# Prioritization of genes driving congenital phenotypes of patients with *de novo* genomic structural variants

*Sjors Middelkamp[1,†], Judith M. Vlaar[1,†], Jacques Giltay[2], Jerome Korzelius[1,3], Nicolle Besselink[1], Sander Boymans[1], Roel Janssen[1], Lisanne de la Fonteijne[1], Ellen van Binsbergen[2], Markus J. van Roosmalen[1], Ron Hochstenbach[2], Daniela Giachino[5], Michael E. Talkowski[6,7,8], Wigard P. Kloosterman[2], Edwin Cuppen[1,*]*

[1]Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Utrecht 3584 CX, the Netherlands

[2]Department of Genetics, University Medical Center Utrecht, Utrecht 3584 EA, the Netherlands

[3]Max Planck Institute for Biology of Aging, Cologne, Germany

[5]Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Orbassano 10043, Italy.

[6]Molecular Neurogenetics Unit, Center for Human Genetic Research, Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA.

[7]Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

[8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[†] These authors contributed equally to this work

*Correspondence ecuppen@umcutrecht.nl

## Abstract

**Background:** Genomic structural variants (SVs) can affect many genes and regulatory elements. Therefore, the molecular mechanisms driving the phenotypes of patients with multiple congenital abnormalities and/or intellectual disability carrying *de novo* SVs are frequently unknown.

**Results:** We applied a combination of systematic experimental and bioinformatic methods to improve the molecular diagnosis of 39 patients with *de novo* SVs and an inconclusive diagnosis after regular genetic testing. In seven of these cases (18%) whole genome sequencing analysis detected disease-relevant complexities of the SVs missed in routine microarray-based analyses. We developed a computational tool to predict effects on genes directly affected by SVs and on genes indirectly affected due to changes in chromatin organization and impact on regulatory mechanisms. By combining these functional predictions with extensive phenotype information, candidate driver genes were identified in 16/39 (41%) patients. In eight cases evidence was found for involvement of multiple candidate drivers contributing to different parts of the phenotypes. Subsequently, we applied this computational method to a collection of 382 patients with previously detected and classified *de novo SVs* and identified candidate driver genes in 210 cases (54%), including 32 cases whose SVs were previously not classified as pathogenic. Pathogenic positional effects were predicted in 25% of the cases with balanced SVs and in 8% of the cases with copy number variants.

**Conclusions:** These results show that driver gene prioritization based on integrative analysis of WGS data with phenotype association and chromatin organization datasets can improve the molecular diagnosis of patients with *de novo* SVs.

**Keywords:** Structural variation, Copy number variants, Neurodevelopmental disorders, Intellectual disability, Multiple congenital anomalies, Driver genes, Whole genome sequencing, Transcriptome sequencing, Topologically associated domains, Positional effects

## Background

*De novo* germline structural variations (SVs) including deletions, duplications, inversions, insertions and translocations are important causes of (neuro-)developmental disorders such as intellectual disability and autism [1,2]. Clinical genetic centers routinely use microarrays and sometimes karyotyping to detect SVs at kilo- to mega base resolution [3]. The pathogenicity of an SV is generally determined by finding overlap with SVs in other patients with similar phenotypes [4,5]. SVs can affect large genomic regions which can contain many genes and non-coding regulatory elements [1]. This makes it challenging to determine which and how specific affected gene(s) and regulatory elements contributed to the phenotype of a patient. Therefore, the causative genes driving the phenotype are frequently unknown for patients with *de novo* SVs which can hamper conclusive genetic diagnosis.

SVs can have a direct effect on the expression and functioning of genes by altering their copy number or by truncating their coding sequences [1]. In addition, SVs can also indirectly influence the expression of adjacent genes by disrupting the interactions between genes and their regulatory elements [6]. New developments in chromatin conformation capture (3C) based technologies such as Hi-C have provided the means to study these indirect, positional effects [7]. Most of the genomic interactions (loops) between genes and enhancers occur within megabase-sized topologically associated domains (TADs). These domains are separated from each other by boundary elements characterized by CTCF-binding, which limit interactions between genes and enhancers that are not located within the same TAD [8,9]. For several loci, such as the *EPHA4* [10], *SOX9* [11], *IHH* [12], *Pitx* [13] loci, it has been demonstrated that disruption of TAD boundaries by SVs can cause rewiring of genomic interactions between genes and enhancers, which can lead to altered gene expression during embryonic development and ultimately in disease phenotypes [14]. Although the organization of TADs appears to be stable across cell types, sub-TAD genomic interactions between genes and regulatory elements have been shown to be relatively dynamic and cell type-specific [15]. Disruptions of genomic interactions are therefore optimally studied in disease-relevant cell types, which may be obtained from mouse models or from patient-derived induced pluripotent stem cells. However, it is not feasible to study each individual locus or patient with such

81 elaborate approaches and disease-relevant tissues derived from patients are usually not
82 available. Therefore, it is not yet precisely known how frequently positional effects contribute
83 to the phenotypes of patients with developmental disorders.

84 It has been shown that the use of computational methods based on combining
85 phenotypic information from the Human Phenotype Ontology (HPO) database
86 ("phenomatching") with previously published chromatin interactions datasets can improve the
87 interpretation of the molecular consequences of *de novo* SVs [16–18]. These approaches
88 have largely been based on data derived from a small set of cell types and techniques. Here,
89 we further expand these *in silico* approaches by integrating detailed phenotype information
90 with genome-wide chromatin conformation datasets of many different cell types. By combining
91 this method with whole genome and transcriptome sequencing we predicted which genes are
92 affected by the SVs and which of these genes have likely been involved in the development
93 of the disease phenotype (e.g. candidate driver genes). Accurate characterization of the
94 effects of SVs on genes can be beneficial for the prediction of potential clinical relevance of
95 the SVs. Detailed interpretation of the molecular effects of the SVs helped to identify candidate
96 driver genes in 16 out of 39 included patients who had an inconclusive diagnosis after regular
97 genetic testing. By applying the computational method on larger cohorts of patients with *de*
98 *novo* SVs we estimated the contribution of positional effects for both balanced and unbalanced
99 SVs.
100
101 **Results**
102 **WGS reveals hidden complexity of *de novo* SVs**
103 We aimed to improve the genetic diagnosis of 39 individuals with multiple congenital
104 abnormalities and/or intellectual disability (MCA/ID) who had an inconclusive diagnosis after
105 regular genetic testing or who have complex genomic rearrangements. The phenotypes of the
106 individuals were systematically described by Human Phenotype Ontology (HPO) terms [19–
107 21]. The included individuals displayed a wide range of phenotypic features and most
108 individuals (82%) presented neurological abnormalities including intellectual disability (Fig. 1a,

Additional File 2: Table S1). The parents of each of the patients were healthy, suggesting a *de novo* or recessive origin of the disease phenotypes. All individuals carried *de novo* SVs which were previously detected by ArrayCGH, SNP arrays, karyotyping or long-insert mate-pair sequencing (Additional File 1: Fig. S1a). First, we performed whole genome sequencing (WGS) for all individuals in the cohort to screen for potential pathogenic genetic variants that were not detected by the previously performed genetic tests. No known pathogenic single nucleotide variants (SNVs) were detected in the individuals analysed by patient-parents trio-based WGS (individuals P1 to P20), except for one pathogenic SNV that is associated with a part (haemophilia) of the phenotype of individual P1. A total of 46 unbalanced and 219 balanced *de novo* SVs were identified in the genomes of the individuals (Fig.1b, Additional File 1: Fig. S1b, Additional File 2: Table S2). The detected SVs ranged from simple SVs to very complex genomic rearrangements that ranged from 4 to 40 breakpoint junctions per individual. Importantly, WGS confirmed all previously detected *de novo* SVs and revealed additional complexity of the SVs in 7 (39%) of the 18 cases who were not studied by WGS-based techniques before (Fig. 1c-d, Additional File 2: Table S2). In half of the cases with previously identified *de novo* copy number gains (4/8), the gains were not arranged in a tandem orientation, but instead they were inserted in another genomic region, which can have far-reaching consequences for accurate interpretation of the pathogenetic mechanisms in these individuals (Fig. 1d) [22–24]. This suggests that the complexity of copy number gains in particular is frequently underestimated by microarray analysis. For example, in one case (P11) a previously detected 170 kb copy number gain from chromosome 9 was actually inserted into chromosome X, 82 kb upstream of the *SOX3* gene (Fig. 1d, Additional File 1: Fig. S2). This inserted fragment contains a super-enhancer region that is active in craniofacial development [25] (Additional File 1: Fig. S2). The insertion of the super-enhancer may have disturbed the regulation of *SOX3* expression during palate development, which may have possibly caused the orofacial clefting in this individual [26–30]. The detection of these additional complexities in these seven patients exemplifies the added value that WGS analyses can have for cases that remain unresolved after standard array diagnostics [24].

137

138 ***In silico* phenomatching approach links directly affected genes to phenotypes**

139 Subsequently, we determined if the phenotypes of the patients could be explained by direct
140 effects of the *de novo* SVs, most of which were previously classified as a variant of unknown
141 significance (VUS), on genes. In total, 332 genes are directly affected (deleted, duplicated or
142 truncated) by the SVs in the cohort (Additional File 1: Fig. S1d). The Phenomatch tool was
143 used to match the HPO terms associated with these genes with the HPO terms used to
144 describe the phenotypes of the individuals [16,17]. Genes were considered as candidate driver
145 genes based on the height of their Phenomatch score, the number of phenomatches between
146 the HPO terms of the gene and the patient, recessive or dominant mode of inheritance, Loss
147 of Function constraint score (pLI) [31], Residual Variation Intolerance Score (RVIS) [32] and
148 the presence in OMIM and/or DD2GP [33] databases (Table 1). Directly affected genes
149 strongly or moderately associated with the phenotype are classified as tier 1 (T1) and tier 2
150 (T2) candidate driver genes, respectively (Fig. 2a, Table 1). Genes with limited evidence for
151 contribution to the phenotype are reported as tier 3 (T3) genes. In the cohort of 39 patients,
152 this approach prioritized 2 and 13 of the 332 directly affected genes as T1 and T2 candidate
153 drivers, respectively (Fig. 2b). In three cases, the HPO terms of the identified T1/T2 candidate
154 driver genes could be matched to more than 75% of the HPO terms assigned to the patients,
155 indicating that the effects of the SVs on these genes can explain most of the phenotypes of
156 these patients. (Additional file 2: Table S3). In six other cases, directly affected T1/T2
157 candidate drivers were identified that were only associated with a part of the patient's
158 phenotypes (Additional file 2: Table S3).

