Structural disruption of genomic regions containing ultraconserved elements is associated with neurodevelopmental phenotypes

Ruth B. McCole[1], Wren Saylor[1], Claire Redin[2,3,4], Chamith Y. Fonseka[1], Harrison Brand[2,3,4], Jelena Erceg[1], Michael E. Talkowski[2,3,4], and C.-ting Wu[1]*

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

[2]Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

[3]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

[4]Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02141, USA.

*Corresponding author. Email: twu@genetics.med.harvard.edu. Address: Department of Genetics, 77 Avenue Louis Pasteur, NRB 264, Boston, MA 02115, USA. Phone: (617) 432-4431. Fax: (617) 432-7663.

Short title: Ultraconserved elements in neurodevelopment

Abstract

The development of the human brain and nervous system can be affected by genetic or environmental factors. Here we focus on characterizing the genetic perturbations that accompany and may contribute to neurodevelopmental phenotypes. Specifically, we examine two types of structural variants, namely, copy number variation and balanced chromosome rearrangements, discovered in subjects with neurodevelopmental disorders and related phenotypes. We find that a feature uniting these types of genetic aberrations is a proximity to ultraconserved elements (UCEs), which are sequences that are perfectly conserved between the reference genomes of distantly related species. In particular, while UCEs are generally depleted from copy number variant regions in healthy individuals, they are, on the whole, enriched in genomic regions disrupted by copy number variants or breakpoints of balanced rearrangements in affected individuals. Additionally, while genes associated with neurodevelopmental disorders are enriched in UCEs, this does not account for the excess of UCEs either in copy number variants or close to the breakpoints of balanced rearrangements in affected individuals. Indeed, our data are consistent with some manifestations of neurodevelopmental disorders resulting from a disruption of genome integrity in the vicinity of UCEs.

Introduction

The etiology of neurodevelopmental disorders (NDDs) involves both genetic and environmental factors. In this study, we are concerned with the genetic, rather than environmental, disruptions that may result in NDDs including autism spectrum disorder (ASD) and other neurodevelopmental phenotypes. Evidence for a large genetic component to ASD etiology comes from studies over many decades showing a much higher concordance for ASD in monozygotic twins when compared with dizygotic twins (reviewed in Huguet *et al.*[1]). More generally, NDDs have been associated with an increasingly large range of genetic disruptions that are highly diverse in genomic location, type, and origin and include single nucleotide changes, copy number variations (CNVs), and complex, sometimes copy number neutral, chromosomal rearrangements (reviewed in Kloosterman *et al.*[2], Hu *et al.*[3], and Sahin *et al.*[4]). With regard to the origin of NDD-associated genomic aberrations, *de novo* events have recently been a major focus, although somatic events during development and inherited variation have also been examined (reviewed in Ronemus *et al.*[5] and Hu *et al.*[3]).

Since genetic alterations found in subjects with NDDs and related phenotypes vary in genomic position, aberration type, and origin, we wondered whether examining these aberrations from the perspective of comparative genomics might unveil commonalities among this set of genomic abnormalities. We chose this approach in light of the fact that high sequence conservation is a hallmark of neurodevelopmental[6], presynaptic, and synaptic genes[7,8]. Specifically, this study queries the relationship between genomic rearrangements in individuals with NDDs or related phenotypes and a set of human sequences known as ultraconserved elements (UCEs). Examples of highly

conserved genes with relevance to neurodevelopment and that also contain UCEs include AUTS2[9], FOXP2[10], and BCL11A[11].

First reported in 2004, human UCEs are defined as regions of the human genome displaying extreme sequence conservation with distantly related vertebrates[12-14]. They comprise one of the most enigmatic aspects of our genomes, as the underlying reasons for ultraconservation are still being debated; while UCEs contribute to gene regulation, contain many transcription factor binding motifs, are often transcribed, and may produce phenotypes when mutated or deleted, these properties are not universally considered sufficient to explain ultraconservation[14-30], reviewed in Elgar *et al.* [31] and Harmston *et al.*[32]. We suggest that ultraconservation may derive from a mechanism in which maternal and paternal copies of UCEs are compared within the nucleus, such that individuals bearing discrepancies in sequence or copy number of UCEs or rearrangements that disrupt the comparison have reduced fitness[33-35] (see also Elgar *et al.*[31] and Kritsas *et al.*[36]). Such a process has the potential of reducing the burden in germline, embryonic, and adult somatic cell lineages of cells carrying deleterious genomic aberrations. Interestingly, the positioning and clustering of UCEs along chromosomes can be highly conserved[13,37-41], and regions of the genome that include highly conserved elements appear to interact in three-dimensional space[42]. These observations are in line with the possibility that UCEs have a function that requires their specific three-dimensional positioning within the nucleus, such as the pairing of homologous genomic regions containing UCEs. Importantly, ample evidence exists for the capacity of vertebrate cells to support homolog pairing outside as well as during meiosis (reviewed in Joyce *et al.*[43] and references within) and a role for pairing in gene regulation, sequence comparison,

and copy counting (e.g. Joyce *et al.*[43] Hammond *et al*.[44], Gladyshev *et al*.[45], and references within). Our model is consistent with the viability and fertility observed by Ahituv *et al*.[46] of mice lacking both copies of any of four noncoding UCEs, as deletion of both copies of a UCE would preclude detection of these deletions. It is also compatible with additional, gene regulatory functions for UCEs and therefore is not contradicted by discoveries of abnormal phenotypes, such as improper eye development[25], that arise from mutations in UCEs.

This model predicts that changes to UCE copy number would be disfavored in healthy cells, and consistent with this, meta-analyses of over two dozen datasets of copy number variations (CNVs) representing healthy cells showed a strong depletion of UCEs from deleted or duplicated regions[33-35,47], while a study of plants has shown that sequences conserved in distantly related plant genomes are depleted from segmental duplications[36]. The model also predicts that disruption of UCE copy number above the low levels seen in healthy tissues would be associated with disease, and we have found evidence for this from our meta-analysis of seventeen datasets of CNVs derived from cancer tissues. In particular, we observed a higher level of UCE disruption by cancer-specific CNVs than by CNVs in healthy tissues, with a number of cancer CNV datasets showing an enrichment for UCEs[35]. Of special relevance to NDD, Martinez *et al*.[48] pursued our prediction that abnormal dosage of UCEs would be associated with lowered fitness[33-35] by querying whether the relationship between CNVs and UCEs in individuals with mental delay and congenital abnormalities would differ from that in healthy individuals. They showed that there was indeed an association between UCE position and the structural variants discovered in the genomes of these patients.

Our previous studies all focused on elucidating the positional relationship between UCEs and CNVs, but did not address the relationship between UCEs and copy number neutral rearrangements. The present study also examines CNVs and then, for the first time, addresses copy number neutral rearrangements, taking advantage of the large number of new datasets describing structural variants in subjects with NDD or related phenotypes. Thus, in addition to elucidating the positional relationship between UCEs and rearrangement breakpoints, this study tackles the prediction that genomic disruptions that compromise the comparison process of UCEs will be associated with disease. We begin by analyzing CNVs in a first cohort with respect to their origin, separating inherited CNVs from *de novo* CNVs that are, as a group, associated with NDD causation[5,49,50]. We then compare *de novo* CNVs in a separate ASD cohort to *de novo* CNVs discovered in unaffected siblings. Finally, we examine inversion, translocation, and complex rearrangements in subjects with NDDs or related phenotypes.

Our results demonstrate a striking departure from the depletion of UCEs observed in regions affected by CNVs in healthy individuals. Indeed, individuals with NDD or related phenotypes reveal an enrichment of UCEs in genomic regions encompassed by CNVs or flanking the breakpoints of balanced rearrangements. Importantly, while UCEs are also prevalent within genes whose disruption is thought to elevate NDD risk, this prevalence is not sufficient to explain the excess of UCEs encompassed in the CNVs, nor the extent of the enrichment of UCEs near the breakpoints of balanced rearrangements, in affected cohorts. In summary, our findings suggest three non-conflicting possibilities: 1) UCEs may signpost new candidate genes and critical regions for NDD involvement, as has been suggested previously[48]; 2) disruption of either UCE dosage or nuclear position

134 may itself be a causal factor in NDD etiology; and 3) UCEs may function to support

135 genomic integrity, with NDD being one possible outcome when they are compromised.

Materials and Methods

Datasets

All datasets are available from the publications or web resources listed and *per* these sources the procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation and proper informed consent was obtained.

Vulto-van Silfhout ASD *de novo* CNVs and Vulto-van Silfhout ASD inherited CNVs (See Figure S1): CNV regions corresponding to subjects with autism spectrum disorder were taken from Vulto van-Silfhout *et al*.[51]. Out of 1663 events, 206 events corresponded to subjects with ASD phenotype. Of these, 54 were *de novo*, and 94 were inherited. When overlapping events were combined, 45 regions made up our dataset 'Vulto-van Silfhout ASD *de novo* CNVs', and 81 regions comprised our dataset 'Vulto-van Silfhout ASD inherited CNVs'.

Sanders ASD *de novo* CNVs and Sanders sibling *de novo* CNVs (See Figure S1): CNVs were drawn from Sanders *et al*. [52]. Inherited CNVs were not considered, because only rare inherited CNVs were available from Sanders *et al*. and considering only a subset of all inherited CNVs would not provide us with a full representation of their relationships to UCEs. For *de novo* CNVs, there were 495 events for cases and 121 events for controls. These were filtered to retain validated events, leaving 328 case events and 90 control events. Whole chromosome aneuploidies, as well as one event covering >94% of the Y chromosome q arm, were removed. For cases, 321 events remained,

159 which when overlapping events were combined, produced 191 regions, which we call
160 'Sanders ASD *de novo* CNVs'. For controls, 86 events remained, which, when
161 overlapping events were combined, produced 79 regions that we refer to as 'Sanders
162 sibling *de novo* CNVs' (See Figure S1).
163
164 **NDD breakpoints set 1** (See Figure S3):
165 Breakpoints were taken from 5 publications: Granot-Hershkovitz *et al*. 2011 [53],
166 Kloosterman *et al*. 2011 [54], Chiang *et al*. 2012 [55], Talkowski *et al*. 2012 [56], and Nazaryan
167 *et al*. 2014 [57] (see Table S1 and Figure S3). Only breakpoints documented to accompany
168 ≤1kb of inserted or deleted DNA were included; we did not place any limits on the
169 amount of DNA rearranged by the variant. Duplicate breakpoints and breakpoints from
170 subjects without neurodevelopmental phenotypes were removed. Where necessary,
171 coordinates were converted to genome build hg18 using the UCSC liftover tool
172 (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The final 'NDD breakpoints set 1' dataset
173 comprised 157 breakpoints. Flanking regions of 100kb on either side of each breakpoint
174 were added using the Bedtools [58] slop tool, implemented using Pybedtools [59], producing
175 dataset 'NDD breakpoints set 1 with 100kb flanks', consisting of 76 regions.
176
177 **NDD breakpoints Set 2** (See Figure S3):
178 Breakpoints were taken from Redin *et al*.[60]. 1858 breakpoints were filtered to retain only
179 those where the total genomic imbalance per individual is ≤1kb; we did not place any
180 limits on the amount of DNA rearranged by the variant. These breakpoints were filtered
181 to retain only those where the breakpoint was confirmed by capillary sequencing,

182 resulting in 744 breakpoints. To produce a set of breakpoints that are different from those
183 in NDD breakpoints Set 1, we removed any breakpoints listed by Redin *et al*.[60] that were
184 included in NDD breakpoints Set 1, leaving 590 breakpoints. We then filtered to obtain
185 only breakpoints from patients reported as having an ear, eye, or nervous system
186 phenotype, leaving 478 breakpoints. We also filtered to retain breakpoints from subjects
187 with neither ear, nor eye, nor nervous system phenotypes, and these breakpoints formed
188 our 'Non-NDD breakpoint' dataset, described below. Duplicate breakpoints were
189 removed, leaving 453 breakpoints, which comprised our final dataset 'NDD breakpoints
190 set 2'. Flanking regions of 100kb on either side of each breakpoint were added using the
191 Bedtools [58] slop tool, implemented using Pybedtools [59], and overlapping regions were
192 combined to produce 224 regions, which we refer to as 'NDD breakpoints Set 2 with
193 100kb flanks' .
194
195 **Pooled NDD breakpoints** (See Figure S3): We combined the 157 breakpoints in NDD
196 breakpoints set 1 with the 453 breakpoints in NDD breakpoint set 2 to create our dataset
197 Pooled NDD breakpoints' of 610 breakpoints. Flanking regions of 20kb, 50kb, 100kb,
198 500kb, and 1Mb on either side of each breakpoint were added using the Bedtools[58] slop
199 tool, implemented using Pybedtools[59], and overlapping regions were combined to
200 produce five new datasets, 'Pooled NDD breakpoints with 20kb flanks' containing 316
201 regions, 'Pooled NDD breakpoints with 50kb flanks' with 310 regions, 'Pooled NDD
202 breakpoints with 100kb flanks' with 296 regions, 'Pooled NDD breakpoints with 500kb
203 flanks' with 250 regions, and 'Pooled NDD breakpoints with 1Mb flanks', with 219
204 regions. To examine the distances between UCEs and Pooled NDD breakpoints, we

