.
13. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80:27–38.
14. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B* 1995;57:289–300.
15. Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, Ledbetter DH, Greenberg F, Patel PI. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. *Nat Genet* 1992;1:29–33.
16. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, Martinet D, Shen Y, Valsesia A, Beckmann ND, Thorleifsson G, Belfiore M, Bouquillon S, Campion D, de Leeuw N, de Vries BB, Esko T, Fernandez BA, Fernández-Aranda F, Fernández-Real JM, Gratacòs M, Guilmatre A, Hoyer J, Jarvelin MR, Kooy RF, Kurg A, Le Caignec C, Männik K, Platt OS, Sanlaville D, Van Haelst MM, Villatoro Gomez S, Walha F, Wu BL, Yu Y, Aboura A, Addor MC, Alembik Y, Antonarakis SE, Arveiler B, Barth M, Bednarek N, Béna F, Bergmann S, Beri M, Bernardini L, Blaumeiser B, Bonneau D, Bottani A, Boute O, Brunner HG, Cailley D, Callier P, Chiesa J, Chrast J, Coin L, Coutton C, Cuisset JM, Cuvellier JC, David A, de Freminville B, Delobel B, Delrue MA, Demeer B, Descamps D, Didelot G, Dieterich K, Disciglio V, Doco-Fenzy M, Drunat S, Duban-Bedu B, Dubourg C, El-Sayed Moustafa JS, Elliott P, Faas BH, Faivre L, Faudet A, Fellmann F, Ferrarini A, Fisher R, Flori E, Forer L, Gaillard D, Gerard M, Gieger C, Gimelli S, Gimelli G, Grabe HJ, Guichet A, Guillin O, Hartikainen AL, Heron D, Hippolyte L, Holder M, Homuth G, Isidor B, Jaillard S, Jaros Z, Jiménez-Murcia S, Helas GJ, Jonveaux P, Kaksonen S, Keren B, Kloss-Brandstätter A, Knoers NV, Koolen DA, Kroisel PM, Kronenberg F, Labalme A, Landais E, Lapi E, Layet V, Legallic S, Leheup B, Leube B, Lewis S, Lucas J, MacDermot KD, Magnusson P, Marshall C, Mathieu-Dramard M, McCarthy MI, Meitinger T, Mencarelli MA, Merla G, Moerman A, Mooser V, Morice-Picard F, Mucciolo M, Nauck M, Ndiaye NC, Nordgren A, Pasquier L, Petit F, Pfundt R, Plessis G, Rajcan-Separovic E, Ramelli GP, Rauch A, Ravazzolo R, Reis A, Renieri A, Richart C, Ried JS, Rieubland C, Roberts W, Roetzer KM, Rooryck C, Rossi M, Saemundsen E, Satre V, Schurmann C, Sigurdsson E, Stavropoulos DJ, Stefansson H, Tengström C, Thorsteinsdóttir U, Tinahones FJ, Touraine R, Vallée L, van Binsbergen E, Van der Aa N, Vincent-Delorme C, Visvikis-Siest S, Vollenweider P, Völzke H, Vulto-van Silfhout AT, Waeber G, Wallgren-Pettersson C, Witwicki RM, Zwolinksi S, Andrieux J, Estivill X, Gusella JF, Gustafsson O, Metspalu A, Scherer SW, Stefansson K, Blakemore AI, Beckmann JS, Froguel P. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 2011;478:97–102.
17. Bachmann-Gagescu R, Mefford HC, Cowan C, Glew GM, Hing AV, Wallace S, Bader PI, Hamati A, Reitnauer PJ, Smith R, Stockton DW, Muhle H, Helbig I, Eichler EE, Ballif BC, Rosenfeld J, Tsuchiya KD. Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. *Genet Med* 2010;12:641–7.
18. Mefford HC, Clauin S, Sharp AJ, Moller RS, Ullmann R, Kapur R, Pinkel D, Cooper GM, Ventura M, Ropers HH, Tommerup N, Eichler EE, Bellanne-Chantelot C. Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am J Hum Genet* 2007;81:1057–69.
19. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *Am J Epidemiol* 2017;186:1026–34.