159 Subsequently, we performed RNA sequencing on primary blood cells or
160 lymphoblastoid cell lines derived from the individuals to determine the impact of *de novo* SVs
161 on RNA expression of candidate driver genes. RNA sequencing confirmed that most
162 expressed genes directly affected by *de novo* deletions show a reduced RNA expression (97
163 of 107 genes with a median reduction of 0.46-fold compared to non-affected individuals) (Fig.
164 2d). Although duplicated genes show a median 1.44-fold increase in expression, only 14 of 43

(~30%) of them are significantly overexpressed compared to expression levels in non-affected individuals. In total, 87 genes are truncated by SVs and four of these are classified as T1/T2 candidate drivers. The genomic rearrangements lead to 12 possible fusions of truncated genes and RNA-seq showed an increased expression for two gene fragments due to formation of a fusion gene (Additional file 1: Fig. S3, Additional file 2: Table S4). However, none of the genes involved in the formation of fusion genes were associated with the phenotypes of the patients, although we cannot exclude an unknown pathogenic effect of the newly identified fusion genes. We could detect expression for three deleted and two duplicated T1/T2 candidate drivers and these were differentially expressed when compared to controls. The RNA sequencing data suggests that most genes affected by *de novo* deletions show reduced RNA expression levels and limited dosage compensation. However, the observed increased gene dosage by *de novo* duplications does not always lead to increased RNA expression, at least in blood cells of patients.

**Prediction of positional effects of *de novo* SVs on neighbouring genes**

In 28 of the included cases (72%) our prioritization method did not predict T1/T2 candidate driver genes that are directly affected by the *de novo* SVs. Therefore, we investigated positional effects on the genes surrounding the *de novo* SVs to explain the phenotypes in those cases that were not fully explained by directly affected candidate driver genes. We extended our candidate driver gene prioritization analysis by including all the protein-coding genes located within 2 Mb of the breakpoint junctions, as most chromatin interactions are formed between loci that are less than 2Mb apart from each other [34]. Of the 2,754 genes adjacent to the SVs, 117 are moderately to strongly associated with the specific phenotypes of the individuals based on phenotype association analysis. However, this association with the phenotype does not necessarily mean that these genes located within 2Mb of the breakpoint junctions are really affected by the SVs and thus contributing to the phenotype. To determine if the regulation of these genes was affected, we first evaluated RNA expression levels of those genes. Three-quarters (81/117) of the genes linked to the phenotypes were expressed,

193 but only 9 of these showed reduced or increased expression (Fig. 2d). However, RNA
194 expression in blood may not always be a relevant proxy for most neurodevelopmental
195 phenotypes [35,36]. Therefore, we developed an extensive *in silico* strategy to predict potential
196 disruption of the regulatory landscape of the genes surrounding the SVs (Additional file 1: Fig.
197 S4). Because the interactions between genes and their regulatory elements are cell-type
198 specific a large collection of tissue-specific Hi-C, TAD, promoter capture Hi-C (PCHiC), DNase
199 hypersensitivity site (DHS), RNA and ChIP-seq datasets was included (Additional file 2: Table
200 S5). Several embryonic and neural cell type (such as fetal brain and neural progenitor cells)
201 datasets were included that may be especially relevant to study the neurodevelopmental
202 phenotypes in our cohort.

203 For each gene, we selected the TAD it is located in [37–39], the PCHiC interactions
204 with its transcription start site [40–43] and its DHS connections [44] for each assessed cell
205 type and overlapped these features with the breakpoint junctions of the SVs to determine the
206 proportion of disrupted genomic interactions (Methods, Additional file 1: Fig. S4). We also
207 counted the number of enhancers (which are active in cell types in which the genes show the
208 highest RNA expression [45]) that are located on disrupted portions of the TADs. Additionally,
209 we performed virtual 4C (v4C) for each gene by selecting the rows of the normalized Hi-C
210 matrices containing the transcription start site coordinates of the genes as viewpoints,
211 because the coordinates of TAD boundaries can be dependent on the calling method and the
212 resolution of the Hi-C [46–48] and because a significant portion of genomic interactions
213 crosses TAD boundaries [9]. Integrated scores for TAD disruption, v4C disruption, potential
214 enhancer loss, disruption of PCHiC interactions and DHS connections were used to calculate
215 a positional effect support score for each gene (Additional file 1: Fig. S4). Finally, indirectly
216 affected genes were classified as tier 1, 2 or 3 candidate drivers based on a combination of
217 their association with the phenotype and their support score (Fig. 2a, Table 1).

218 Of the 117 genes that were associated with the phenotypes and located within 2 Mb
219 of the SVs, 16 genes were predicted to be affected by the SVs based on the *in silico* analysis
220 and therefore classified as T1/T2 candidate driver gene (Fig. 2b). The validity of the approach

was supported by the detection of pathogenic positional effects identified in previous studies. For example, the regulatory landscape of *SOX9* was predicted to be disturbed by a translocation 721 Kb upstream of the gene in individual P5, whose phenotype is mainly characterized by acampomelic campomelic dysplasia with Pierre-Robin Syndrome (PRS) including a cleft palate (Additional file 1: Fig. S6). SVs in this region have been predicted to disrupt interactions of *SOX9* with several of its enhancers further upstream, leading to phenotypes similar to the phenotype of individual P5 [49,50]. In individual P39, who has been previously included in other studies, our method predicted a disruption of *FOXG1* expression regulation due to a translocation (Additional file 1: Fig. S4), further supporting the hypothesis that deregulation of *FOXG1* caused the phenotype of this individual [51,52].

Another example of a predicted positional effect is the disruption of the regulatory landscape of the *HOXD* locus in individual P22. This individual has complex genomic rearrangements consisting of 40 breakpoint junctions on four different chromosomes likely caused by chromothripsis [53]. One of the inversions and one of the translocations are located in the TAD upstream (centromeric) of the *HOXD* gene cluster (Fig. 2c). This TAD contains multiple enhancers that regulate the precise expression patterns of the *HOXD* genes during the development of the digits [54–56]. Deletions of the gene cluster itself, but also deletions upstream of the cluster are associated with hand malformations [57–59]. The translocation in individual P22 disrupts one of the main enhancer regions (the global control region (GCR)), which may have led to altered regulation of the expression of *HOXD* genes, ultimately causing brachydactyly and clinodactyly in this patient.

Our approach predicted positional effects on T1/T2 candidate driver genes in 10 included cases (26%). Most of these predicted positional effects were caused by breakpoint junctions of balanced SVs, suggesting that these effects may be especially important for balanced SVs.

**Prediction of driver genes improves molecular diagnosis**

By combining both directly and indirectly affected candidate drivers per patient we found possible explanations for the phenotypes of 16/39 (41%) complex and/or previously unsolved cases (Fig. 3a, Additional file 2: Table S3). Interestingly, in eight cases we found evidence for multiple candidate drivers that are individually only associated with part of the phenotype, but together may largely explain the phenotype (Fig. 3b). For example, we identified four candidate drivers in individual P25, who has a complex phenotype characterized by developmental delay, autism, seizures, renal agenesis, cryptorchidism and an abnormal facial shape (Fig. 3c). This individual has complex genomic rearrangements consisting of six breakpoint junctions and two deletions of ~10 Mb and ~0.6 Mb on three different chromosomes (Fig. 3d). The 6q13q14.1 deletion of ~10 Mb affects 33 genes including the candidate drivers *PHIP* and *COL12A1,* which have been associated with developmental delay, anxiety and facial dysmorphisms in other patients [60,61]. In addition to the two deleted drivers, two genes associated with other parts of the phenotype were predicted to be affected by positional effects (Fig. 3e). One of these genes is *TFAP2A*, whose TAD (characterized by a large gene desert) and long-range interactions overlap with a translocation breakpoint junction. Rearrangements affecting the genomic interactions between *TFAP2A* and enhancers active in neural crest cells located in the *TFAP2A* TAD have recently been implicated in branchio-oculofacial syndrome [62]. The regulation of *BMP2*, a gene linked to agenesis of the ribs and cardiac features, is also predicted to be disturbed by a complex SV upstream of this gene [63,64]. Altogether, these candidate driver genes may have jointly contributed to the phenotype of this individual (Fig. 3d). This case illustrates the challenge of identifying the causal genes driving the phenotypes of patients with structural rearrangements and highlights the notion that multiple genes should be considered for understanding the underlying molecular processes and explaining the patient's phenotype.

### ***In silico*** **driver gene prediction in larger patient cohorts**

Our candidate driver prioritization approach identified many candidate drivers in previously unresolved cases, but these complex cases may not be fully representative for the general

patient population seen in clinical genetic diagnostics. Therefore, we applied our prediction method to two larger sets of patients with *de novo* SVs to further assess the validity and value of the approach. We focused on the genes located at or within 1 Mb of the SVs, because most of the candidate driver genes we identified in our own patient cohort were located within 1 Mb of an SV breakpoint junction (Additional file 1: Fig. S5b). First, we determined the effects of largely balanced structural variants in 228 previously described patients (Additional file 1: Fig. S7a) [51]. In 101 of the 228 (44%) cases the detected *de novo* SVs were previously classified as pathogenic or likely pathogenic and in all but four of these diagnosed cases one or more candidate driver genes have been proposed (Additional file 1: Fig. S7b). Our approach identified 46 T1 and 92 T2 candidate drivers out of 7406 genes located within 1 Mb of the SVs (Additional file 1: Fig. S7c,d, Additional file 2: Table S7). More than half (85/138) of the identified T1/T2 candidate drivers were not previously described as driver genes. In contrast, 23/114 (22%) previously described pathogenic or likely pathogenic drivers were classified as T3 candidates and 38/114 (33%) were not reported as driver by our approach (Fig. 4a), mostly because the Phenomatch scores were below the threshold (46%) or because the genes were not associated with HPO terms (41%) (Additional file 1: Fig. S7e). T1/T2 candidate drivers were identified in 99/225 (44%) of the individuals with mostly balanced SVs, including 31 individuals with SVs that were previously classified as VUS (Fig. 4b, Additional file 1: Fig. S8). Positional effect on genes moderately to strongly associated with the phenotypes were predicted in 63 (28%) of the cases with balanced SVs.