205 allocated any breakpoint in our Pooled NDD breakpoint set into a cluster if it occurred
206 within 1kb of any other breakpoint. We then retained just one breakpoint from each
207 cluster, producing our dataset 'Pooled NDD breakpoints within 1kb clustered', which
208 contained 331 breakpoints.
209
210 **Pathogenic, likely pathogenic, and VUS breakpoints** (See Figure S3): We divided our
211 NDD breakpoints set 2 according to the category of pathogenicity the breakpoints for
212 each patient fell into. The categories were provided in Redin *et al*.[60], and consisted of
213 Pathogenic, Likely Pathogenic, and Variants of Unknown Significance (VUS). In total,
214 64 were from patients classified as having pathogenic variants, 150 were classified as
215 having likely pathogenic variants, and 264 were from patients with variants of unknown
216 significance. When 500kb flanks were added to either side of each breakpoint and
217 overlapping regions were combined, the breakpoints made up the datasets 'Pathogenic
218 breakpoints with 500kb flanks' with 32 regions, 'Likely pathogenic breakpoints with
219 500kb flanks' with 62 regions, and 'VUS breakpoints with 500kb flanks' with 120
220 regions.
221
222 **Non-NDD breakpoints** (See Figure S3): Breakpoints from Redin *et al*.[60] were filtered as
223 described above and by phenotype to retain 76 breakpoints from subjects with neither ear,
224 nor eye, nor nervous system phenotypes. 500kb flanks were added to these breakpoints to
225 produce 38 regions.
226

Loss-of-function (LoF) genes: Genes with two or more loss-of-function mutations in ASD patients were collated by Redin *et al.* [60]. The coordinates were converted from build hg19 to build hg18 using the UCSC liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The final gene list comprised 304 genes and is referred to as 'LoF genes' (see Table S1).

Constrained genes: We obtained all gene identifiers from Lek *et al.* [61] that had pLI score > 0.9, indicating these genes are loss-of-function intolerant in humans and likely to be selectively constrained. These identifiers were used to obtain coordinates for each gene in human genome build hg19, using the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and track 'GENCODE genes V 19'. The coordinates were then lifted over to build hg18 using the UCSC liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The final list comprised 3249 genes and is called 'Constrained genes' (See Table S1).

Embryonic brain genes: We collected genes that showed higher expression in the embryonic human brain when compared to the postnatal human brain, identified by Iossifov *et al.*[62]. The coordinates were lifted over to build hg18 using the UCSC liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The final dataset, named ' Embryonic brain genes' comprised 1903 genes (See Table S1).

Type 2 diabetes (T2D) genes: We obtained genes associated with type 2 diabetes from Morris *et al.*[63], matching the gene symbols provided with gene coordinates in build hg18 using the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and manual

250 searches on the UCSC genome browser (http://genome.ucsc.edu). In total, our 'T2D
251 genes' dataset is made up of 70 genes (See Table S1).
252
253 **Schizophrenia (SZ) genes:** Genes implicated in schizophrenia risk by means of genome
254 wide association studies (GWAS) were collected from the NHGRI-EBI GWAS catalog[64]
255 at http://www.ebi.ac.uk/gwas/search?query=Schizophrenia. Genes were filtered to retain
256 those labeled 'schizophrenia'; genes with label 'schizophrenia and autism' were not
257 included.
258
259 **Enhancers**: Enhancers dataset was from the ENCODE combined genome segmentation
260 from the ENCODE UCSC hub[65] 'E' (enhancer) class genomic regions for six ENCODE
261 cell/tissue types.
262
263 **Human accelerated regions (HARs):** Genome coordinates for HARs were obtained
264 from Doan *et al*.[66], and lifted over to hg18 using the UCSC liftover tool
265 (http://genome.ucsc.edu/cgi-bin/hgLiftOver).
266
267 **Repetitive elements:** Genomic coordinates for repetitive elements were obtained from
268 the UCSC genome browser Repeat Masker track.
269
270 **UCEs:** Our UCE dataset comprises 896 regions that display 100% DNA conservation
271 between all members of these three groups: 1) human, mouse, and rat; 2) human, dog,
272 and mouse, and 3) human and chicken. These sequences are available in McCole *et al.*[35]

and are provided in Table S1. We divide the UCEs into intergenic, intronic, and exonic elements, defining introns and exons using the UCSC known genes track for genome build hg18. Exonic elements are those UCEs that wholly or partially overlap exons, and comprise 179 elements. Intronic elements are any of the non-exonic UCEs that wholly or partially overlap introns, and comprise 416 elements. The remaining 301 UCEs are categorized as intergenic.

Depletion and enrichment analysis

To determine depletion or enrichment of UCEs in genomic regions of interest, we use a similar method as previously described[33-35]. We begin by merging all overlapping genomic regions of interest, using a custom python script, which functions almost identically to the BEDtools[58] merge function. We then remove from these merged genomic regions of interest any overlaps with unsequenced portions of the human genome. Next we calculate the overlap in bp between merged genomic regions of interest and each of the three categories of UCEs (intergenic, intronic, exonic). For example, for the intergenic UCEs, the pipeline first calculates the overlap in basepairs between the intergenic UCEs and CNVs (or other regions of interest), which we refer to as the 'observed overlap'. Next, the pipeline randomly permutes the positions of the UCEs within the appropriate portion of the genome; with, for example, the intergenic UCEs being permuted only within intergenic regions. The pipeline then calculates the overlap between the CNVs and the randomly positioned elements that are matched in number, length, and genomic region, and we call this the 'expected overlap'. The randomization process is then repeated 1,000 times to generate a distribution of expected overlaps.

When analyzing the entire UCE set, this process is repeated for intronic and exonic UCEs and then the intergenic, intronic, and exonic observed overlaps are added to produce the observed overlap for the entire UCE set, and similarly, expected overlaps for intergenic, intronic, and exonic regions within each of the 1,000 permutations are added to produce a distribution of expected overlaps for the entire UCE set. The ratio of observed/expected overlap is reported as 'obs/exp'. To assess statistical significance, the distribution of expected overlaps is assessed for normality using a Kolomogorov-Smirnov (KS) test, and the resulting P-value is reported. If the distribution is consistent with normality, a Z-test is used to determine the statistical significance of the depletion (obs/exp < 1) or enrichment (obs/exp > 1) of observed UCE overlap with genomic regions of interest, compared with expected overlap, and the resulting P value is reported. P values below 10^{-17} are not calculated exactly by the test and so are reported as here as $<1\times10^{-17}$. Occasionally, our distribution of expected overlaps is not normally distributed, precluding the generation of a P-value and, in this situation, we provide instead the proportion of expected overlaps that equaled or exceeded the observed overlap. To allow for the possibility of either enrichment or depletion, a two-tailed test is employed with a combined α of 0.05.

Binomial test for overlap of Pooled NDD breakpoints with UCEs

To calculate the probability of observing 1 or more breakpoints overlapping UCEs, we calculated that UCEs cover 241,287bp. Assuming a 3 Gb human genome, the proportion of the human genome that UCEs make up is 0.00008. We then performed a binomial test

with the probability of success set to 0.00008 and the number of trials equal to the number of breakpoints.

Distances from breakpoints to UCEs

UCEs within 1kb of each other were grouped and, for each group, one member of the group was chosen randomly for analysis, to avoid biasing the results with very closely clustered UCEs. Distances between UCEs and the nearest breakpoint were calculated using a custom python script utilizing Pybedtools[59]. The resulting distances were then expressed as frequencies in the form of a histogram. Additionally, 1,000 sets of regions matched to UCEs in terms of length, number, and genomic portion (intergenic, intronic, or exonic), which we call 'control regions' were generated using our depletion/enrichment analysis pipeline. The distances between these control regions and their nearest breakpoint was then calculated. The distribution of UCE-breakpoint distances, and of control-breakpoint distances, was compared using an Anderson-Darling test implemented in python using the scipy function anderson_ksamp, which provides exact significance thresholds and an approximated P value (stated). The maximum distance allowed was set at 5.2Mb, to match the distribution displayed in Figure 2E. For visualization of the results and comparison to the histogram of UCE-breakpoint distances, the distances in each of the 1,000 trials were expressed as frequencies and the interquartile range between the 1,000 trials at each distance was plotted.

339 Depletion and enrichment analysis with gene sets excluded

340 We carried out this analysis using the same pipeline as for 'Depletion and enrichment
341 analysis', above. We excluded the regions of the genome corresponding to our gene set of
342 interest, as well as in some experiments the regions 100kb upstream and downstream of
343 these genes. This means that any UCEs and regions of interest such as CNVs that fall
344 within the excluded regions were not included in the analysis. Additionally, no UCE-
345 matched random control regions were allowed to fall within the excluded regions.

346

347 Partial correlation analysis

348 Data for enhancers, HARs, and repetitive elements was obtained as described above. GC
349 percentage was calculated using the hg18 human reference genome and bedtools[58,59] 'nuc'
350 function. Custom python scripts were used to calculate the density of genomic regions of
351 interest in bp within bins of 100kb by summing the number of bp covered by the regions
352 and dividing this by the genomic coverage of the bin (100kb). Bins containing
353 unsequenced regions of the reference genome were excluded from the analysis. The
354 Spearman partial correlation coefficient and accompanying P-value, which represents the
355 partial correlation between UCE density and a feature of interest such as Pooled NDD
356 breakpoints after removal of any contribution of any correlation with a third feature, such
357 as enhancers, was then calculated using matlab function *partialcorr* implemented within
358 custom python scripts.

359

360 UCEs overlapping regions of interest (Table S3)

361 For each UCE, its ID, coordinates, type (intergenic, intronic, exonic), and the UCSC and

362 HGNC gene symbol for any overlapping gene is listed. Then, for each dataset listed in

363 Table S1, if the UCE is overlapped by this dataset, a '1' is listed next to this UCE. If the

364 UCE does not overlap this dataset, a '0' is listed.

365

366 Scripts

367 Custom python scripts associated with this study can be accessed at

368 https://github.com/rmccole/UCEs_and_neurodevelopment

369

Results

De novo CNVs in cases of autism spectrum disorder disrupt UCE dosage.

Our first analysis compared how *de novo* and inherited CNVs in the genomes of (ASD) probands are positioned relative to UCEs, noting that the bulk of the CNV-mediated genetic etiology of ASD is thought to be borne by *de novo*, rather than inherited, CNVs[5,49,50,52,67-72]. *De novo* CNVs are defined as being present in the genomes of the proband but not either parent, while inherited CNVs are defined as identifiable in the genomes of the proband and at least one parent. From Vulto-van Silfhout *et al*.[51], we obtained 206 CNVs from patients with an ASD phenotype, of which 54 arose *de novo* and 94 were inherited; 58 were classed as "inheritance unknown" and were not examined further (see Figure S1 for details of filtering steps). When segments of overlapping *de novo* CNVs were combined, a final dataset of 45 distinct 'Vulto-van Silfhout ASD *de novo* CNVs' covering 5.91% of the genome was obtained (Table S1). Similarly, by merging overlapping inherited CNVs, we obtained 81 CNV regions covering 3.19% of the genome (Table S1), and we call these CNVs 'Vulto-van Silfhout ASD inherited CNVs'.

We queried our datasets of CNVs for their relationship to the positions of UCEs using an analysis pipeline that builds upon our previously published methods[33-35] (Methods, "Depletion and enrichment analysis"). The UCEs are first partitioned into intergenic, intronic, and exonic subsets of 301, 416, and 179 UCEs respectively[35], and then the pipeline operates on the three categories of UCEs separately. Briefly, we calculate a distribution of expected overlaps that reflects the overlaps that would arise between the CNVs and UCEs if the UCEs were distributed randomly within the

393 appropriate genomic subset (exonic UCEs being randomly placed within exons, for
394 example). The observed overlap between UCEs and CNVs is then compared to the
395 distribution of expected overlaps using a Z-test, with the outcome accompanied by a P-
396 value and ratio of observed/expected (obs/exp) overlaps. Occasionally, our distribution of
397 expected overlaps is not normally distributed, in which case we provide the proportion of
398 expected overlaps that equaled or exceeded the observed overlap.