Subsequently, we also assessed the value of our driver prioritization approach for individuals with unbalanced copy number variants. We collected genetic and phenotypic information of 154 individuals with *de novo* copy number variants (<10 Mb) identified by clinical array-based copy number profiling (Additional file 1: Fig. S7a,b, Additional file 2: Table S6). The CNVs in the majority (83%) of these individuals have been previously classified as pathogenic according to clinical genetic diagnostic criteria (Additional file 1: Fig. S7b). These criteria are mostly based on the overlap of the CNVs with CNVs of other individuals with similar phenotypes and the causative driver genes were typically not previously specified. Our method

identified T1/T2 candidate driver genes in 87/154 (56%) individuals, including 9/26 individuals with CNVs previously classified as VUS (Fig. 4a, Additional file 2: Table S7). Interestingly, support for positional effects on candidate drivers was only found in 11% of the cases with CNVs, suggesting that pathogenic positional effects are more common in patients with balanced SVs than in patients with unbalanced SVs (Fig. 4b). No driver genes were identified for 39% of the previously considered pathogenic CNVs (based on recurrence in other patients). In some cases, potential drivers may remain unidentified because of incompleteness of the HPO database or insufficient description of the patient's phenotypes. However, given the WGS results described for our patient cohort, it is also likely that some complexities of the CNVs may have been missed by the array-based detection method. The data also suggests that many disease-causing genes or mechanisms are still not known, and that some SVs are incorrectly classified as pathogenic.

## Discussion

About half of the patients with neurodevelopmental disorders do not receive a diagnosis after regular genetic testing based on whole exome sequencing and microarray-based copy number profiling [3]. Furthermore, the molecular mechanisms underlying the disease phenotype often remain unknown, even when a genetic variant is diagnosed as (potentially) pathogenic in an individual, as this is often only based on recurrence in patients with a similar phenotype. Here, we applied an integrative method based on WGS, computational phenomatching and prediction of positional effects to improve the diagnosis and molecular understanding of disease aetiology of individuals with *de novo* SVs.

Our WGS-approach identified additional complexities of the *de novo* SVs previously missed by array-based analysis in 7 of 18 cases, supporting previous findings that WGS can have an added value in identifying additional SVs that are not routinely detected by microarrays [24]. Our results indicate that duplications in particular are often more complex than interpreted by microarrays, which is in line with previous studies [22,65]. WGS can therefore be a valuable follow-up method to improve the diagnosis particularly of patients with

copy number gains classified as VUS. Knowing the exact genomic location and orientation of SVs is important for the identification of possible positional effects.

To systematically dissect and understand the impact of *de novo* SVs, we developed a computational tool based on integration of HiC, RNA-seq and ChIP-seq datasets to predict positional effects of SVs on the regulation of gene expression. We combined these predictions with phenotype association information to identify candidate driver genes. In 9/39 of the complex cases we identified candidate drivers that are directly affected by the break-junctions of the SVs. Positional effects of SVs have been shown to cause congenital disorders, but their significance is still unclear and they are not yet routinely screened for in genetic diagnostics [14]. Our method predicted positional effects on genes associated with the phenotype in 25% and 8% of all studied cases with balanced and unbalanced *de novo* SVs, respectively. Previous studies estimated that disruptions of TAD boundaries may be the underlying cause of the phenotypes of ~7.3% patients with balanced rearrangements [51] and of ~11.8% of patients with large rare deletions [16]. Our method identified a higher contribution of positional effects in patients with balanced rearrangements mainly because our method included more extensive chromatin conformation datasets and also screened for effects that may explain smaller portions of the phenotypes. Our method, although it incorporates most of all published chromatin conformation datasets on untransformed human cells, focuses on disruptions of interactions, which is a simplification of the complex nature of positional effects. It gives an insight in the potential effects that lead to the phenotypes and prioritizes candidates that need to be followed up experimentally, ideally in a developmental context for proofing causality.

SVs can affect many genes and multiple ‘disturbed’ genes may together contribute to the phenotype. Indeed, in eight cases we found support for the involvement of multiple candidate drivers that were affected by one or more *de novo* SVs. Nevertheless, in many of the studied cases our method did not detect candidate drivers. This may be due to insufficient data or knowledge about the genes and regulatory elements in the affected locus and/or due to missing disease associations in the used databases. Additionally, *de novo* SVs are also frequently identified in healthy individuals [66–68] and some of the detected SVs of unknown

significance may actually be benign and the disease caused by other genetic or non-genetic factors. The datasets underlying our computational workflow can be easily updated with more detailed data when emerging in the future, thereby enabling routine reanalysis of previously identified SVs. Moreover, our approach can be extended to study the consequences of SVs in different disease contexts such as cancer, where SVs also play a major causal role.

## Conclusions

Interpretation of SVs is important for clinical diagnosis of patients with developmental disorders, but it remains a challenge because SVs can have many different effects on multiple genes. We developed an approach to gain a detailed overview of the genes and regulatory elements affected by *de novo* SVs in patients with congenital disease. We show that WGS can be useful as a second-tier test to detect variants that are not detected by exome- and array-based approaches.

## Methods

### Patient selection and phenotyping

A total of 39 individuals with *de novo* germline SVs and an inconclusive diagnosis were included in this study. Individuals P1 to P21 and their biological parents were included at the University Medical Center Utrecht (the Netherlands) under study ID NL55260.041.15 15-736/M. Individual P22, previously described by Redin *et al.* as UTR22 [51], and her parents were included at the San Luigi University Hospital (Italy). For individuals P23 to P39, lymphoblastoid cell lines (LCL cell lines) were previously derived as part of the Developmental Genome Anatomy Project (DGAP) of the Brigham and Women's Hospital and Massachusetts General Hospital, Boston, Massachusetts, USA [51]. Written informed consent was obtained for all included individuals and parents and the studies were approved by the respective institutional review boards.

### DNA and RNA extraction

Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood samples of individuals P1 to P22 and their biological parents using a Ficoll-Paque Plus gradient (GE Healthcare Life Sciences) in SepMate tubes (STEMCELL Technologies) according to the manufacturer's protocols. LCL cell lines derived from individuals P23 to P39 were expanded in RPMI 1640 medium supplemented with GlutaMAX (ThermoFisher Scientific), 10% fetal bovine serum, 1% penicillin, 1% streptomycin at 37°C. LCL cultures of each individual were split into three flasks and cultured separately for at least one week to obtain technical replicate samples for RNA isolation. Genomic DNA was isolated from the PBMCs or LCL cell lines using the QIASymphony DNA kit (Qiagen). Total RNA was isolated using the QIAsymphony RNA Kit (Qiagen) and RNA quality (RIN > 8) was determined using the Agilent RNA 6000 Nano Kit.

**Whole-genome sequencing**

Purified DNA was sheared to fragments of 400-500 bp using a Covaris sonicator. WGS libraries were prepared using the TruSeq DNA Nano Library Prep Kit (Illumina). WGS libraries were sequenced on an Illumina Hiseq X instrument generating 2x150 bp paired-end reads to a mean coverage depth of at least 30x. The WGS data was processed using an in-house Illumina analysis pipeline (https://github.com/UMCUGenetics/IAP). Briefly, reads were mapped to the CRCh37/hg19 human reference genome using BWA-0.7.5a using "BWA-MEM -t 12 -c 100 -M -R" [69]. GATK IndelRealigner [70] was used to realign the reads. Duplicated reads were removed using Sambamba markdup [71].

**Structural variant calling and filtering**

Raw SV candidates were called with Manta v0.29.5 using standard settings [72] and Delly v0.7.2 [73] using the following settings: "-q 1 -s 9 -m 13 -u 5". Only Manta calls overlapping with breakpoint junctions called by Delly (+/- 100 basepairs) were selected. Rare SVs were selected by filtering against SV calls of 1000 Genomes [74] and against an in-house database containing raw Manta SV calls of ~120 samples (https://github.com/UMCUGenetics/vcf-explorer). *De novo* SVs were identified in individuals P1 to P22 by filtering the SVs of the

children against the Manta calls (+/- 100 basepairs) of the father and the mother. Filtered SV calls were manually inspected in the Integrative Genome Viewer (IGV). *De novo* breakpoint junctions of individuals P1 to P21 were validated by PCR using AmpliTaq gold (Thermo Scientific) under standard cycling conditions and by Sanger sequencing. Primers were designed using Primer3 software (Additional file S2: Table S2). Breakpoint junction coordinates for individuals P22 to P39 were previously validated by PCR [51,53].

**Single nucleotide variant filtering**

Single nucleotide variants and indels were called using GATK HaplotypeCaller. For individuals P1 to P21 (whose parents were also sequenced), reads overlapping exons were selected and the Bench NGS Lab platform (Agilent-Cartagenia) was used to detect possible pathogenic *de novo* or recessive variants in the exome. *De novo* variants were only analysed if they affect the protein structure of genes that are intolerant to missense and loss-of-function variants. Only putative protein changing homozygous and compound heterozygous variants with an allele frequency of <0.5% in ExAC [31] were reported.

**RNA-sequencing and analysis**

RNA-seq libraries were prepared using TruSeq Stranded Total RNA Library Prep Kit (Illumina) according to the manufacturer's protocol. RNA-seq libraries were pooled and sequenced on a NextSeq500 (Illumina) in 2x75bp paired-end mode. Processing of RNA sequencing data was performed using a custom in-house pipeline (https://github.com/UMCUGenetics/RNASeq). Briefly, reads were aligned to the CRCh37/hg19 human reference genome using STAR 2.4.2a [75]. The number of reads mapping to genes were counted using HTSeq-count 0.6.1 [76]. Genes overlapping with SV breakpoints (eg truncated genes) were also analyzed separately by counting the number of reads mapping to exons per truncated gene fragment (up- and downstream of the breakpoint junction). RNA-seq data obtained from PBMCs (Individuals P1 to P22) and LCL cell lines (Individuals P23 to P39) were processed as separate datasets. The R-package DESeq2 was used to normalize raw read counts and to perform differential gene

444 expression analysis for both datasets separately [77]. Genes with more than 0.5 reads per
445 kilobase per million mapped reads (RPKM) were considered to be expressed.
446
447 **Gene annotation**
448 Gene information (including genomic positions, Ensembl IDs, HGNC symbols and Refseq IDs)
449 was obtained from Ensembl (GRCh37) using the R-package biomaRt (v2.38) [78]. Genes
450 containing a RefSeq mRNA ID and a HGNC symbol were considered as protein-coding genes.
451 Genomic coordinates for the longest transcript were used if genes contained multiple RefSeq
452 mRNA IDs. The list of 19,300 protein-coding genes was further annotated with 1) pLI, 2) RVIS,
453 3) haploinsufficiency (HI) scores, 4) OMIM identifiers and 5) DD2GP information for each gene
454 (see Additional file 2: Table S5 for data sources). A phenotypic score based on these five
455 categories was determined for each gene. Modes of inheritance for each gene were retrieved
456 from the HPO and DD2GP databases.
457
458 **Computational prediction of the effects of SVs on genes**
459 For each patient, the protein-coding genes located at or adjacent (< 2Mb) to the *de novo* SV
460 breakpoint junctions were selected from the list of genes generated as described above. The
461 HPO terms linked to these genes in the HPO database [19–21] were matched to each
462 individual HPO term assigned to the patient and to the combination of the patient's HPO terms
463 using Phenomatch [16,17]. For each gene, the number of phenomatchScores higher than 5
464 ("phenomatches") with each individual HPO term of a patient was calculated. The strength of
465 the association (none, weak, medium or strong) of each selected gene with the phenotype of
466 the patient was determined based on the total phenomatchScore, the number of
467 phenomatches, the mode of inheritance and the phenotypic score (Fig. S4a, Table 1).
468 Subsequently, potential direct and indirect effects of the SVs (none, weak or strong)
469 on the genes were predicted (Fig. S4a, Table 1). The prediction analyses were based on
470 chromatin organization and epigenetic datasets of many different cell types obtained from
471 previous studies (see Additional file 2: Table S5 for data sources).

First, we determined which TADS of 20 different cell types overlapped with the *de novo* SVs and which genes were located within these disrupted TADs [37–39] (Additional file 1: Fig. S4b). To determine if the disrupted portions of the TADs contained regulatory elements that may be relevant for the genes located in the affected TADs, we selected the three cell types in which the gene is highest expressed based on RNA-seq data from the Encode/Roadmap projects [45] reanalysed by Schmitt et al [37] (Additional file 1: Fig S4C). The number of active enhancers (determined by chromHMM analysis of Encode/Roadmap ChIP-seq data [45]) in the TADs up- and downstream of the breakpoint junction in the three selected cell types was counted (Additional file 1: Fig S4D). Virtual 4C was performed by selecting the rows of the normalized Hi-C matrices containing the transcription start site coordinates of the genes. The v4C profiles were overlapped with the breakpoint junctions to determine the portion of interrupted Hi-C interactions of the gene (Additional file 1: Fig. S4e). In addition, promoter capture Hi-C data of 22 tissue types [40–43] and DNAse-hypersensitivity site (DHS) connections [44] were overlapped with the SV breakpoints to predict disruption of long range interactions over the breakpoint junctions (Additional file 1: Fig. S4f).Genes with at least a weak phenotype association and a weak SV effect are considered as T3 candidate genes. Genes were classified as T1 candidate drivers if they have a strong association with the phenotype and are strongly affected by the SV. Genes classified as T2 candidate driver can have a weak/medium phenotype association combined with a strong SV effect or they can have a medium/strong phenotype association with a weak SV effect (Fig. 2a, Table 1).

Table 1: Cut-offs used to classify affected genes as T1, T2 or T3 candidate driver genes.

| **1. Phenotype association** | | | | |
|---|---|---|---|---|
| | | Weak | Medium | Strong |
| Phenotypic score | pLI > 0.9<br>RVIS < 10<br>HI < 10<br>DD2GP<br>OMIM | >0 | >0 | >2 |
| Mode of inheritance | | | AD/XD/XR+XY | AD/XD/XR+XY |
| phenomatchScore | | >0 | >4 | >10 |

| | | | | | |
|---|---|---|---|---|---|
| Phenomatches with individual HPO terms | Score > 1<br><br>Score > 5 | >0 | >10% | >25% | |
| **2. Effect of SV on gene** | | | | | |
| | | Weak | | Strong | |
| Gene location | | Adjacent | Dup | Adjacent | DEL/TRUNC |
| Support score | TAD disrupted<br>V4C disrupted<br>PCHiC disrupted<br>DHS disrupted<br>RNA expression | >1 | NA | >3 | NA |
| **3. Driver classification** | | | | | |
| Classification | | T3 | T2 | | T1 |
| Phenotype association<br>+<br>Effect of SV on gene | | Weak<br>+<br>Weak | Strong<br>+<br>Weak | Medium<br>+<br>Strong | Strong<br>+<br>Strong |

pLI: probability of being loss-of-function intolerant, RVIS: Residual Variation Intolerance Score, HI: Haploinsufficiency, DD2GP: Developmental Disorders Genotype-Phenotype Database, OMIM: Online Mendelian Inheritance in Man, AD: Autosomal dominant, XD: X-linked dominant, XR: X-linked recessive, XY: Y-linked, TAD: topologically associated domain, V4C: virtual 4C, PCHiC: promoter capture Hi-C, DHS: DNase hypersensitivity site.

**SV and phenotype information large patient cohorts**

Breakpoint junction information and HPO terms for 228 individuals (excluding the individuals already included in this study for WGS and RNA-seq analysis) with mostly balanced SVs were obtained from Redin et al [51]. Phenotype and genomic information for 154 patients with *de novo* copy number variants ascertained by clinical genomic arrays were obtained from an inhouse patient database from the University Medical Center Utrecht (the Netherlands).

**List of abbreviations**

HPO: Human Phenotype Ontology; kb: kilobase; RPKM: reads per kilobase per million mapped reads; SNV: Single nucleotide variant; SV: Structural variant; TAD: topologically associated domain; VUS: Variant of unknown significance; WGS: Whole-genome sequencing

## Declarations

### Ethics approval and consent to participate

All individuals or their parents provided written informed consent to participate in the study. The study was approved by the Medical Ethics Committee (METC) of the University Medical Center Utrecht (NL55260.041.15 15-736/M). The study was performed in accordance with the Declaration of Helsinki.

### Consent for publication

All participants in this study provided consent for publication.

### Availability of data and material

Whole genome sequencing and RNA sequencing datasets generated during the study have been deposited in the European Genome-phenome Archive (https://www.ebi.ac.uk/ega) under accession number EGAS00001003489. All custom code used in this study is available on https://github.com/UMCUGenetics/Complex_SVs.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This work is supported by the funding provided by the Netherlands Science Foundation (NWO) Vici grant (865.12.004) to Edwin Cuppen.

### Authors' contributions

SM and JV performed experiments and computational analyses. JG and RH ascertained and enrolled individuals P1 to P21 and provided phenotypic information. JK and SM cultured LCL cell lines. NB and LdlF performed DNA and RNA isolations and lab support. SB, RJ, MJvR and WK provided support for computational analyses. EvB collected genomic and phenotypic information of individuals U1-U111. GP provided material for individual P22. MET provided LCL cell lines and information of individuals P22-P39. SM, JV, JG, JK, and EC designed the study. SM, JV and EC wrote the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

We wish to thank all individuals who participated in this research. We thank Giulia Pregno and Giorgia Mandrile for the clinical and biological study of individual P22. We thank Utrecht Sequencing Facility (USEQ) for providing RNA-sequencing service and data. USEQ is subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University. We would also like to thank the Hartwig Medical Foundation for providing whole genome sequencing services.

## References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.

2. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38.

3. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19:325.

4. Hehir-Kwa JY, Pfundt R, Veltman JA, de Leeuw N. Pathogenic or not? Assessing the clinical relevance of copy number variants. Clin Genet. 2013;84:415–21.

5. Nowakowska B. Clinical interpretation of copy number variants in the human genome. J Appl Genet. 2017;58:449–57.

6. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol. 2016;17:771–82.

7. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14:390–403.

8. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:772.

9. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet. 2018;19:789-800.

10. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.

11. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538:265–9.

12. Will AJ, Cova G, Osterwalder M, Chan W-L, Wittler L, Brieske N, et al. Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). Nat Genet. 2017;49:1539–45.

13. Kragesteen BK, Spielmann M, Paliou C, Heinrich V, Schöpflin R, Esposito A, et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat Genet. 2018;50:1463–73.

14. Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018;19:453-467.

15. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.

16. Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang N, Hurles ME, et al. Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. 2014;15:423.

17. Zepeda-Mendoza CJ, Ibn-Salem J, Kammin T, Harris DJ, Rita D, Gripp KW, et al. Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. Am J Hum Genet. 2017;101:206–17.

18. Yauy K, Gatinois V, Guignard T, Sati S, Puechberty J, Gaillard JB, et al. Looking for Broken TAD Boundaries and Changes on DNA Interactions: Clinical Guide to 3D Chromatin Change Analysis in Complex Chromosomal Rearrangements and Chromothripsis. Methods Mol Biol. 2018;1769:353–61.

19. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85:457–64.

20. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45:D865–76.

21. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2018;47:D1018-27.

22. Brand H, Collins RL, Hanscom C, Rosenfeld JA, Pillalamarri V, Stone MR, et al. Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. Am J Hum Genet. 2015;97:170–6.

23. Nazaryan-Petersen L, Eisfeldt J, Pettersson M, Lundin J, Nilsson D, Wincent J, et al. Replicative and non-replicative mechanisms in the formation of clustered CNVs are indicated by whole genome characterization. PLoS Genet. 2018;14:e1007780.

24. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511:344–7.

25. Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. Cell Rep. 2018;23:1581–97.