399 Because we generate a bespoke expected overlap distribution for each CNV
400 dataset of interest, our analysis is suitable for analyzing and comparing datasets that
401 differ widely with respect to their genomic coverage, as is common for datasets of CNVs
402 and other structural variants. Additionally, our expected overlap distribution reflects the
403 division of UCEs between the intergenic, intronic, and exonic regions of the genome,
404 ensuring that any interesting differences between the observed and expected overlaps
405 cannot merely be due to overall trends in UCE occupancy of different genomic regions
406 (for example, 20% of the UCEs are exonic even though only approximately 3% of the
407 human genome is composed of exons[73]). As for our dataset of UCEs, we made use of a
408 previously defined set of 896 UCEs[33-35] that are ≥200bp in length and 100% identical in
409 DNA sequence between the reference genomes of human, mouse, and rat, or of human,
410 dog, and mouse, or of human and chicken (Table S1) all of which represent unique
411 sequences in the genome[33].

412 We began our study with the expectation that overall depletion of UCEs from a
413 set of CNVs is an indicator that such a set of CNVs is likely benign[33-35]. For example, our
414 most recent analysis of healthy individuals, including 8 datasets of predominantly
415 inherited CNVs covering 51.4% of the genome and 4 datasets of *de novo* CNVs covering

416 0.9% of the genome yielded obs/exp ratios of 0.771 ($P = 1.7 \times 10^{-12}$) and 0.395 (P =
417 0.044), respectively[35].

418 Here, we used a 2-tailed α of 0.025 and found that the Vulto-van Silfhout ASD
419 CNVs inherited from unaffected parents show statistically significant depletion (Figure
420 1A, Table S2A; P = 0.003, obs/exp = 0.493), with only 17 UCEs overlapping CNVs. This
421 CNV profile is similar to our findings for CNVs representing healthy individuals[35] (). In
422 stark contrast, Vulto-van Silfhout ASD *de novo* CNVs in affected probands are not
423 depleted of UCEs; rather they are enriched due to overlap with 68 UCEs (Figure 1A,
424 Table S2A; P = 0.024, obs/exp = 1.272). Further, the UCE overlap for Vulto-van Silfhout
425 *de novo* CNVs is significantly different from that of the Vulto-van Silfhout inherited
426 CNVs (Figure 1A, Table S2A; P = 0.002). Therefore, in the ASD patients examined by
427 Vulto-van Silfhout *et al.*[51], *de novo* CNVs in ASD probands affect an excess of UCEs as
428 compared to inherited CNVs. Finally, we examined the relationship between Vulto-van
429 Silfhout ASD *de novo* CNVs and intergenic, intronic, and exonic UCEs since our
430 previous studies had suggested that intergenic, intronic, and exonic UCEs may differ
431 somewhat in their relationship to CNVs. We found that the enrichment for UCEs is
432 driven by intronic elements, with 36 UCEs overlapping the *de novo* CNVs (Figure 1C,
433 Table S2A; P = 0.010, obs/exp = 1.487).

434 We next turned to data from Sanders *et al.*[74], which presents *de novo* CNV data
435 for autism simplex cases (where only one family member is affected) and their unaffected
436 siblings. We obtained 321 'Sanders ASD *de novo* CNVs' from ASD cases (see Figure S1
437 for details of filtering steps), which collapsed into 191 distinct non-overlapping CNV
438 regions that, together, covered 8.49% of the genome (Table S1). We likewise retained 86

439 CNVs from siblings of ASD cases (Figure S1), which collapsed into 79 non-overlapping
440 CNV regions covering 1.15% of the genome, and these CNVs comprise our 'Sanders
441 sibling *de novo* CNV' dataset (Table S1).

442 We found that Sanders ASD *de novo* CNVs and Sanders sibling *de novo* CNVs
443 both fail to show significant depletion of UCEs and thus deviate from the profile that is
444 generally obtained for CNVs representing healthy individuals. Specifically, Sanders ASD
445 *de novo* CNVs overlap 78 UCEs (Table S2A; P = 0.278, obs/exp = 0.938), and Sanders
446 sibling *de novo* CNVs overlap 7 UCEs (Table S2A; P = 0.056, obs/exp = 0.538), with the
447 difference between their degrees of overlap being insignificant (Table S2A; P = 0.123).
448 The result for Sanders sibling *de novo* CNVs presents an intriguing picture. On the one
449 hand, they do not show significant depletion of UCEs and thus resemble the Sanders
450 ASD *de novo* CNVs. On the other hand, their obs/exp ratio of 0.538 approaches the ratio
451 of 0.395 we observed in our previous analysis of *de novo* CNVs representing unaffected
452 individuals[35].

453 To further investigate differences between the Sanders sibling and ASD *de novo*
454 CNVs, we next examined intergenic, intronic and exonic UCEs separately. With respect
455 to intergenic UCEs, neither Sanders sibling *de novo* CNVs nor Sanders ASD *de novo*
456 CNVs show depletion (Table S2A; Proportion = 0.354, obs/exp = 1.197, P = 0.425,
457 obs/exp = 1.039 for Sanders sibling and ASD *de novo* CNVs, respectively). This may
458 suggest that the intergenic UCE profiles of both sets of CNVs are not commensurate with
459 that of unaffected individuals. For intronic and exonic UCEs, respectively, Sanders
460 sibling *de novo* CNVs showed low obs/exp values of 0.303 and 0.298 (Table S2A; P =
461 0.047 for intronic UCEs, proportion = 0.096 for exonic UCEs), while the analogous

462 values for Sanders ASD *de novo* CNVs were higher at 0.835 and 1.054, respectively
463 (Table S2A; P = 0.137 and P = 0.402). Similarly, Sanders sibling *de novo* CNVs are
464 depleted for all genic UCEs, that is, intronic and exonic UCEs, combined (Figure 1B,
465 Table S2A; P = 0.019, obs/exp = 0.302). In contrast, not only do Sanders ASD *de novo*
466 CNVs fail to show depletion of genic UCEs (Figure 1B, Table S2A, P = 0.212, obs/exp =
467 0.900) they overlap genic UCEs significantly more than do Sanders sibling *de novo*
468 CNVs (Figure 1B, Table S2A; P = 0.032). This suggests that, for UCEs in genes, *de novo*
469 CNVs in ASD probands show a greater propensity to disrupt UCE copy number than do
470 *de novo* CNVs in unaffected siblings.

471 In sum, we have found that disruption of UCE copy number by *de novo* CNVs is
472 elevated in two cohorts of individuals with ASD. In one cohort, *de novo* CNVs in ASD
473 probands disrupt UCEs more than do inherited CNVs in the same individuals while, in a
474 second cohort, *de novo* CNVs in ASD probands disrupt UCEs in genes more than do *de*
475 *novo* CNVs in their unaffected siblings.

UCEs are strongly enriched in proximity to the breakpoints of balanced rearrangements in subjects with neurodevelopmental phenotypes.

Given the importance of balanced rearrangements in the genetic etiology of NDDs,[55,56,75-77] we asked whether the connection between UCEs and NDD-associated genomic rearrangements is limited to CNVs or whether it extends to copy number neutral, or 'balanced', rearrangements. Defining balanced rearrangements as those that did not delete or duplicate more than 1 kb of sequence around each breakpoint, we drew our first set of balanced rearrangements drew from five publications that provided the information on breakpoints and phenotypes that is required for our filtration steps: Granot-Hershkovitz *et al*.[53], Kloosterman *et al*.[54], Chiang *et al*.[55], Talkowski *et al*.[56], and Nazaryan *et al*.[57] (see Methods, Table S1, and Figure S3 for filtering steps and a list of breakpoints with their respective publication sources).

As expected, no UCEs were positioned directly over any of the breakpoints (likelihood of overlap = 0.012; Methods). We therefore asked whether UCEs are enriched in the vicinity of the breakpoints by querying the 100kb genomic regions flanking either side of the breakpoints. A flank size of 100kb was chosen because it is the smallest flank size that that produces a normal distribution of expected overlaps and can thus provide a P-value for depletion or enrichment of UCEs. When flanks surrounding different breakpoints overlapped, we merged them into contiguous regions. The resulting dataset, called 'NDD breakpoints with 100kb flanks set 1' (Table S1, Figure S3) and encompassing 76 genomic regions and 0.51% of the genome, overlaps 22 UCEs and shows strong enrichment for UCEs (Figure 2A, Table S2B; $P = 2.22\times10^{-16}$, obs/exp = 4.795).

We next gathered a set of completely independent breakpoints, again in subjects with NDD and related phenotypes, from Redin *et al*.[60] (Methods, Figure S3). Again, as no direct overlaps between these breakpoints and UCEs were observed (likelihood of overlap = 0.034; Methods), 100kb flanks were added on either side of each breakpoint. The resulting dataset, called 'NDD breakpoints with 100kb flanks set 2' (Table S1, Figure S3) and comprising 224 regions and 1.52% of the genome, is also enriched in UCEs, overlapping 36 UCEs (Figure 2B, Table S2B; $P=5.7\times10^{-8}$, obs/exp = 2.429). Combining NDD breakpoint set 1 and 2 into 'Pooled NDD breakpoints' (Table S1, Figure S3) produced a dataset of 296 regions and covering 2.01% of the genome that is also significantly enriched for UCEs, overlapping 58 elements (Figure 2C, Table S2B; $P<1.0\times10^{-17}$, obs/exp = 3.049).

Because the size of the flanks placed around each breakpoint might affect the degree and significance of UCE enrichment, the Pooled NDD breakpoint dataset was analyzed with four additional flank sizes, two less than and two greater than the original 100kb to give the following series of flank sizes: 20kb, 50kb, 100kb, 500kb, and 1Mb (Table S1, Figure S3). Note that the larger number of breakpoints in the Pooled NDD breakpoints dataset produced normally distributed overlaps even with smaller flanks. With respect to the number of regions covered, the resulting datasets ranged from 316 regions for 20kb flanks to 219 for 1Mb flanks, the larger flanks giving rise to a smaller number of regions because they are more likely to overlap and thus be combined into contiguous regions. With respect to the percentage of genome covered, the datasets ranged from 0.42% for 20kb flanks to 16.59% for 1Mb flanks. In all cases, the datasets were significantly enriched for UCEs (Figure 2D, Table S2B; $1.0\times10^{-17} \leq P \leq 1.2\times10^{-10}$,

522 $1.719 \leq \text{obs/exp} \leq 4.367$). Interestingly, the highest obs/exp ratio was observed for the
523 smallest flank size (Figure 2D; obs/exp = 4.367 for 20kb flanks), and the obs/exp ratios
524 decreased as flank size increased, indicating that the enrichment of UCEs near Pooled
525 NDD breakpoints is most extreme when regions closest to the breakpoints themselves are
526 examined.

527 Our observations of enrichment suggest that UCEs and Pooled NDD breakpoints
528 tend to be more closely positioned than would be expected by chance. To explore this
529 further, we first calculated the distances from UCEs to the nearest Pooled NDD
530 breakpoint. We then obtained 1,000 sets of control genomic regions, matched to UCEs in
531 terms of number, length, and genomic region of interest (intergenic, intronic, or exonic)
532 using our analysis pipeline described previously. Next, for each of the 1,000 sets, we
533 calculated the distances from the control regions to the nearest breakpoint to produce a
534 distribution of expected distance measurements. We then compared the distributions of
535 observed (UCE-to-breakpoint) distances to the expected (control region-to-breakpoint
536 distances), using an Anderson-Darling test (Methods, "Distances from breakpoints to
537 UCEs"), and found the distributions to be significantly different (Figure 2E; $P = 2.7\times10^{-4}$),
538 reflecting that, at smaller distances, UCEs are more prevalent than are control regions
539 (Figure 2E). These findings argue that UCEs are located significantly closer to Pooled
540 NDD breakpoints than are randomly placed control regions (Figure 2E).

541 As the subjects described by Redin *et al*.[60] were classified into three categories
542 based on the predicted pathogenicity of their genomic rearrangements, we asked whether
543 balanced rearrangement breakpoints are differentially positioned with respect to UCEs
544 depending on the classification of the individual. The three categories are: 1)

545 'pathogenic', having strong evidence that the rearrangement is involved in NDD etiology;
546 2) 'likely pathogenic', having somewhat weaker evidence that the rearrangement is
547 involved in NDD etiology; and 3) variants of unknown significance ('VUS'). For each of
548 the three categories, we added 500kb flanks on either side of all breakpoints; flank sizes
549 of 50kb and 100kb were unable to produce normally distributed expected overlaps due to
550 the small number of breakpoints in each category. The resulting 'pathogenic', 'likely
551 pathogenic, and 'VUS' datasets respectively contained 32, 62, and 120 regions and
552 covered 1.04%, 2.11%, and 4.07% of the genome (Table S1; Figure S3). Regions
553 surrounding breakpoints in all three categories are enriched with UCEs, with the
554 pathogenic regions overlapping 25 UCEs (Figure 2F, Table S2B; $P = 4.1\times10^{-5}$, obs/exp =
555 2.281), likely pathogenic affecting 33 UCEs (Figure 2G, Table S2B; $P = 7.3\times10^{-4}$,
556 obs/exp = 1.732), and VUS encompassing 64 UCEs (Figure 2H, Table S2B; $P = 1.2\times10^{-6}$,
557 obs/exp = 1.797). That even variants of unknown significance display an association with
558 UCE position suggests that proximity to UCEs could be investigated as a novel way to
559 classify and understand these rearrangements.