26. Brewer MH, Chaudhry R, Qi J, Kidambi A, Drew AP, Menezes MP, et al. Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3. PLoS Genet. 2016;12:e1006177.

27. Haines B, Hughes J, Corbett M, Shaw M, Innes J, Patel L, et al. Interchromosomal Insertional Translocation at Xq26.3 Alters SOX3 Expression in an Individual With XX Male

617 Sex Reversal. J Clin Endocrinol Metab. 2015;100:E815–20.

618 28. Bunyan DJ, Robinson DO, Tyers AG, Huang S, Maloney VK, Grand FH, et al. X-Linked
619 Dominant Congenital Ptosis Cosegregating with an Interstitial Insertion of a Chromosome
620 1p21.3 Fragment into a Quasipalindromic Sequence in Xq27.1. Open Journal of Genetics.
621 2014;04:415–25.

622 29. DeStefano GM, Fantauzzo KA, Petukhova L, Kurban M, Tadin-Strapps M, Levy B, et al.
623 Position effect on FGF13 associated with X-linked congenital generalized hypertrichosis.
624 Proc Natl Acad Sci U S A. 2013;110:7790–5.

625 30. Zhu H, Shang D, Sun M, Choi S, Liu Q, Hao J, et al. X-linked congenital hypertrichosis
626 syndrome is associated with interchromosomal insertions mediated by a human-specific
627 palindrome near SOX3. Am J Hum Genet. 2011;88:819–26.

628 31. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of
629 protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

630 32. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to
631 functional variation and the interpretation of personal genomes. PLoS Genet.
632 2013;9:e1003709.

633 33. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER:
634 Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl
635 Resources. Am J Hum Genet. 2009;84:524–33.

636 34. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A
637 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.
638 Cell. 2014;159:1665–80.

639 35. Cai C, Langfelder P, Fuller TF, Oldham MC, Luo R, van den Berg LH, et al. Is human
640 blood a good surrogate for brain tissue in transcriptional studies? BMC Genomics.
641 2010;11:589.

642 36. Tylee DS, Kawaguchi DM, Glatt SJ. On the outside, looking in: A review and evaluation
643 of the comparability of blood and brain “-omes.” Am J Med Genet B Neuropsychiatr Genet.
644 2013;162:595–603.

645 37. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin
646 Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep.
647 2016;17:2042–59.

648 38. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, et al.
649 Chromosome conformation elucidates regulatory relationships in developing human brain.
650 Nature. 2016;538:523–7.

651 39. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional
652 genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.

653 40. Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, et al.
654 CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol.
655 2016;17:127.

656 41. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-
657 Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target
658 Gene Promoters. Cell. 2016;167:1369–84.e19.

659 42. Freire-Pritchett P, Schoenfelder S, Várnai C, Wingett SW, Cairns J, Collier AJ, et al.
660 Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic
661 stem cells. Elife. 2017;6:e21926.

662 43. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, et
663 al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in
664 terminal differentiation. Nat Genet. 2017;49:1522–8.

665 44. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The
666 accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

667 45. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen
668 A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

669 46. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods
670 for the identification of topologically associating domains. Genome Biol. 2018;19:217.

671 47. Dali R, Blanchette M. A critical assessment of topologically associating domain
672 prediction tools. Nucleic Acids Res. 2017;45:2994–3005.

673 48. Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, et al. Measuring the
674 reproducibility and quality of Hi-C data. Genome Biol. 2019;20:57.

675 49. Benko S, Fantes JA, Amiel J, Kleinjan D-J, Thomas S, Ramsay J, et al. Highly
676 conserved non-coding elements on either side of SOX9 associated with Pierre Robin
677 sequence. Nat Genet. 2009;41:359–64.

678 50. Amarillo IE, Dipple KM, Quintero-Rivera F. Familial microdeletion of 17q24.3 upstream of
679 SOX9 is associated with isolated Pierre Robin sequence due to position effect. Am J Med
680 Genet A. 2013;161A:1167–72.

681 51. Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, et al. The genomic
682 landscape of balanced cytogenetic abnormalities associated with human congenital
683 anomalies. Nat Genet. 2017;49:36–45.

684 52. Mehrjouy MM, Fonseca ACS, Ehmke N, Paskulin G, Novelli A, Benedicenti F, et al.
685 Regulatory variants of FOXG1 in the context of its topological domain organisation. Eur J
686 Hum Genet. 2018;26:186–96.

687 53. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J,
688 et al. Mapping and phasing of structural variation in patient genomes using nanopore
689 sequencing. Nat Commun. 2017;8:1326.

690 54. Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, et al. A
691 Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs.
692 Science. 2013;340:1234167–1234167.

693 55. Rodríguez-Carballo E, Lopez-Delisle L, Zhan Y, Fabre PJ, Beccari L, El-Idrissi I, et al.
694 The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of
695 antagonistic regulatory landscapes. Genes Dev. 2017;31:2264–81.

696 56. Fabre PJ, Leleu M, Mormann BH, Lopez-Delisle L, Noordermeer D, Beccari L, et al.
697 Large scale genomic reorganization of topological domains at the HoxD locus. Genome Biol.
698 2017;18:149.

699 57. Svensson AM, Curry CJ, South ST, Whitby H, Maxwell TM, Aston E, et al. Detection of a
700 de novo interstitial 2q microdeletion by CGH microarray analysis in a patient with limb

701 malformations, microcephaly and mental retardation. Am J Med Genet A. 2007;143A:1348–
702 53.

703 58. Mitter D, Chiaie BD, Lüdecke H-J, Gillessen-Kaesbach G, Bohring A, Kohlhase J, et al.
704 Genotype-phenotype correlation in eight new patients with a deletion encompassing 2q31.1.
705 Am J Med Genet A. 2010;152A:1213–24.

706 59. Montavon T, Thevenet L, Duboule D. Impact of copy number variations (CNVs) on long-
707 range gene regulation at the HoxD locus. Proc Natl Acad Sci U S A. 2012;109:20204–11.

708 60. Webster E, Cho MT, Alexander N, Desai S, Naidu S, Bekheirnia MR, et al. De novo
709 PHIP-predicted deleterious variants are associated with developmental delay, intellectual
710 disability, obesity, and dysmorphic features. Molecular Case Studies. 2016;2:a001172.

711 61. Engwerda A, Frentz B, den Ouden AL, Flapper BCT, Swertz MA, Gerkes EH, et al. The
712 phenotypic spectrum of proximal 6q deletions based on a large cohort derived from social
713 media and literature reports. Eur J Hum Genet. 2018;26:1478–89.

714 62. Laugsch M, Bartusel M, Rehimi R, Alirzayeva H, Karaolidou A, Crispatzu G, et al.
715 Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation
716 with Patient-Specific hiPSCs. Cell Stem Cell. 2019;24:736–52.e12.

717 63. Tan TY, Gonzaga-Jauregui C, Bhoj EJ, Strauss KA, Brigatti K, Puffenberger E, et al.
718 Monoallelic BMP2 Variants Predicted to Result in Haploinsufficiency Cause Craniofacial,
719 Skeletal, and Cardiac Features Overlapping Those of 20p12 Deletions. Am J Hum Genet.
720 2017;101:985–94.

721 64. Kostina A, Bjork H, Ignatieva E, Irtyuga O, Uspensky V, Semenova D, et al. Notch, BMP
722 and WNT/β-catenin network is impaired in endothelial cells of the patients with thoracic
723 aortic aneurysm. Atheroscler Suppl. 2018;35:e6–13.

724 65. Newman S, Hermetz KE, Weckselblatt B, Rudd MK. Next-generation sequencing of
725 duplication CNVs reveals that most are tandem and some create fusion genes at
726 breakpoints. Am J Hum Genet. 2015;96:208–20.

727 66. Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A,
728 et al. Characteristics of de novo structural changes in the human genome. Genome Res.
729 2015;25:792–801.

730 67. Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, et al. Paternally
731 inherited cis-regulatory structural variants are associated with autism. Science.
732 2018;360:327–31.

733 68. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera AV, et al. An open
734 resource of structural variation for medical and population genetics. bioRxiv. 2019:578674.

735 69. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
736 Bioinformatics. 2009;25:1754–60.

737 70. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
738 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
739 sequencing data. Genome Res. 2010;20:1297–303.

740 71. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS
741 alignment formats. Bioinformatics. 2015;31:2032–4.

742 72. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta:

743 rapid detection of structural variants and indels for germline and cancer sequencing
744 applications. Bioinformatics. 2016;32:1220–2.

745 73. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural
746 variant discovery by integrated paired-end and split-read analysis. Bioinformatics.
747 2012;28:i333–9.

748 74. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An
749 integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.

750 75. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
751 universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

752 76. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput
753 sequencing data. Bioinformatics. 2015;31:166–9.

754 77. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
755 RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

756 78. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of
757 genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols.
758 2009;4:1184–91.

759 79. Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, et al. The evolution of lineage-
760 specific regulatory activities in the human embryonic limb. Cell. 2013;154:185–96.