560 Finally, we examined the breakpoints of rearrangements from Redin *et al.*[60] from
561 subjects that did not present with phenotypes related to neurodevelopment, but displayed,
562 instead, other congenital or developmental phenotypes (Methods). Calling this dataset
563 'non-NDD breakpoints', we added 500kb flanks on either side of each breakpoint, as
564 50kb and 100kb flanks did not produce a normally distributed set of expected overlaps
565 (Figure S3, Table S1). Interestingly, and unlike all datasets representing patients with
566 NDD and related phenotypes, this dataset containing 38 regions and covering 1.25% of
567 the genome is not enriched for UCEs (Figure 2I, Table S2B; P = 0.187, obs/exp = 0.721).

That not all copy number neutral breakpoints have elevated numbers of UCEs in their vicinity suggests that local enrichment for UCEs is not a universal feature of breakpoints, and that the enrichment of UCEs surrounding the Pooled NDD breakpoints may be a particular feature of subjects with NDD and related neurodevelopmental phenotypes. In sum, these studies of balanced rearrangements argue that association between UCE position and structural variants is not limited to those that disrupt UCE dosage.

Gene sets related to neurodevelopment do not explain the excess of UCEs affected by Vulto-van Silfhout ASD *de novo* CNVs or the enrichment of UCEs near breakpoints of balanced rearrangements.

We also considered the possibility that the enrichment for UCEs in our Vulto-van Silfhout ASD *de novo* CNV and Pooled NDD breakpoint datasets is driven by UCEs within or near specific subsets of genes. Five groups of genes were tested: 1) 'LoF genes', which are defined as those with two or more *de novo* loss-of-function mutations discovered across a cohort of NDD patients and considered by Redin *et al.*[60] as strong candidates for NDD causation or involvement; 2) 'Constrained genes', whose degree of loss-of-function variation in healthy individuals is less than would be predicted by models of mutation rates and thus are considered strong candidates for disease association[61]; 3) 'Embryonic brain genes', which are expressed specifically in the embryonic human brain and thus are considered a proxy for genes involved in neurodevelopment[62]; 4) 'SZ genes', which are associated with schizophrenia through genome-wide association studies[64]; 5) 'T2D genes', which are associated with type 2 diabetes through genome wide association studies[63]. The LoF and Embryonic brain genes

were chosen because they are directly connected to processes and pathways relevant to neurodevelopment, and the Constrained and SZ genes were chosen because they have been previously documented to overlap with genes involved in NDD[56,61,70]. The T2D genes were chosen as a control for our studies, as they represent a disease that is not thought to involve genomic regions relevant to neurodevelopment. As described below, our studies involved three steps.

Our first step of analysis assessed the frequency with which UCEs are found within the five gene sets, considering intronic and exonic portions of the genome only, using our previously described pipeline (Methods, 'Depletion and enrichment analysis with gene sets excluded'). We first found that LoF, Constrained, and Embryonic brain gene sets are significantly enriched for UCEs (Figure 3A, Table S2C; $P < 1.0\times10^{-17}$, obs/exp = 4.377 for LoF genes, $P < 1.0\times10^{-17}$, obs/exp = 1.910 for Constrained genes, and $P = 3.7\times10^{-8}$, obs/exp = 1.704 for Embryonic brain genes). For SZ genes, we observed nominal enrichment (Figure 3A, Table S2C; P = 0.025, obs/exp = 1.419). T2D genes do not show significant enrichment, albeit with an obs/exp ratio above one (Figure 3A, Table S2C; P = 0.156, obs/exp = 1.467).

Because the LoF, Constrained, and Embryonic brain genes display the strongest enrichment for UCEs, we reasoned that they are the most likely to drive enrichment in Vulto-van Silfhout ASD *de novo* CNVs. Here, the genomic regions occupied by, and extending 100kb on either side of, LoF, Constrained, and Embryonic brain gene sets were excluded, both in turn and all at once, from our standard analysis pipeline (Methods 'Depletion and enrichment analysis'). If the gene sets were responsible for enrichment, then exclusion of these genes and their flanking regions from our analyses would result in

614 loss of enrichment. If such a result were obtained, our studies would further reveal
615 whether the Vulto-van Silfhout ASD *de novo* CNVs would, under these circumstances,
616 resemble that of unaffected individuals and be depleted of UCEs.
617 Our studies showed that excluding LoF genes and their 100kb flanks caused UCE
618 enrichment to be lost from Vulto-van Silfhout ASD *de novo* CNVs (Figure 3B, Table
619 S2D; P = 0.111, obs/exp = 1.174). A similar outcome was obtained for Constrained genes.
620 (Figure 3B, Table S2D; P = 0.042, obs/exp = 1.306). For Embryonic brain genes and
621 their 100kb flanks, enrichment remained (Figure 3B, Table S2D; P = 0.023, obs/exp =
622 1.294). When we removed all three gene sets at once, including 100kb on either side of
623 all genes, our expected overlaps were not normally distributed, and so we did not
624 calculate a P-value using a Z-test. Instead, we found that the proportion of expected
625 overlaps that equaled or exceeded the observed overlap is inconsistent with enrichment
626 (Figure 3B, Table S2D; Proportion = 0.166, obs/exp = 1.191), as is the result of a χ^2 test
627 to discern whether the number of UCEs overlapped differs from expectation calculated
628 from the genome coverage of the CNVs (Table S2D; P = 0.676). Enrichment was also
629 lost when we excluded the gene sets, either alone or in combination, in all cases without
630 excluding the regions 100kb on either side of the genes (Table S2D). Taken together,
631 these outcomes indicate that the enrichment of UCEs in Vulto-van Silfhout ASD *de novo*
632 CNVs is due in large part to UCEs inside and within 100kb of LoF and Constrained
633 genes. Importantly, however, none of these studies caused the profile of CNVs to be
634 depleted of UCEs (obs/exp $\geq$ 1.095 in all cases), reinforcing our observations that Vulto-
635 van Silfhout ASD *de novo* CNVs are distinct from CNV datasets representing unaffected

individuals. As such, the excess of UCEs affected by the Vulto-van Silfhout ASD *de novo* CNVs continues to suggest an importance of these UCEs in neurodevelopment.

Finally, we turned to the Pooled NDD breakpoints with 100kb flanks. Here, we found that enrichment for UCEs remained after exclusion of LoF, Constrained, or Embryonic brain genes, or all three sets combined, together with their 100kb flanks. While these analyses resulted in larger P-values and smaller obs/exp ratios, suggesting that UCEs in and flanking these genes do contribute to enrichment, they nevertheless demonstrate that such UCEs do not, alone, explain enrichment (Figure 3C, Table S2E; P = 0.002, obs/exp = 1.756 for LoF genes, $P = 1.6\times10^{-4}$, obs/exp = 2.231 for Constrained genes, $P = <1.0\times10^{-17}$, obs/exp = 3.143 for Embryonic brain genes, and P = 0.001, obs/exp = 2.031 for all three sets combined). Enrichment was also maintained when we excluded just the gene sets, alone or in combination, but did not extend the exclusion 100kb on either side of the genes (Table S2E). In summary, the enrichment of UCEs in the genomic regions around the breakpoints of balanced rearrangements associated with NDD cannot be explained by UCEs residing in, or within 100kb of, genes implicated in NDD.

Positive correlation between UCEs and both LoF genes and Pooled NDD breakpoints is robust to other genomic features.

We next queried the underlying basis for the enrichment of UCEs in Vulto-van Silfhout ASD *de novo* CNVs (Figure 1A and 3B), LoF genes (Figure 3A), and the vicinity of Pooled NDD breakpoints (Figure 2C and 3C). In particular, partial correlation analyses (Methods) were implemented to examine the contributions of four genomic

features: enhancers, regions of the genome that have experienced accelerated sequence change, known as human accelerated regions[78] (HARs; described further in Discussion), repetitive elements, and GC percentage. We included enhancers because they can overlap UCEs[12,13,17] and HARs because they have been associated with NDD[66]. GC content and repetitive elements were included because they can affect the formation of structural variants[79,80]. Dividing the genome into 100kb bins and considering $\alpha = 0.01$ due to Bonferroni correction for 5 tests, we found that, for Vulto-van Silfhout ASD *de novo* CNVs, correlation with UCEs remained significant when controlling for co-correlation with enhancers, HARs, and GC percentage, but was lost when controlling for repetitive elements either alone or in combination with enhancers and HARs (Figure 3D). These results suggest that the association between UCEs and Vulto-van Silfhout ASD *de novo* CNVs is robust to the positions of enhancers, HARs, and to GC percentage, but not to repetitive element position, possibly due to the fact that UCEs are anti-correlated with repetitive elements[33] and CNVs are positively correlated with repetitive elements[79].

Having observed an impressive enrichment for UCEs in LoF genes, additional analyses were conducted to address the potential contribution of other genomic features in this context, too. The correlation between UCEs and LoF genes remained significant in all cases, including when the enhancers, HARs, and repetitive elements were combined together. Finally, we examined Pooled NDD breakpoints with 100kb flanks, and found that correlation with UCEs was again significant when controlling for co-correlation with enhancers, HARs, repetitive elements, these three features combined, and GC percentage. This finding indicates that the enrichment for UCEs near Pooled NDD breakpoints is not

681 obviously influenced by the other genomic features we examined and, instead, may be an

682 inherent feature of the position of UCEs.

683

Discussion

The present study documents an excess of UCEs in regions affected by ASD *de novo* CNVs and an enrichment of UCEs in the vicinity of balanced rearrangement breakpoints in individuals with NDD and related phenotypes. UCEs have long been associated with gene regulatory functions, particularly for developmentally important genes[12-14] (reviewed in Baira *et al*.[81], Harmston *et al*.[32], and Fabris *et al*.[82]). Indeed, our results point to the importance of considering longer range interactions (over 100kb) and spatial genome organization[83,84] when exploring how UCEs may regulate genes in neurodevelopment.

We[33-35] and others[31,36] have also speculated on a different and non-mutually exclusive model for UCE conservation, wherein UCEs contribute to genome integrity. Here, cells would assess the integrity of UCE-containing genomic regions, perhaps by physically pairing allelic UCEs such that disruptions of pairing compromise cell viability and even organismal fitness. As such, *c*ellular *a*ssessment and *r*esponse to *d*isruption (CARD) would provide selective advantage to organisms by reducing their burden of deleterious chromosomal changes[33-35]. Indeed, such a mechanism is consistent with, and may help to explain, the enrichment of UCEs in NDD-relevant genomic regions (Figure S4). Specifically, elimination of deleterious mutations would be most advantageous in regions that are functionally important and/or highly vulnerable to mutation, and NDD-associated regions meet both criteria; the functional importance of NDD-associated regions is evidenced by the severity of many NDD phenotypes, while the vulnerability of genes involved in neuronal and brain development and function is exacerbated by their

706 longer than average length[85-87], subjecting them to increased replication stress and thus
707 chromosome rearrangements (Wilson *et al*. [88] and references therein).

708 Curiously, a recent study has reported that NDD-relevant regions contain an
709 excess of sequences, known as HARs, that had been highly conserved and then, in
710 humans, experienced accelerated rates of change[66]. How might this paradoxical situation
711 arise, wherein a highly conserved sequence, perhaps a UCE, suddenly begins to change at
712 an accelerated rate? We note first that mutations to UCEs can, in fact, acquire SNPs[89-91],
713 suggesting that evolution can in some cases erode UCEs. Secondly, heterozygosity has
714 been correlated in certain studies with increased mutation rates, with pairing considered
715 as a possible mechanism for the detection of heterozygosity in some instances[92-94]. Thus,
716 in the case that heterozygosity can increase mutation rate and, in addition, considering
717 that the first instance of an altered UCE will almost certainly result in heterozygosity, this
718 initial mutation might be accompanied by accelerated further mutation of the region. In
719 any case, should alteration of a UCE produce an advantageous function, such as a new
720 regulatory activity, selection for the new function may ultimately release that UCE
721 sequence from the its previous constraints, creating a HAR. As this process might be
722 expected to be most prevalent in genomic regions where UCEs are enriched, such as
723 NDD-relevant regions, it could explain the excess of HARs in these same regions.

724 To conclude, our results demonstrate an association between neurodevelopmental
725 phenotypes and an elevated level of structural variation affecting UCE dosage or genomic
726 position. Interestingly, a positive correlation has been observed between NDD and an
727 elevated risk for cancer (Norwood *et al*.[95] and references therein), perhaps reflecting that
728 disruption of UCEs can misregulate genes involved in development and oncogenesis. In

729 light of our previous observations that cancer-specific CNVs are enriched in UCEs[35], it
730 may also be that individuals with NDDs would lack the proposed safeguard of intact and
731 correctly positioned UCEs[33-35] later in life, and therefore show higher cancer prevalence.
732
733
734

Note:

During preparation of this manuscript, two manuscripts of relevance were released on the bioarchive preprint server. Firstly, Short *et al*. [96] report an enrichment of *de novo* single nucleotide variants (SNVs) in conserved, putatively regulatory, noncoding genomic regions in the genomes of probands with developmental disorders. Secondly, Werling *et al*. [97] describe an intriguing excess of *de novo* SNVs and small (<50bp) insertion-deletion (indel) variants within conserved regions in ASD cases compared to controls, though the statistical significance of this observation did not survive multiple testing correction.

744 **Supplementary Data**

745 The supplementary data consists of four figures and three excel tables.

746

747 **Acknowledgements**

748 We thank D. Balick, R. Collins, S. Erdin, K. Mattioli, V. Pillalamarri, S. Sunyaev,
749 R. Tarnita, T. Tullius, D. Weghorn, and all members of the Wu laboratory for helpful and
750 insightful discussions. This work was supported by a William Randolph Hearst
751 Foundation Award to R.B.M., an EMBO Long Term Fellowship to J.E., awards to M.E.T.
752 from NIH (MH095867 and GM061354), the Simons Foundation for Autism Research
753 (SFARI #346042), and the March of Dimes, and awards to Ct.W. from NIH
754 (DP1GM106412; R01HD091797) and Harvard Medical School.

755 The authors declare that no conflicts of interest exist.

756 We apologize to authors whose work we were unable to cite due to limits on
757 citation number.

758

759 **Web resources**

760 Custom python scripts associated with this study can be accessed at
761 https://github.com/rmccole/UCEs_and_neurodevelopment

References

1. Huguet, G., Ey, E., and Bourgeron, T. (2013). The genetic landscapes of autism spectrum disorders. Annu Rev Genomics Hum Genet *14*, 191–213.

2. Kloosterman, W.P., and Cuppen, E. (2013). Chromothripsis in congenital disorders and cancer: similarities and differences. Curr Opin Cell Biol *25*, 341–348.

3. Hu, W.F., Chahrour, M.H., and Walsh, C.A. (2014). The diverse genetic landscape of neurodevelopmental disorders. Annu Rev Genomics Hum Genet *15*, 195–213.

4. Sahin, M., and Sur, M. (2015). Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders. Science *350*, aab3897–aab3897.

5. Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet *15*, 133–141.

6. Georgi, B., Voight, B.F., and Bucan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genetics *9*, e1003484.

7. Hadley, D., Murphy, T., Valladares, O., Hannenhalli, S., Ungar, L., Kim, J., and Bucan, M. (2006). Patterns of sequence conservation in presynaptic neural genes. Genome Biol *7*, R105.

8. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

9. Sultana, R., Yu, C.-E., Yu, J., Munson, J., Chen, D., Hua, W., Estes, A., Cortes, F., la Barra, de, F., Yu, D., et al. (2002). Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. Genomics *80*, 129–134.

10. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., and Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. Nature *418*, 869–872.

11. Dias, C., Estruch, S.B., Graham, S.A., McRae, J., Sawiak, S.J., Hurst, J.A., Joss, S.K., Holder, S.E., Morton, J.E.V., Turner, C., et al. (2016). BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. Am J Hum Genet *99*, 253–274.

12. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

13. Sandelin, A., Bailey, P., Bruce, S., Engström, P.G., Klos, J.M., Wasserman, W.W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span

798 the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

799 14. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith,
800 S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding
801 sequences are associated with vertebrate development. PLoS Biol *3*, e7.

802 15. Poulin, F., Nobrega, M.A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E.M., and
803 Pennacchio, L.A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer.
804 Genomics *85*, 774–781.

805 16. Lampe, X., Samad, O.A., Guiguen, A., Matis, C., Remacle, S., Picard, J.J., Rijli, F.M.,
806 and Rezsohazy, R. (2008). An ultraconserved Hox-Pbx responsive element resides in the
807 coding sequence of Hoxa2 and is active in rhombomere 4. *36*, 3214–3225.

808 17. Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-
809 Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2008). Ultraconservation
810 identifies a small subset of extremely constrained developmental enhancers. Nat Genet *40*,
811 158–160.

812 18. Dong, X., Navratilova, P., Fredman, D., Drivenes, Ø., Becker, T.S., and Lenhard, B.
813 (2010). Exonic remnants of whole-genome duplication reveal cis-regulatory function of
814 coding exons. Nucleic Acids Research *38*, 1071–1085.

815 19. Jaeger, S.A., Chan, E.T., Berger, M.F., Stottmann, R., Hughes, T.R., and Bulyk, M.L.
816 (2010). Conservation and regulatory associations of a wide affinity range of mouse
817 transcription factor binding sites. Genomics *95*, 185–195.

818 20. Lujambio, A., Portela, A., Liz, J., Melo, S.A., Rossi, S., Spizzo, R., Croce, C.M.,
819 Calin, G.A., and Esteller, M. (2010). CpG island hypermethylation-associated silencing
820 of non-coding RNAs transcribed from ultraconserved regions in human cancer.
821 Oncogene *29*, 6390–6401.

822 21. Poitras, L., Yu, M., Lesage-Pelletier, C., Macdonald, R.B., Gagné, J.-P., Hatch, G.,
823 Kelly, I., Hamilton, S.P., Rubenstein, J.L.R., Poirier, G.G., et al. (2010). An SNP in an
824 ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain.
825 Development *137*, 3089–3097.

826 22. Braconi, C., Valeri, N., Kogure, T., Gasparini, P., Huang, N., Nuovo, G.J.,
827 Terracciano, L., Croce, C.M., and Patel, T. (2011). Expression and functional role of a
828 transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma.
829 Proc Natl Acad Sci USA *108*, 786–791.

830 23. McBride, D.J., Buckle, A., van Heyningen, V., and Kleinjan, D.A. (2011). DNaseI
831 Hypersensitivity and Ultraconservation Reveal Novel, Interdependent Long-Range
832 Enhancers at the Complex Pax6 Cis-Regulatory Region. PLoS ONE *6*, e28616.

833 24. Sana, J., Hankeova, S., Svoboda, M., Kiss, I., Vyzula, R., and Slaby, O. (2012).
834 Expression Levels of Transcribed Ultraconserved Regions uc.73 and uc.388 Are Altered

in Colorectal Cancer. Oncology *82*, 114–118.

25. Bhatia, S., Bengani, H., Fish, M., Brown, A., Divizia, M.T., de Marco, R., Damante, G., Grainger, R., van Heyningen, V., and Kleinjan, D.A. (2013). Disruption of Autoregulatory Feedback by a Mutation in a Remote, Ultraconserved PAX6 Enhancer Causes Aniridia. Am J Hum Genet *93*, 1126–1134.

26. Ferdin, J., Nishida, N., Wu, X., Nicoloso, M.S., Shah, M.Y., Devlin, C., Ling, H., Shimizu, M., Kumar, K., Cortez, M.A., et al. (2013). HINCUTs in cancer: hypoxia-induced noncoding ultraconserved transcripts. Cell Death Differ. *20*, 1675–1687.

27. Hudson, R.S., Yi, M., Volfovsky, N., Prueitt, R.L., Esposito, D., Volinia, S., Liu, C.-G., Schetter, A.J., Roosbroeck, K., Stephens, R.M., et al. (2013). Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer. Mol. Cancer *12*, 13.

28. Liz, J., Portela, A., Soler, M., Gómez, A., Ling, H., Michlewski, G., Calin, G.A., Guil, S., and Esteller, M. (2014). Regulation of pri-miRNA Processing by a Long Noncoding RNA Transcribed from an Ultraconserved Region. Mol Cell *55*, 138–147.

29. Silla, T., Kepp, K., Tai, E.S., Goh, L., Davila, S., Catela Ivkovic, T., Calin, G.A., and Voorhoeve, P.M. (2014). Allele frequencies of variants in ultra conserved elements identify selective pressure on transcription factor binding. PLoS ONE *9*, e110692.

30. Olivieri, M., Ferro, M., Terreri, S., Durso, M., Romanelli, A., Avitabile, C., De Cobelli, O., Messere, A., Bruzzese, D., Vannini, I., et al. (2016). Long non-coding RNA containing ultraconserved genomic region 8 promotes bladder cancer tumorigenesis. Oncotarget *7*, 2063620654.

31. Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet *24*, 344–352.

32. Harmston, N., Baresic, A., and Lenhard, B. (2013). The mystery of extreme non-coding conservation. Philos Trans R Soc Lond B Biol Sci *368*, 20130021.

33. Derti, A., Roth, F.P., Church, G.M., and Wu, C.-T. (2006). Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet *38*, 1216–1220.

34. Chiang, C.W.K., Derti, A., Schwartz, D., Chou, M.F., Hirschhorn, J.N., and Wu, C.T. (2008). Ultraconserved Elements: Analyses of Dosage Sensitivity, Motifs and Boundaries. Genetics *180*, 2277–2293.

35. McCole, R.B., Fonseka, C.Y., Koren, A., and Wu, C.-T. (2014). Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. PLoS Genetics *10*, e1004646.

36. Kritsas, K., Wuest, S.E., Hupalo, D., Kern, A.D., Wicker, T., and Grossniklaus, U.

871 (2012). Computational analysis and characterization of UCE-like elements (ULEs) in
872 plant genomes. Genome Research *22*, 2455–2466.

873 37. Sun, H., Skogerbø, G., and Chen, R. (2006). Conserved distances between vertebrate
874 highly conserved elements. Hum Mol Genet *15*, 2911–2922.

875 38. Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G.,
876 Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007). Genomic
877 regulatory blocks encompass multiple neighboring genes and maintain conserved synteny
878 in vertebrates. Genome Res *17*, 545–555.

879 39. Dimitrieva, S., and Bucher, P. (2012). Genomic context analysis reveals dense
880 interaction network between vertebrate ultraconserved non-coding elements.
881 Bioinformatics *28*, i395–i401.

882 40. Polychronopoulos, D., Sellis, D., and Almirantis, Y. (2014). Conserved noncoding
883 elements follow power-law-like distributions in several genomes as a result of genome
884 dynamics. PLoS ONE *9*, e95437.

885 41. Polychronopoulos, D., Athanasopoulou, L., and Almirantis, Y. (2016). Fractality and
886 entropic scaling in the chromosomal distribution of conserved noncoding elements in the
887 human genome. Gene 148–160.

888 42. Robyr, D., Friedli, M., Gehrig, C., Arcangeli, M., Marin, M., Guipponi, M., Farinelli,
889 L., Barde, I., Verp, S., Trono, D., et al. (2011). Chromosome conformation capture
890 uncovers potential genome-wide interactions between human conserved non-coding
891 sequences. PLoS ONE *6*, e17634.

892 43. Joyce, E.F., Erceg, J., and Wu, C.-T. (2016). Pairing and anti-pairing: a balancing act
893 in the diploid genome. Curr Opin Genet Dev *37*, 119–128.

894 44. Hammond, T.M. (2017). Sixteen Years of Meiotic Silencing by Unpaired DNA.
895 Advances in Genetics *97*, 1–42.

896 45. Gladyshev, E. (2017). Repeat-Induced Point Mutation and Other Genome Defense
897 Mechanisms in Fungi. Microbiol Spectr *5*.

898 46. Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M.
899 (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol *5*, e234.

900 47. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J.,
901 Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of
902 copy number variation in the human genome. Nature *464*, 704–712.

903 48. Martínez, F., Monfort, S., Roselló, M., Oltra, S., Blesa, D., Quiroga, R., Mayo, S.,
904 and Orellana, C. (2010). Enrichment of ultraconserved elements among genomic
905 imbalances causing mental delay and congenital anomalies. BMC Med Genomics *3*, 54.

906 49. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom,
907 B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy
908 number mutations with autism. Science *316*, 445–449.

909 50. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom,
910 B., Lee, Y.-H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children
911 on the autistic spectrum. Neuron *74*, 285–299.

912 51. Vulto-van Silfhout, A.T., Hehir-Kwa, J.Y., van Bon, B.W.M., Schuurs-Hoeijmakers,
913 J.H.M., Meader, S., Hellebrekers, C.J.M., Thoonen, I.J.M., de Brouwer, A.P.M., Brunner,
914 H.G., Webber, C., et al. (2013). Clinical significance of de novo and inherited copy-
915 number variation. Hum Mutat *34*, 1679–1687.

916 52. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey,
917 A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De
918 novo mutations revealed by whole-exome sequencing are strongly associated with autism.
919 Nature *485*, 237–241.

920 53. Granot-Hershkovitz, E., Raas-Rothschild, A., Frumkin, A., Granot, D., Silverstein, S.,
921 and Abeliovich, D. (2011). Complex chromosomal rearrangement in a girl with
922 psychomotor-retardation and a de novo inversion: inv(2)(p15;q24.2). Am J Med Genet A
923 *155A*, 1825–1832.