761

762 **Figures, tables and additional files**

763 **Fig. 1** Characterization of *de novo* SVs in a cohort of individuals with neurodevelopmental
764 disorders. **a** Frequencies of clinical phenotypic categories described for the 39 included
765 individuals based on categories defined by HPO. Nervous system abnormalities are divided
766 into four subcategories. **b** Number of *de novo* break junctions per SV type identified by WGS
767 of 39 included patients. Most detected *de novo* SVs are part of complex genomic
768 rearrangements involving more than three breakpoint junctions. **c** Number of cases in which
769 WGS analysis identified new, additional or similar SVs compared to microarray-based copy
770 number profiling. **d** Schematic representation of additional genomic rearrangements that were
771 observed by WGS in five individuals. For each patient, the top panel shows the *de novo* SVs
772 identified by arrays or karyotyping and bottom panel shows the structures of the SVs detected
773 by WGS. The WGS data of individual P8 revealed complex chromoanasynthesis
774 rearrangements involving multiple duplications and an insertion of a fragment from chr14 into
775 chr3. Individual P11 has an insertion of a fragment of chr9 into chrX that was detected as a

776 copy number gain by array-based analysis (Fig. S2). The detected copy number gains in
777 individuals P12 and P21 show an interspersed orientation instead of a tandem orientation.
778 The translocation in patient P20 appeared to be more complex than previously anticipated
779 based on karyotyping results, showing 11 breakpoint junctions on three chromosomes.
780
781 **Fig. 2** Prediction of candidate driver genes directly and indirectly affected by the SVs. **a**
782 Schematic overview of the computational workflow developed to detect candidate driver
783 genes. Classification of genes at (direct) or surrounding (indirect) the *de novo* SVs is based
784 on association of the gene with the phenotype and the predicted direct or indirect effect on the
785 gene (Table 1). **b** Total number of identified tier 1, 2 and 3 candidate driver genes predicted
786 to be directly or indirectly affected by an SV. **c** Genome browser overview showing the
787 predicted disruption of regulatory landscape of the *HOXD* locus in individual P22. A 107 kb
788 fragment (red shading) upstream of the *HOXD* locus (green shading) is translocated to a
789 different chromosome and a 106 kb fragment (yellow shading) is inverted. The SVs affect the
790 TAD centromeric of the *HOXD* locus which is involved in the regulation of gene expression in
791 developing digits. The translocated and inverted fragments contain multiple mouse [55] and
792 human (day E41) [79] embryonic limb enhancers, including the global control region (GCR).
793 Disruptions of these developmental enhancers likely contributed to the limb phenotype of the
794 patient. The virtual V4C track shows the HiC interactions per 10kb bin in germinal zone (GZ)
795 cells using the *HOXD13* gene as viewpoint [38]. The bottom track displays the PCHiC
796 interactions of the *HOXD13* gene in neuroectodermal cells [42]. UCSC Liftover was used to
797 convert mm10 coordinates to hg19. **d** RNA expression levels of genes at or adjacent to *de*
798 *novo* SVs. Log2 fold RNA expression changes compared to controls (see methods)
799 determined by RNA sequencing for expressed genes (RPKM >0.5) that are located within 2
800 Mb of SV breakpoint junctions (FLANK) or that are inverted (INV), duplicated (DUP), deleted
801 (DEL) or truncated (TRUNC). Differentially expressed genes ($p < 0.05$, calculated by DESeq2)
802 are displayed in red.
803

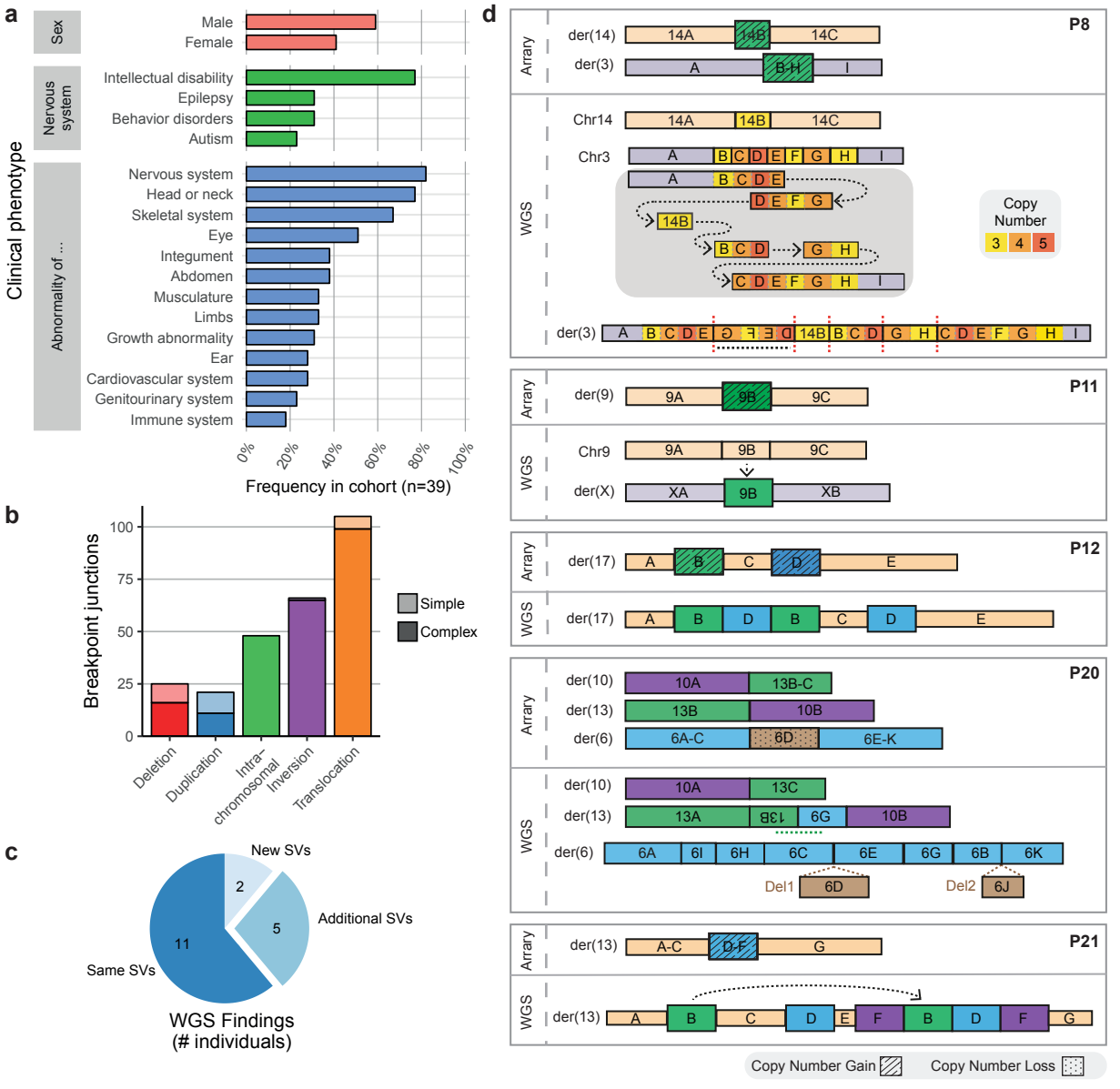
804 **Fig. 3** SVs can affect multiple candidate drivers which jointly contribute to a phenotype. **a**
805 Number of patients whose phenotype can be partially or largely explained by the predicted
806 T1/T2 candidate drivers. These molecular diagnoses are based on the fraction of HPO terms
807 assigned to the patients that have a phenomatchScore of more than five with at least one
808 T1/T2 driver gene. **b** Scatterplot showing the number of predicted T1/T2 candidate drivers
809 compared to the total number of genes at or adjacent (< 2Mb) to the *de novo* SVs per patient.
810 **c** Heatmap showing the association of the four predicted T1/T2 candidate drivers with the
811 phenotypic features (described by HPO terms) of individual P25. The numbers correspond to
812 score determined by Phenomatch. The four genes are associated with different parts of the
813 complex phenotype of the patient. **d** Ideogram of the derivative (der) chromosomes 6, 12 and
814 20 in individual P25 reconstructed from the WGS data. WGS detected complex
815 rearrangements with six breakpoint junctions and two deletions on chr6 and chr20 of
816 respectively ~10 Mb and ~0.6 Mb. **e** Circos plot showing the genomic regions and candidate
817 drivers affected by the complex rearrangements in individual P25. Gene symbols of T1/T2 and
818 T3 candidate drivers are shown in respectively red and black. The break junctions are
819 visualized by the lines in the inner region of the plot (red lines and highlights indicate the
820 deletions). The middle ring shows the log2 fold change RNA expression changes in
821 lymphoblastoid cells derived from the patient compared to controls measured by RNA
822 sequencing. Genes differentially expressed ($p<0.05$) are indicated by red (log2 fold change <
823 -0.5) and blue (log2 fold change > 0.5) bars. The inner ring shows the organization of the TADs
824 and their boundaries (indicated by vertical black lines) in germinal zone (GZ) brain cells [38].
825 TADs overlapping with the *de novo* SVs are highlighted in red.
826
827 **Fig. 4** *In silico* prediction of candidate drivers in larger cohorts of patients with *de novo* SVs. **a**
828 Comparison between previous SV classifications with the strongest candidate driver (located
829 at or adjacent (<1 Mb) to these SVs) predicted by our approach. Two different patient cohorts,
830 one containing mostly balanced SVs [51] and one containing copy number variants, were
831 screened for candidate drivers. Our method identified T1/T2 candidate drivers for most SVs

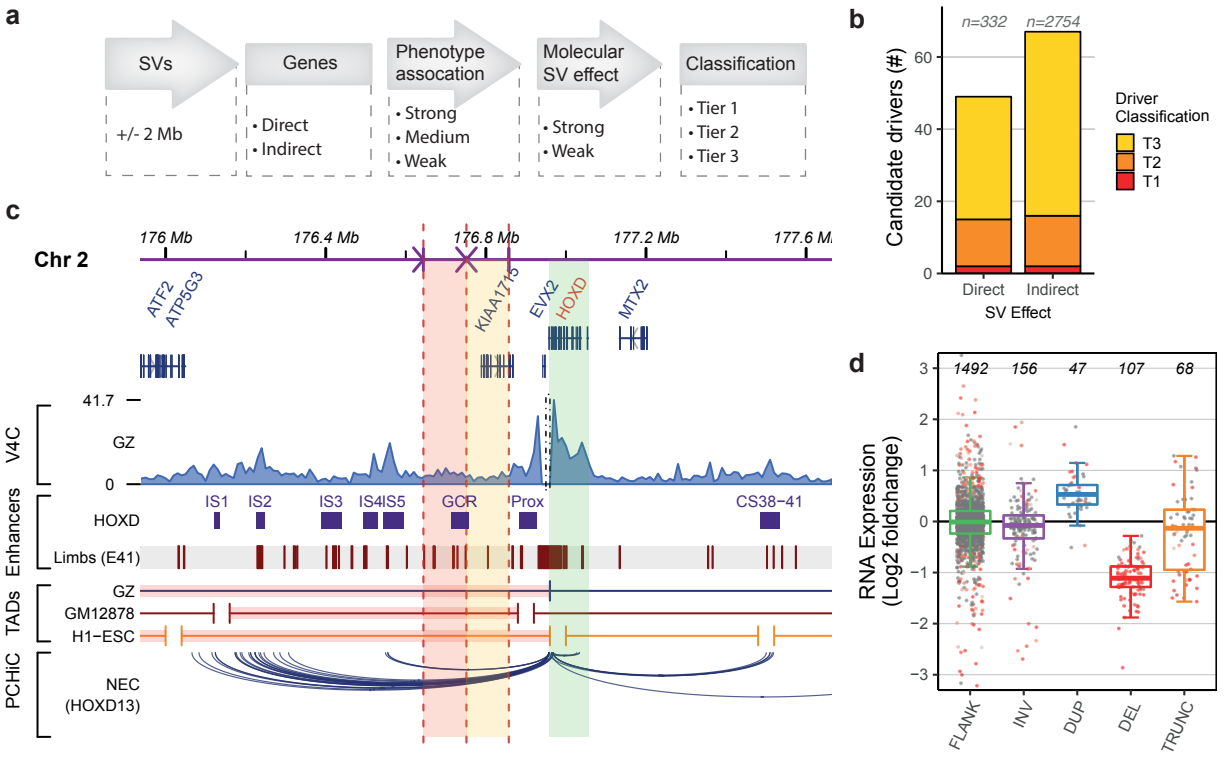
previously classified as pathogenic or likely pathogenic. Additionally, the method detected T1/T2 candidate drivers for some SVs previously classified as VUS, which may lead to a new molecular diagnosis. **b** Quantification of the predicted effects of the SVs on proposed T1/T2 candidate driver genes per cohort. Individuals with multiple directly and indirectly affected candidate drivers are grouped in the category described as “Both”. Indirect positional effects of SVs on genes contributing to phenotypes appear to be more common in patients with balanced SVs compared to patients with copy number variants.