924 54. Kloosterman, W.P., Guryev, V., van Roosmalen, M., Duran, K.J., de Bruijn, E.,
925 Bakker, S.C.M., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M., et al.
926 (2011). Chromothripsis as a mechanism driving complex de novo structural
927 rearrangements in the germline. Human Molecular Genetics *20*, 1916–1924.

928 55. Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills,
929 R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R., et al. (2012). Complex reorganization
930 and predominant non-homologous repair following chromosomal breakage in
931 karyotypically balanced germline rearrangements and transgenic integration. Nat Genet
932 *44*, 390–7–S1.

933 56. Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C.,
934 Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M., et al. (2012).
935 Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer
936 Risk across Diagnostic Boundaries. Cell *149*, 525–537.

937 57. Nazaryan, L., Stefanou, E.G., Hansen, C., Kosyakova, N., Bak, M., Sharkey, F.H.,
938 Mantziou, T., Papanastasiou, A.D., Velissariou, V., Liehr, T., et al. (2014). The strength
939 of combined cytogenetic and mate-pair sequencing techniques illustrated by a germline
940 chromothripsis rearrangement involving FOXP2. Eur J Hum Genet *22*, 338–343.

941 58. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for
942 comparing genomic features. Bioinformatics *26*, 841–842.

943 59. Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python

944 library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423–
945 3424.

946 60. Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom,
947 C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., et al. (2017). The genomic landscape of
948 balanced cytogenetic abnormalities associated with human congenital anomalies. Nat
949 Genet *49*, 36–45.

950 61. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T.,
951 O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of
952 protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

953 62. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D.,
954 Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The
955 contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–
956 221.

957 63. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir,
958 V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale
959 association analysis provides insights into the genetic architecture and pathophysiology
960 of type 2 diabetes. Nat Genet *44*, 981–990.

961 64. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A.,
962 Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated
963 resource of SNP-trait associations. Nucleic Acids Research *42*, D1001–D1006.

964 65. ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D.,
965 Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the
966 human genome. Nature *489*, 57–74.

967 66. Doan, R.N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S.,
968 Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., et al. (2016). Mutations in Human
969 Accelerated Regions Disrupt Cognition and Social Behavior. Cell *167*, 1–14.

970 67. Jacquemont, M.-L., Sanlaville, D., Redon, R., Raoul, O., Cormier-Daire, V., Lyonnet,
971 S., Amiel, J., Le Merrer, M., Heron, D., de Blois, M.-C., et al. (2006). Array-based
972 comparative genomic hybridisation identifies high frequency of cryptic chromosomal
973 rearrangements in patients with syndromic autism spectrum disorders. J Med Genet *43*,
974 843–849.

975 68. Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P.,
976 van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm
977 for mental retardation. Nat Genet *42*, 1109–1112.

978 69. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J.,
979 Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global
980 rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.

981 70. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M.,
982 Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the
983 interpretation of de novo mutation in human disease. Nat Genet *46*, 944–950.

984 71. Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., Marks, S.,
985 Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare de novo and transmitted copy-number
986 variation in autistic spectrum disorders. Neuron *70*, 886–897.

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

991 73. Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006).
992 The UCSC Known Genes. Bioinformatics *22*, 1036–1046.

993 74. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek,
994 A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism
995 Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*,
996 1215–1233.

997 75. Brand, H., Pillalamarri, V., Collins, R.L., Eggert, S., O'Dushlaine, C., Braaten, E.B.,
998 Stone, M.R., Chambert, K., Doty, N.D., Hanscom, C., et al. (2014). Cryptic and complex
999 chromosomal aberrations in early-onset neuropsychiatric disorders. Am J Hum Genet *95*,
1000 454–461.

1001 76. van Heesch, S., Simonis, M., van Roosmalen, M.J., Pillalamarri, V., Brand, H., Kuijk,
1002 E.W., de Luca, K.L., Lansu, N., Braat, A.K., Menelaou, A., et al. (2014). Genomic and
1003 Functional Overlap between Somatic and Germline Chromosomal Rearrangements. Cell
1004 Reports *9*, 1–10.

1005 77. Vergult, S., van Binsbergen, E., Sante, T., Nowak, S., Vanakker, O., Claes, K., Poppe,
1006 B., Van der Aa, N., van Roosmalen, M.J., Duran, K., et al. (2014). Mate pair sequencing
1007 for the detection of chromosomal aberrations in patients with intellectual disability and
1008 congenital malformations. Eur J Hum Genet *22*, 652–659.

1009 78. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S.,
1010 Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed
1011 during cortical development evolved rapidly in humans. Nature *443*, 167–172.

1012 79. Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and
1013 McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different
1014 forms of human mutation and variation. Am J Hum Genet *91*, 1033–1040.

1015 80. Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant
1016 formation in genomic disorders. Nat Rev Genet *17*, 224–238.

1017 81. Baira, E., Greshock, J., Coukos, G., and Zhang, L. (2008). Ultraconserved elements:

1018 genomics, function and disease. RNA Biol *5*, 132–134.

1019 82. Fabris, L., and Calin, G.A. (2017). Understanding the Genomic Ultraconservations:
1020 T-UCRs and Cancer. Int Rev Cell Mol Biol *333*, 159–172.

1021 83. Zepeda-Mendoza, C.J., Ibn-Salem, J., Kammin, T., Harris, D.J., Rita, D., Gripp, K.W.,
1022 MacKenzie, J.J., Gropman, A., Graham, B., Shaheen, R., et al. (2017). Computational
1023 Prediction of Position Effects of Apparently Balanced Human Chromosomal
1024 Rearrangements. Am J Hum Genet *101*, 206–217.

1025 84. Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How
1026 Alterations of Chromatin Domains Result in Disease. Trends Genet *32*, 225–237.

1027 85. Smith, D.I., Zhu, Y., McAvoy, S., and Kuhn, R. (2006). Common fragile sites,
1028 extremely large genes, neural development and cancer. Cancer Lett. *232*, 48–57.

1029 86. King, I.F., Yandava, C.N., Mabb, A.M., Hsiao, J.S., Huang, H.-S., Pearson, B.L.,
1030 Calabrese, J.M., Starmer, J., Parker, J.S., Magnuson, T., et al. (2013). Topoisomerases
1031 facilitate transcription of long genes linked to autism. Nature *501*, 58–62.

1032 87. Wei, P.-C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M., Alt, F.W., and Schwer, B.
1033 (2016). Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural
1034 Stem/Progenitor Cells. Cell *164*, 644–655.

1035 88. Wilson, T.E., Arlt, M.F., Park, S.H., Rajendran, S., Paulsen, M., Ljungman, M., and
1036 Glover, T.W. (2015). Large transcription units unify copy number variants and common
1037 fragile sites arising under replication stress. Genome Research *25*, 189–200.

1038 89. Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama,
1039 S.R., and Haussler, D. (2007). Human genome ultraconserved elements are ultraselected.
1040 Science *317*, 915.

1041 90. Halligan, D.L., Oliver, F., Guthrie, J., Stemshorn, K.C., Harr, B., and Keightley, P.D.
1042 (2011). Positive and negative selection in murine ultra-conserved noncoding elements. *28*,
1043 2651–2660.

1044 91. Chiang, C.W.K., Liu, C.-T., Lettre, G., Lange, L.A., Jorgensen, N.W., Keating, B.J.,
1045 Vedantam, S., Nock, N.L., Franceschini, N., Reiner, A.P., et al. (2012). Ultraconserved
1046 Elements in the Human Genome: Association and Transmission Analyses of Highly
1047 Constrained SNPs. Genetics *192*, 253–266.

1048 92. Yang, H., Zhong, Y., Peng, C., Chen, J.-Q., and Tian, D. (2010). Important role of
1049 indels in somatic mutations of human cancer genes. BMC Med. Genet. *11*, 128.

1050 93. Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L.D., and
1051 Tian, D. (2015). Parent-progeny sequencing indicates higher mutation rates in
1052 heterozygotes. Nature *523*, 463–467.

1053 94. Amos, W. (2016). Heterozygosity increases microsatellite mutation rate. Biol Lett *12*,
1054 20150929.

1055 95. Norwood, M.S., Lupo, P.J., Chow, E.J., Scheurer, M.E., Plon, S.E., Danysh, H.E.,
1056 Spector, L.G., Carozza, S.E., Doody, D.R., and Mueller, B.A. (2017). Childhood cancer
1057 risk in those with chromosomal and non-chromosomal congenital anomalies in
1058 Washington State: 1984-2013. PLoS ONE *12*, e0179006.

1059 96. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright,
1060 C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. De novo mutations in regulatory
1061 elements cause neurodevelopmental disorders. *10.1101/112896*.

1062 97. Werling, D.M., Brand, H., An, J.-Y., Stone, M.R., Glessner, J.T., Zhu, L., Collins,
1063 R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E.-C., et al. Limited
1064 contribution of rare, noncoding variation to autism spectrum disorder from sequencing of
1065 2,076 genomes in quartet families. *http://biorxiv.org/lookup/doi/10.1101/127043*.

1066

1067

Figure legends

Figure 1: *De novo* CNVs in cases of autism spectrum disorder disrupt UCE dosage.

(A): UCEs are depleted from Vulto-van Silfhout ASD inherited CNVs and enriched in Vulto-van Silfhout ASD *de novo* CNVs. Vulto-van Silfhout ASD *de novo* CNVs are represented in lilac throughout the figures.

(B): UCEs in genes are depleted from Sanders sibling *de novo* CNVs, and not depleted from Sanders ASD *de novo* CNVs. A X^2 test for the observed and mean expected numbers of UCE overlaps in both conditions shows a significant difference.

(C): Enrichment of UCEs in Vulto-van Silfhout ASD *de novo* CNVs is driven by enrichment of intronic UCEs; intergenic and exonic UCEs were not significantly enriched.

All panels: Large dots indicate observed overlap of UCEs, red for enriched, blue for depleted, and black for neither. Boxes indicate median and interquartile range for overlap with UCE-matched random regions, whiskers indicate 1.5× interquartile range, with all other points plotted as outliers. Brackets: A X^2 test for the observed and mean expected numbers of UCE overlaps in both conditions shows a significant difference.

Figure 2: UCEs are enriched in proximity to NDD breakpoints.

(A-C): UCEs are enriched in the regions extending 100kb on either side of NDD breakpoints set 1 (A), independent NDD breakpoints set 2 (B) and the combination of the two sets (C). Breakpoints shown in (C) are referred to as 'Pooled NDD breakpoints', and analyses using this dataset are in orange throughout. Histogram bars show distribution of

expected overlaps, red bars show observed overlap between UCEs and breakpoint flanking regions.

(D): UCEs are enriched around pooled NDD breakpoints when five different flank sizes are chosen. P values for the analysis are displayed above the bars.

(E): UCEs (orange) are closer to breakpoints than random regions matched to UCEs (grey: inter-quartile range for 1,000 iterations of random regions matched to UCEs). P value is for Anderson-Darling test of distribution of distances from UCEs to breakpoints compared with the distribution of random regions to breakpoints, with a maximum distance of 5.2Mb.

(F-H): UCEs were enriched in the regions extending 500kb on either side of breakpoints from set 2, whether the subject is designated as having a 'pathogenic' variant (F),'likely pathogenic' variant (G), or 'variant of unknown significance' (H).

(I): UCEs were not enriched in the regions extending 500kb on either side of NDD breakpoints described by Redin *et al*. [60] that came from subjects with phenotypes unrelated to neurodevelopment.

Figure 3: Gene sets related to neurodevelopment do not explain the excess of UCEs affected by Vulto-van Silfhout ASD *de novo* CNVs or the enrichment of UCEs near Pooled NDD breakpoints.

(A) UCEs are enriched in genes with 2 or more loss-of-function mutations in ASD subjects (LoF, shown in green in panels A and D), genes with high loss-of-function intolerance scores (constrained), and genes specifically expressed in embryonic brain.