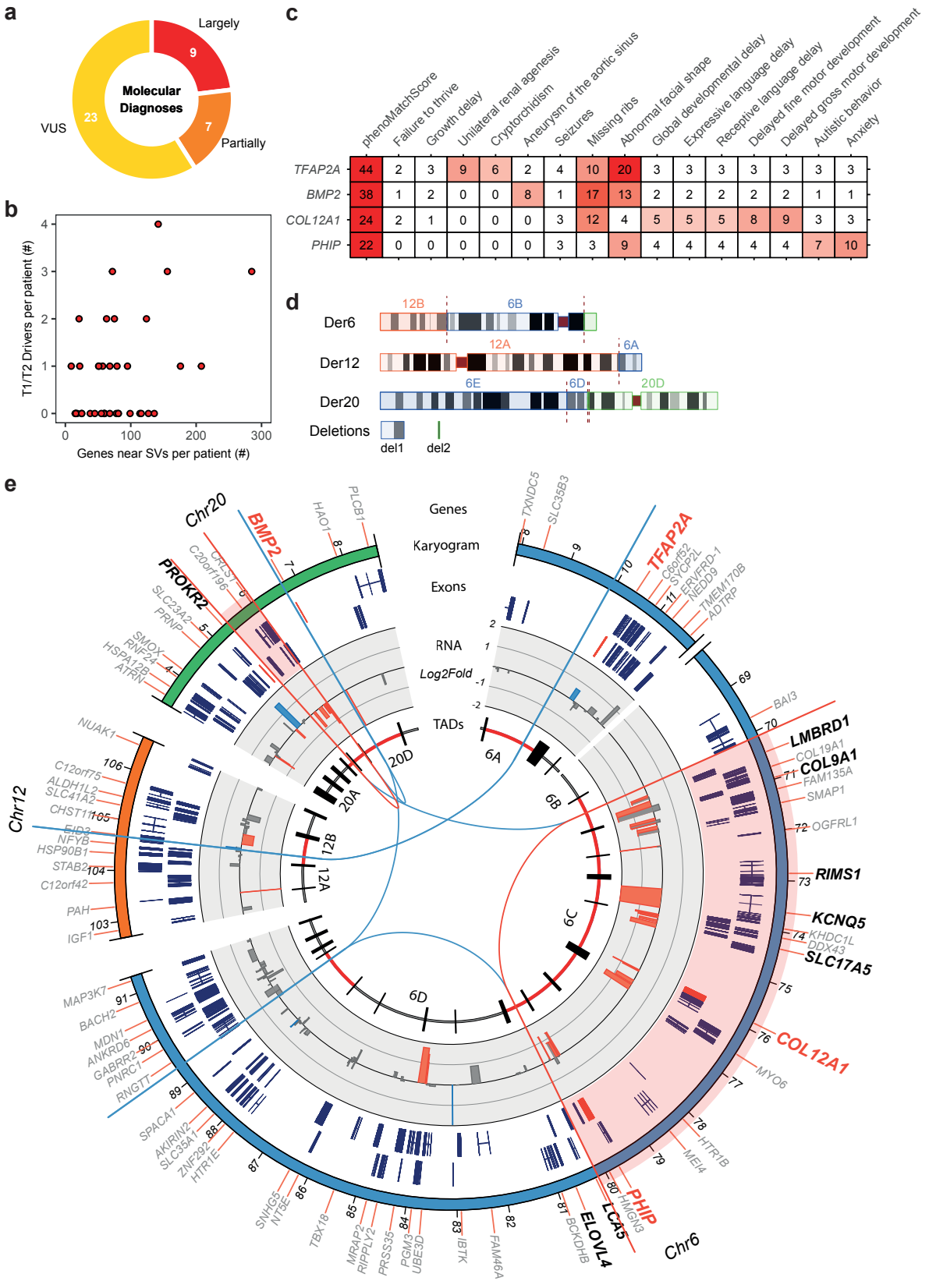
**Additional files**

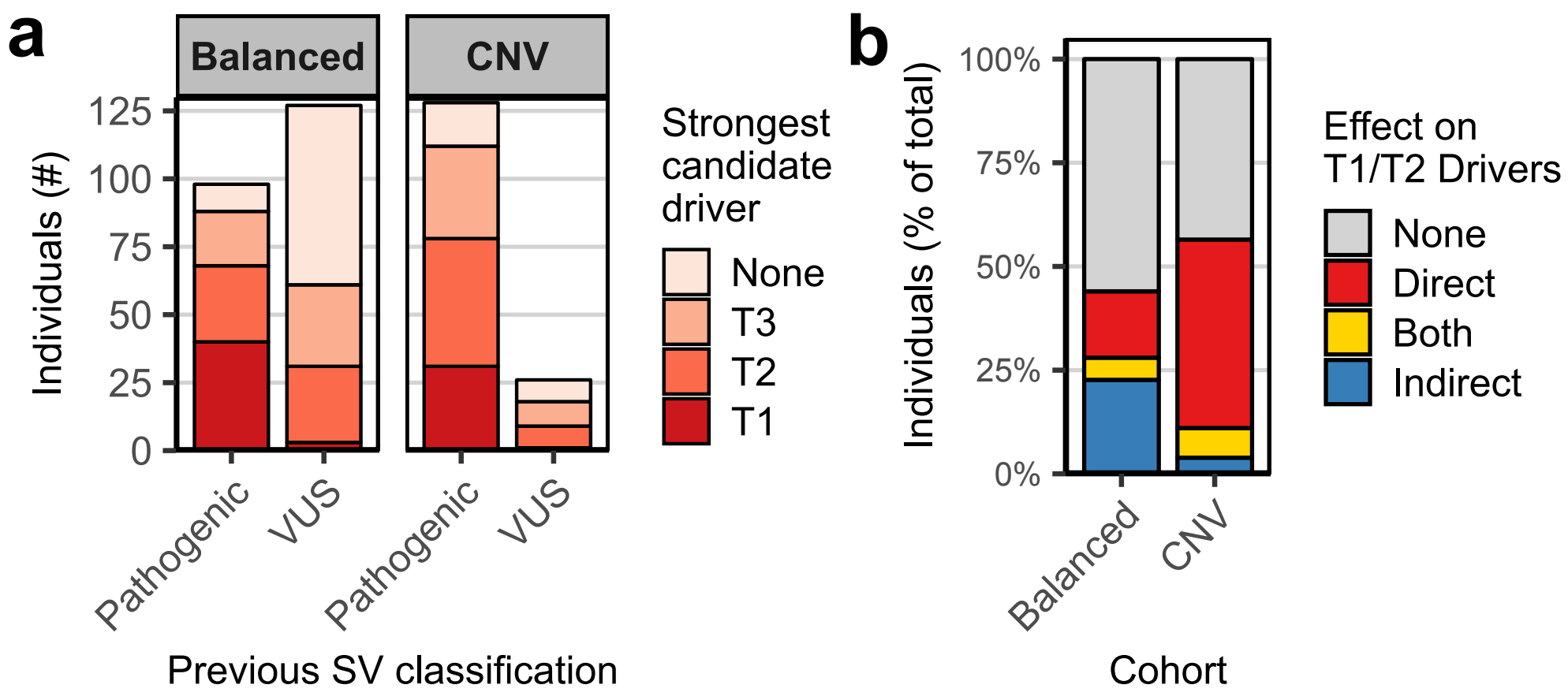
**Additional file 1 (pdf):** Figures S1 to S8, including figure legends and supplemental references.