1112 UCEs were nominally enriched in genes associated with schizophrenia (SZ) and not
1113 enriched in genes associated with type 2 diabetes (T2D). Bar height shows obs/exp ratio.
1114 P values are displayed above the bars. Red indicates significant enrichment.
1115 (B) UCEs are enriched in Vulto-van Silfhout ASD *de novo* CNVs (lilac box). When
1116 regions corresponding to the three most enriched gene sets from panel (A) plus 100kb
1117 flanks upstream and downstream of the genes are excluded from the analysis, UCE
1118 enrichment is lost in the cases of LoF and constrained genes. For embryonic brain genes,
1119 enrichment remains. When all three gene sets are combined, a normal distribution of
1120 expected overlaps was not produced, but the proportion of expected overlaps that was
1121 greater than or equal to the observed overlap is not consistent with enrichment († symbol).
1122 (C) UCEs are enriched in Pooled NDD breakpoint 100kb flanks (orange box). When
1123 regions corresponding to the three most enriched gene sets from panel A plus 100kb
1124 flanks upstream and downstream of the genes are excluded from analysis, UCE
1125 enrichment remains significant.
1126 Panels B and C: Large dots indicate observed overlap of UCEs, red indicates significant
1127 enrichment, black indicates lack of significant enrichment. Boxes indicate median and
1128 interquartile range for overlap with UCE-matched random control regions. Whiskers
1129 indicate 1.5x interquartile range, with all other points plotted as outliers.
1130 (D) Partial correlation analysis shows that positive correlation between the density of
1131 UCEs and LoF genes, and between the density of UCEs and Pooled NDD breakpoints
1132 with 100kb flanks is robust to co-correlation with enhancers, human accelerated regions
1133 (HARs), repetitive elements, these three combined, and GC content. For Vulto-van
1134 Silfhout ASD *de novo* CNVs, correlation with UCEs remained significant when

1135 controlling for co-correlation with enhancers, HARs, and GC percentage, but significance
1136 was lost when controlling for repetitive elements and for combined enhancers, HARs,
1137 and repetitive elements. Heatmap color represents shows spearman partial correlation
1138 coefficients; P values are noted within the cells; genome divided into 100kb bins; α =
1139 0.01 considering Bonferonni correction for 5 tests.
1140 All panels: green - LoF genes; lilac - Vulto-van Silfhout ASD *de novo* CNVs; orange -
1141 Pooled NDD breakpoints.
1142

Supplementary information legends

The supplementary information consists of four figures and three excel files.

Figures.

Figure S1: Explanation of methods to obtain datasets. A: Datasets from Vulto-van Silfhout *et al*. B: datasets from Sanders *et al*.

Figure S2: CNVs from healthy individuals are depleted of UCEs. CNV datasets from eight separate studies of healthy individuals were previously analyzed in McCole *et al*. and consistently showed significant depletion for UCEs.

Figure S3: Explanation of methods to obtain breakpoint datasets.
A: Processing for NDD breakpoint set 1. *Breakpoints derived from patients where their total genomic imbalance was ≤1kb. **Breakpoints also described in Chiang *et al*. 2012 were removed from the set taken from Talkowski *et al*. 2012. B: Processing for NDD breakpoint set 2.

Figure S4: Explanation of the 'CARD' mechanism.
Cell populations include those with no genomic rearrangements ('N', white), with benign rearrangements ('B', blue), or with deleterious genomic disruptions ('D', pink). Benign rearrangements are positioned where they do not disrupt homologous UCE pairing and comparison (orange lines between yellow UCEs), while deleterious disruptions are those that interfere with UCE pairing and are detected (orange exclamation point). Cells containing deleterious rearrangements are culled.

Excel files.

Table S1: Datasets. Tab A: Information on all datasets studied. Subsequent tabs: The coordinates for each dataset listed in (A) are contained in a tab, with the dataset name corresponding to the tab title. Last tab: detailed source information is given for Pooled NDD breakpoints.

Table S2: Enrichment analysis. A: CNVs in ASD subjects and siblings B: Balanced rearrangement breakpoints C: Gene sets. D: Vulto-van Silfhout ASD *de novo* CNVs, excluding gene sets from the analysis. E: Pooled NDD breakpoints with 100kb flanks, excluding gene sets from the analysis. All tabs: Proportion: of 1,000 expected overlap iterations, the number of times the expected overlap generated was equal to, or more extreme than, the observed UCE overlap (bp), divided by the total number of iterations, which was always 1,000. P-value: significance of whether the observed overlap (bp) differs from the expected overlaps, as determined by a Z-test. obs/exp: observed overlap (bp) divided by mean of expected overlaps (bp). Outcome: Determined with a two-tailed test ($P \leq 0.025$ in each tail for an overall a of 0.05).

Table S3: List of UCE coordinates together with their overlaps with genes and with all datasets analyzed in this study.

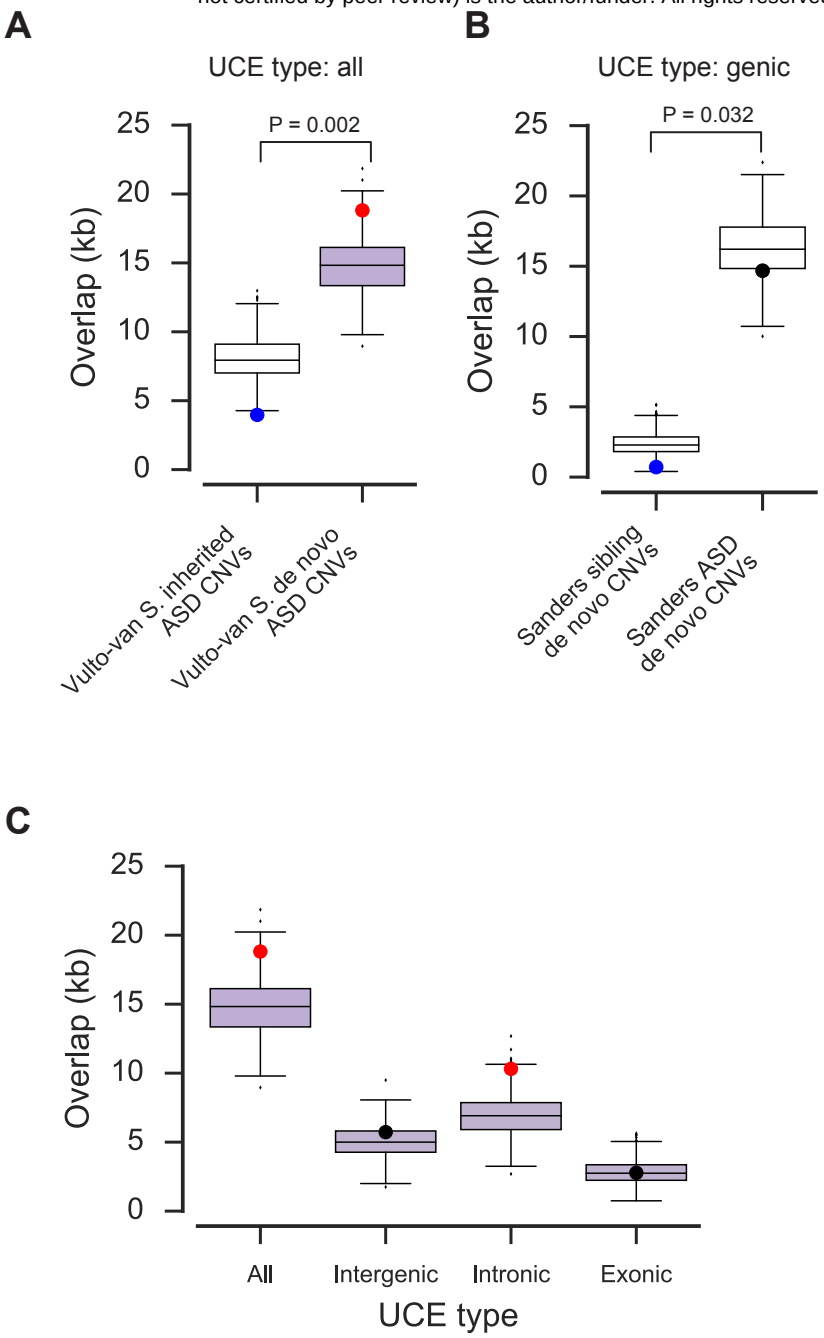

Figure 1

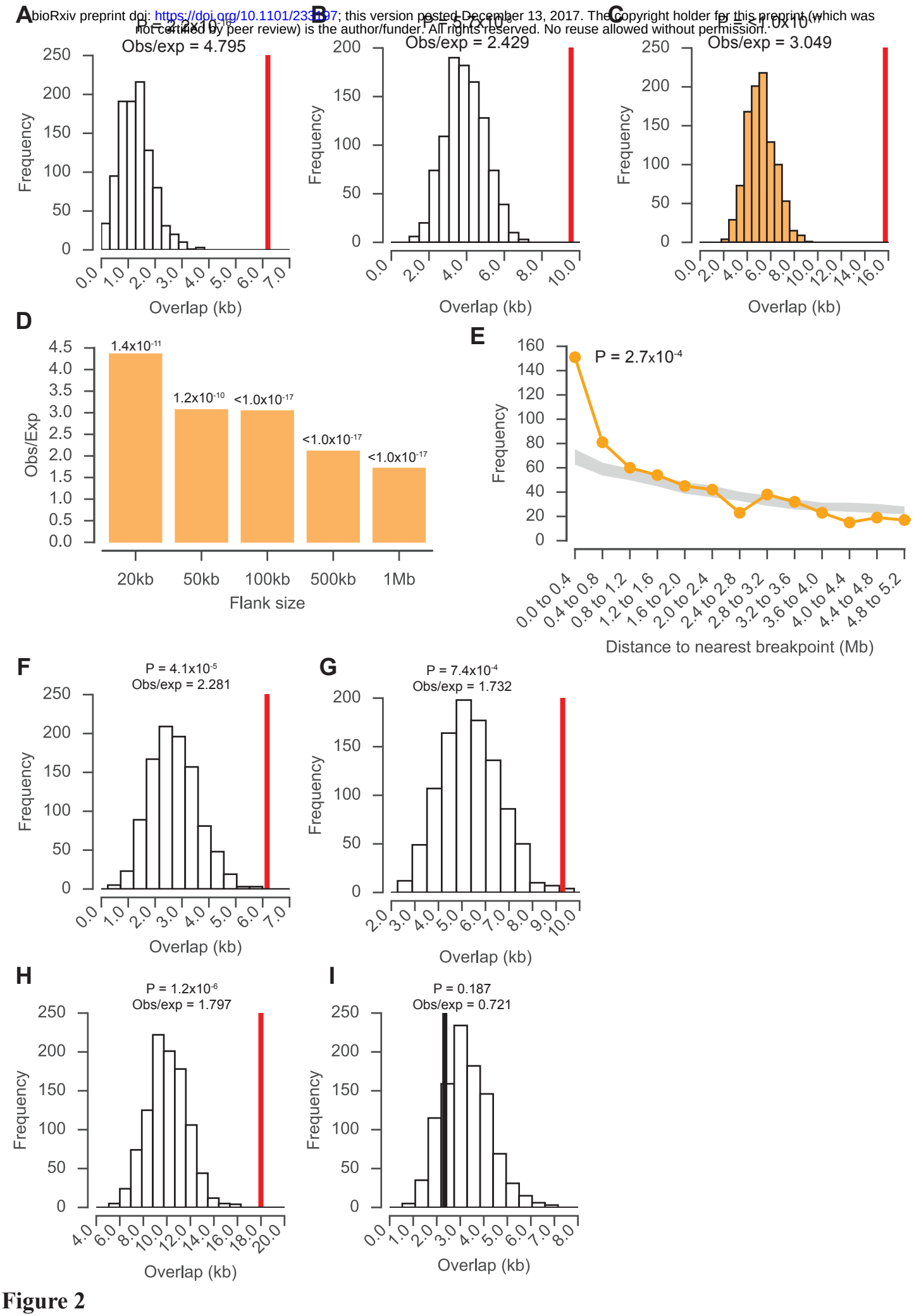

Figure 2

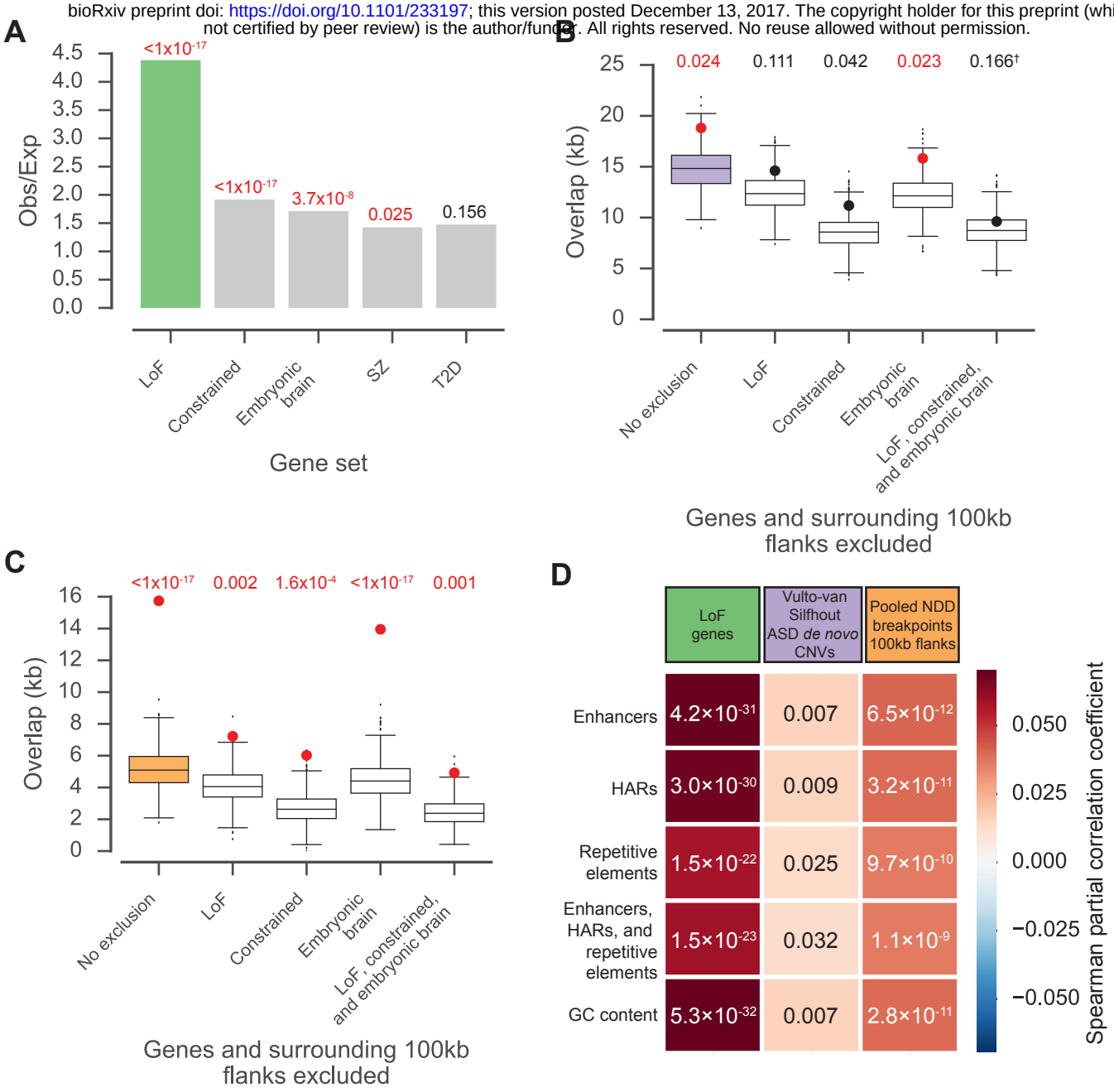

Figure 3

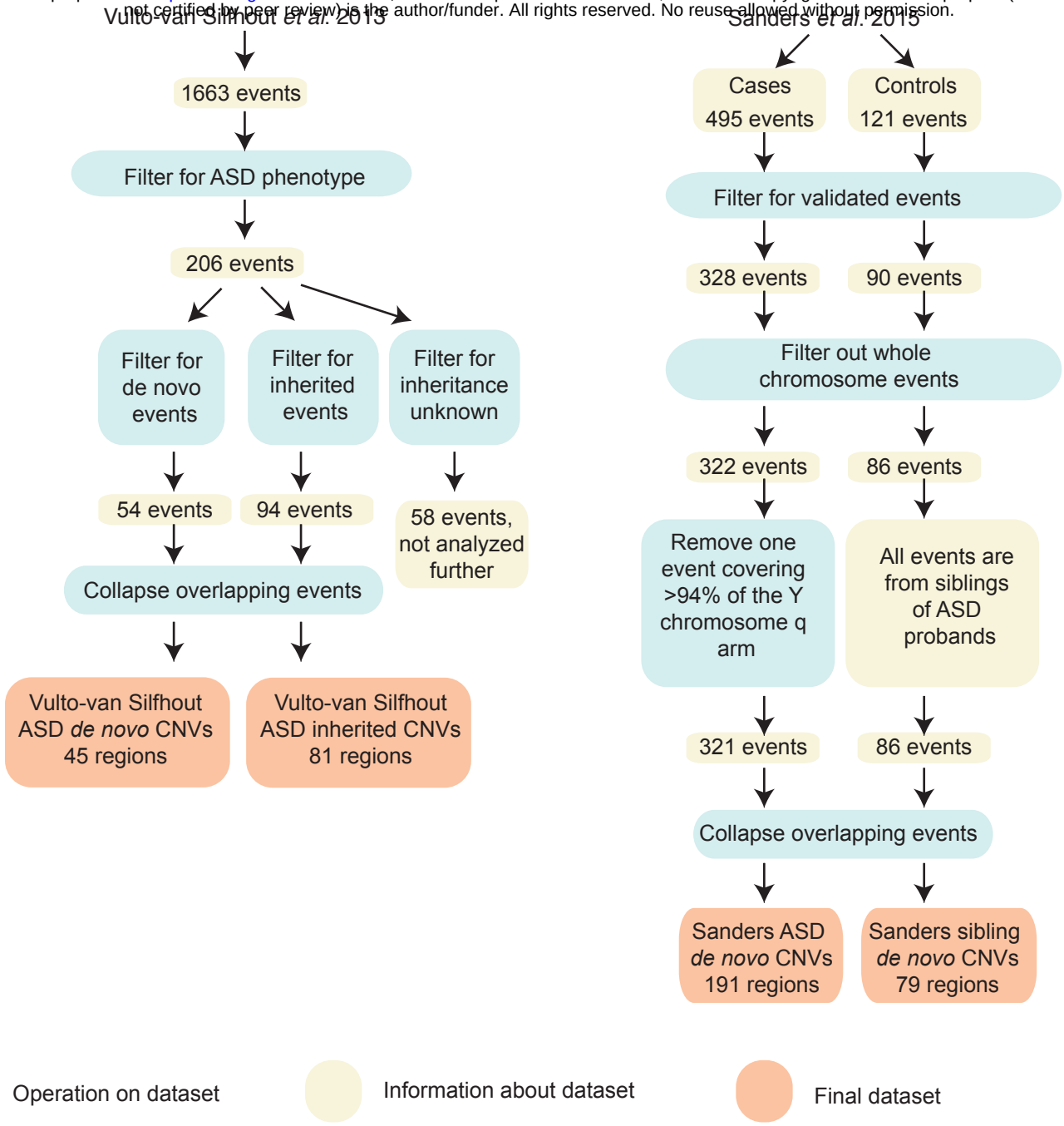

Figure S1: Explanation of methods to obtain CNV datasets.

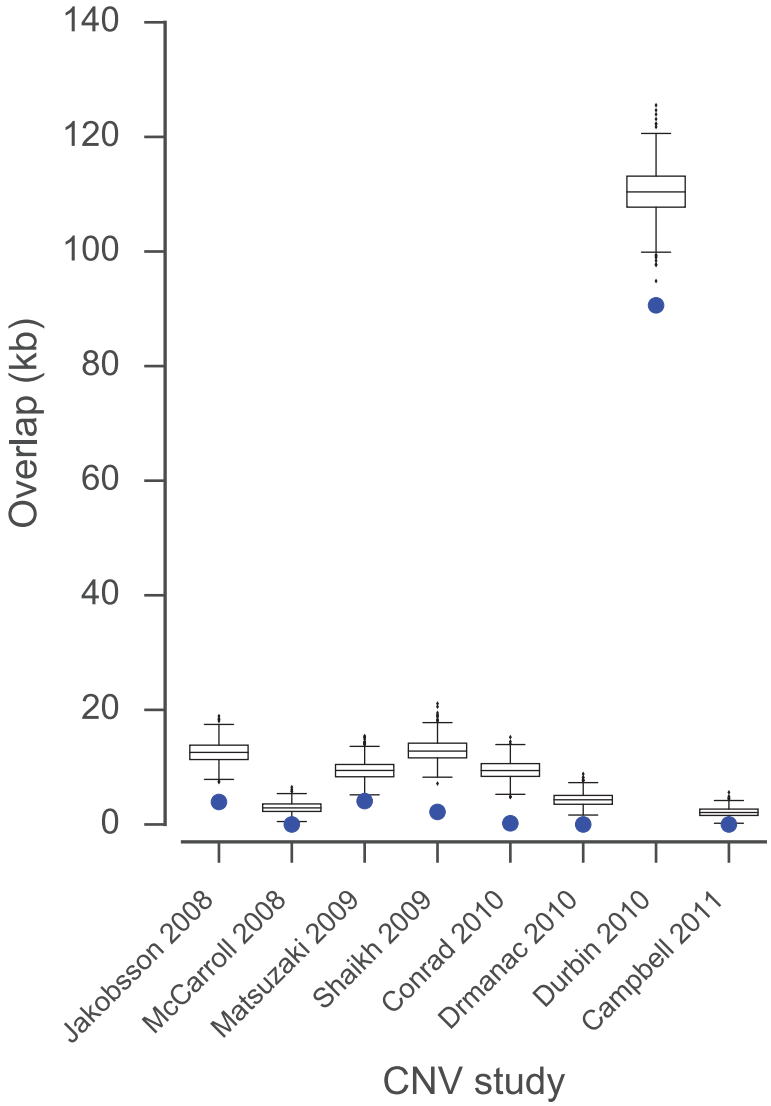

Figure S2. CNVs from healthy individuals are depleted of UCEs.

A: Processing for NDD breakpoints Set 1

Study	Number of breakpoints	Number of breakpoints with ≤1kb imbalance	Number of non-duplicate breakpoints	Number of breakpoints from subject with neurodevelop-metal phenotype
Granot-Hershkovitz *et al*. 2011	10	10	9	9
Kloosterman *et al*. 2011	24	24	23	23
Chiang *et al*. 2012	282	114*	108	100
Talkowski *et al*. 2012	38**	16	8	4
Nazaryan *et al*. 2014	24	24	21	21

Combine breakpoints
No duplicate breakpoints
NDD breakpoints Set 1 157 breakpoints
Add 100kb flanks and combine overlapping regions
NDD breakpoints Set 1 with 100kb flanks 76 regions

Combine breakpoints
Pooled NDD breakpoints 610 breakpoints‡
Cluster breakpoints occurring within 1kb
Retain one breakpoint from each cluster
Pooled NDD breakpoints within 1kb clustered 331 breakpoints

Add flanks and combine overlapping regions

20kb flanks → Pooled NDD breakpoints 20kb flanks 316 regions
50kb flanks → Pooled NDD breakpoints 50kb flanks 310 regions
100kb flanks → Pooled NDD breakpoints 100kb flanks 296 regions
500kb flanks → Pooled NDD breakpoints 500kb flanks 250 regions
1Mb flanks → Pooled NDD breakpoints 1Mb flanks 219 regions

Sort by pathogenicity category assigned to subject

Pathogenic 64 breakpoints
Likely pathogenic 150 breakpoints
Variants of Unknown Significance (VUS) 264 breakpoints

Add 500kb flanks and combine overlapping regions

Pathogenic breakpoints 500kb flanks 32 regions
Likely pathogenic breakpoints 500kb flanks 62 regions
VUS breakpoints 500kb flanks 120 regions

B: Processing for NDD breakpoints Set 2

Redin *et al*. 2017
1858 breakpoints
Filter for breakpoints in subjects with ≤1kb total imbalance
824 breakpoints
Filter for breakpoints pinpointed by capillary sequencing
744 breakpoints
Filter for breakpoints with IDs not matching any in Set 1
626 breakpoints
Remove any breakpoints that overlap those in Set 1
590 breakpoints
Filter for subject phenotype

Subjects with ear, eye, and/or nervous system phenotypes
478 breakpoints

Subjects with neither ear, nor eye, nor nervous system phenotypes
76 breakpoints

Remove duplicates

NDD breakpoint Set 2 453 breakpoints
Add flanks and combine overlapping regions
NDD breakpoints Set 2 with 100kb flanks 224 regions

Non-NDD breakpoints 76 breakpoints
Add flanks and combine overlapping regions
non-NDD breakpoints with 500kb flanks 38 regions

Operation on dataset
Information about dataset
Final dataset

Figure S3: Explanation of methods to obtain breakpoint datasets.

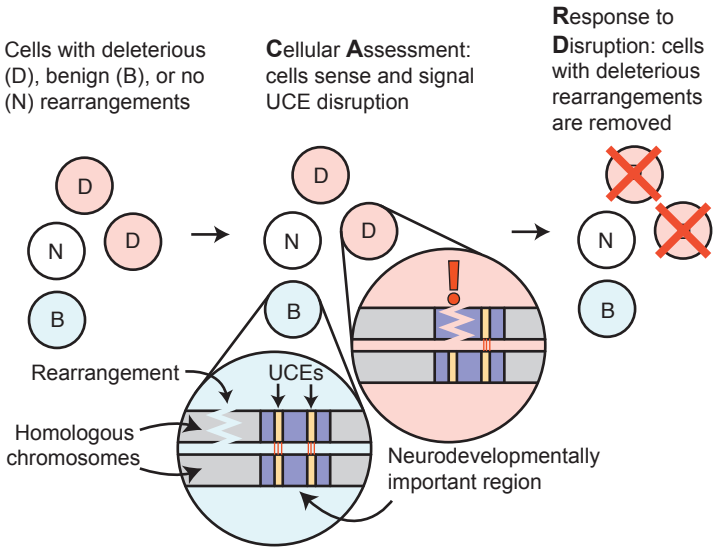

Figure S4. Explanation of the 'CARD' mechanism.