**Additional file 2 (xlsx): Table S1.** Phenotype information of the 39 included patients with *de novo* SVs. **Table S2.** Coordinates of the *de novo* SV breakpoint junctions detected in the 39 individuals by WGS. **Table S3.** Candidate driver genes detected for each included patient. **Table S4.** Fusion genes detected in the patients by RNA sequencing. **Table S5.** List of external data sources used in this study. **Table S6.** Detected *de novo* copy number variants in 154 patients of the diagnostics cohort. **Table S7**. Candidate driver genes detected in patients’ cohorts.


a
Clinical phenotype
Sex
Male
Female
Nervous system
Intellectual disability
Epilepsy
Behavior disorders
Autism
Abnormality of ...
Nervous system
Head or neck
Skeletal system
Eye
Integument
Abdomen
Musculature
Limbs
Growth abnormality
Ear
Cardiovascular system
Genitourinary system
Immune system
0%
20%
40%
60%
80%
100%
Frequency in cohort (n=39)
b
Breakpoint junctions
100
75
50
25
0
Simple
Complex
Deletion
Duplication
Intra-chromosomal
Inversion
Translocation
c
New SVs
2
Additional SVs
5
11
Same SVs
WGS Findings
(# individuals)
d
P8
Array
der(14)
14A
14B
14C
der(3)
A
B-H
I
WGS
Chr14
Chr3
Copy Number
3 4 5
der(3)
P11
der(9)
9A
9B
9C
Chr9
der(X)
XA
XB
P12
der(17)
P20
der(10)
10A
13B-C
der(13)
13B
10B
der(6)
6A-C
6D
6E-K
13C
13A
6G
6A
6I
6H
6C
6E
6B
6K
Del1
Del2
6J
P21
der(13)
A-C
D-F
G
Copy Number Gain
Copy Number Loss


a
SVs
+/- 2 Mb
Genes
• Direct
• Indirect
Phenotype assocation
• Strong
• Medium
• Weak
Molecular SV effect
• Strong
• Weak
Classification
• Tier 1
• Tier 2
• Tier 3
b
n=332
n=2754
Candidate drivers (#)
Direct
Indirect
SV Effect
Driver Classification
T3
T2
T1
c
Chr 2
176 Mb
176.4 Mb
176.8 Mb
177.2 Mb
177.6 Mb
ATF2
ATP5G3
KIAA1715
EVX2
HOXD
MTX2
V4C
41.7
GZ
0
Enhancers
IS1
IS2
IS3
IS4IS5
GCR
Prox
CS38−41
HOXD
Limbs (E41)
TADs
GZ
GM12878
H1−ESC
PCHiC
NEC
(HOXD13)
d
RNA Expression
(Log2 foldchange)
1492
156
47
107
68
FLANK
INV
DUP
DEL
TRUNC


a
Largely
9
Molecular Diagnoses
VUS
23
7
Partially
b
T1/T2 Drivers per patient (#)
Genes near SVs per patient (#)
c
phenoMatchScore
Failure to thrive
Growth delay
Unilateral renal agenesis
Cryptorchidism
Aneurysm of the aortic sinus
Seizures
Missing ribs
Abnormal facial shape
Global developmental delay
Expressive language delay
Receptive language delay
Delayed fine motor development
Delayed gross motor development
Autistic behavior
Anxiety
TFAP2A 44 2 3 9 6 2 4 10 20 3 3 3 3 3 3 3
BMP2 38 1 2 0 0 8 1 17 13 2 2 2 2 2 1 1
COL12A1 24 2 1 0 0 0 3 12 4 5 5 5 8 9 3 3
PHIP 22 0 0 0 0 0 3 3 9 4 4 4 4 4 7 10
d
12B
6B
Der6
12A
6A
Der12
6E
6D
20D
Der20
Deletions
del1
del2
e
Chr20
BMP2
PROKR2
Genes
Karyogram
Exons
RNA
Log2Fold
TADs
TFAP2A
Chr12
Chr6
LMBRD1
COL9A1
RIMS1
KCNQ5
SLC17A5
COL12A1
PHIP
ELOVL4
LCA5


a
Balanced
CNV
Individuals (#)
125
100
75
50
25
0
Pathogenic
VUS
Pathogenic
VUS
Previous SV classification
Strongest candidate driver
None
T3
T2
T1
b
Individuals (% of total)
100%
75%
50%
25%
0%
Balanced
CNV
Cohort
Effect on T1/T2 Drivers
None
Direct
Both
Indirect

# Supplemental materials

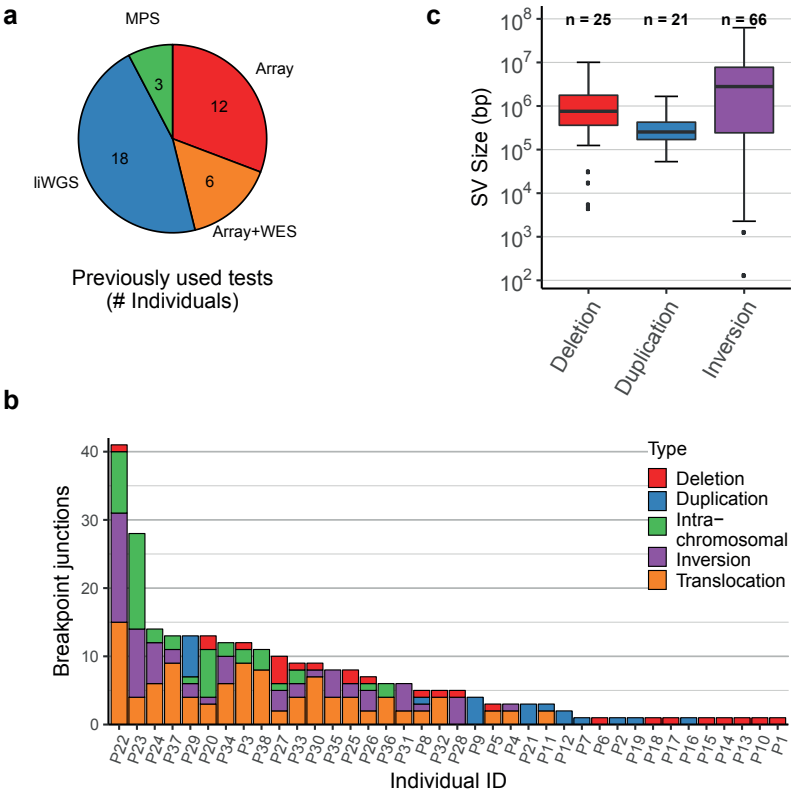



**Fig. S1** Detected *de novo* germline SVs in 39 included patients. **a** Genetic tests previously used in a clinical setting to identify the *de novo* SVs in the included individuals. Microarrays (ArrayCGH or SNP arrays) were used to detect the deletions and duplications in 18 of the included individuals. MPS: Mate-pair sequencing, WES: Whole Exome Sequencing, liWGS: long-insert Whole Genome Sequencing. **b** Number of identified *de novo* SV breakpoint junctions per individual. **c** Size distribution in base pairs (bp) of the identified *de novo* deletions (median size 757,378 bp), duplications (median size 253,729 bp) and inversions (median size 2,295,988 bp).

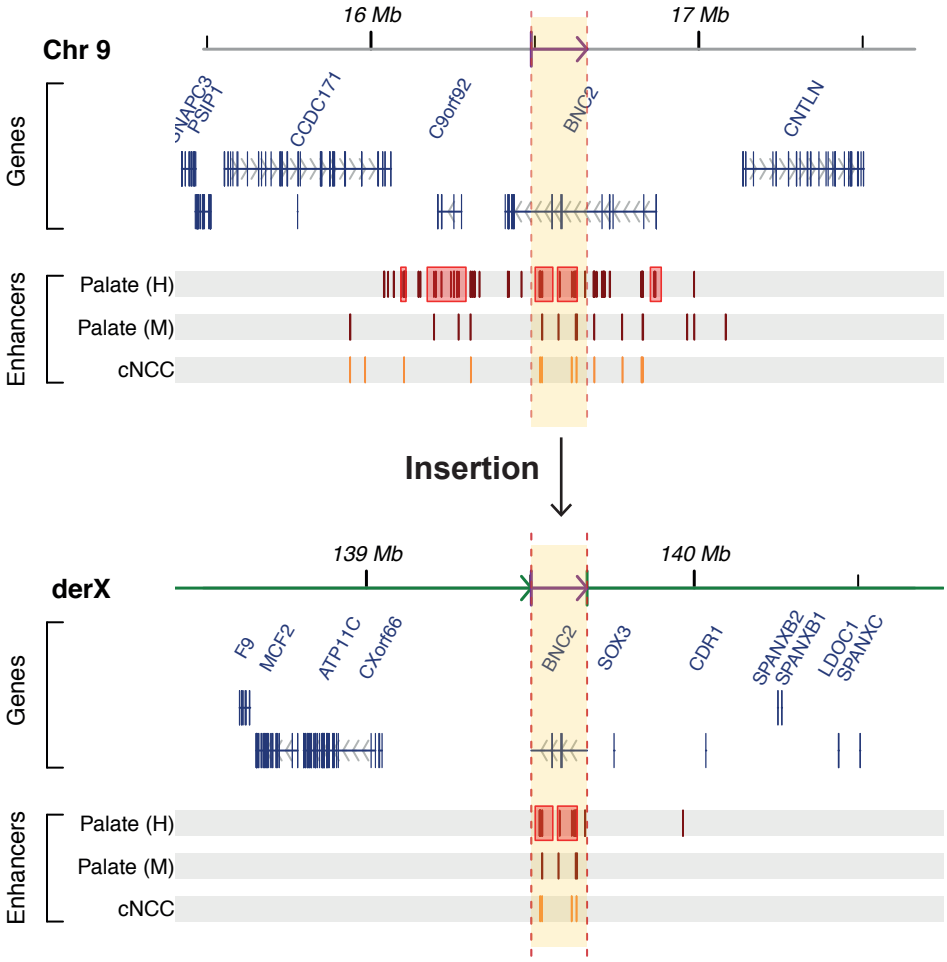



**Fig. S2** Insertion of a super-enhancer region upstream of *SOX3* detected by WGS in individual P11. A 170kb duplication in the *BNC2* gene body at chr9 was reported by array-based analysis (top panel), but WGS detected that this duplication is actually inserted in chrX (bottom panel). The fragment (highlighted in yellow) is inserted 82 kb upstream of the *SOX3* gene. This locus at chrX contains a palindromic sequence that is susceptible for formation of genomic rearrangement. Multiple patients with varying phenotypes and different insertions at this locus have been described [1–5]. The inserted fragment from chr9 contains multiple enhancers, including two previously described super-enhancer clusters (highlighted by red boxes), that are active in human (Palate (H), Carnegie stage 13) and mouse (Palate (M), embryonic day 11.5) craniofacial development and human cultured cranial neural crest cells (cNCC) [6–8]. The inserted enhancers may disturb the normal expression of the *SOX3* gene and/or the surrounding genes, which may have led to the cleft palate phenotype in this patient. Genomic coordinates of mouse (mm9) embryonic craniofacial enhancers (determined by p300 ChIP-seq [6]) were converted to hg19 coordinates using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

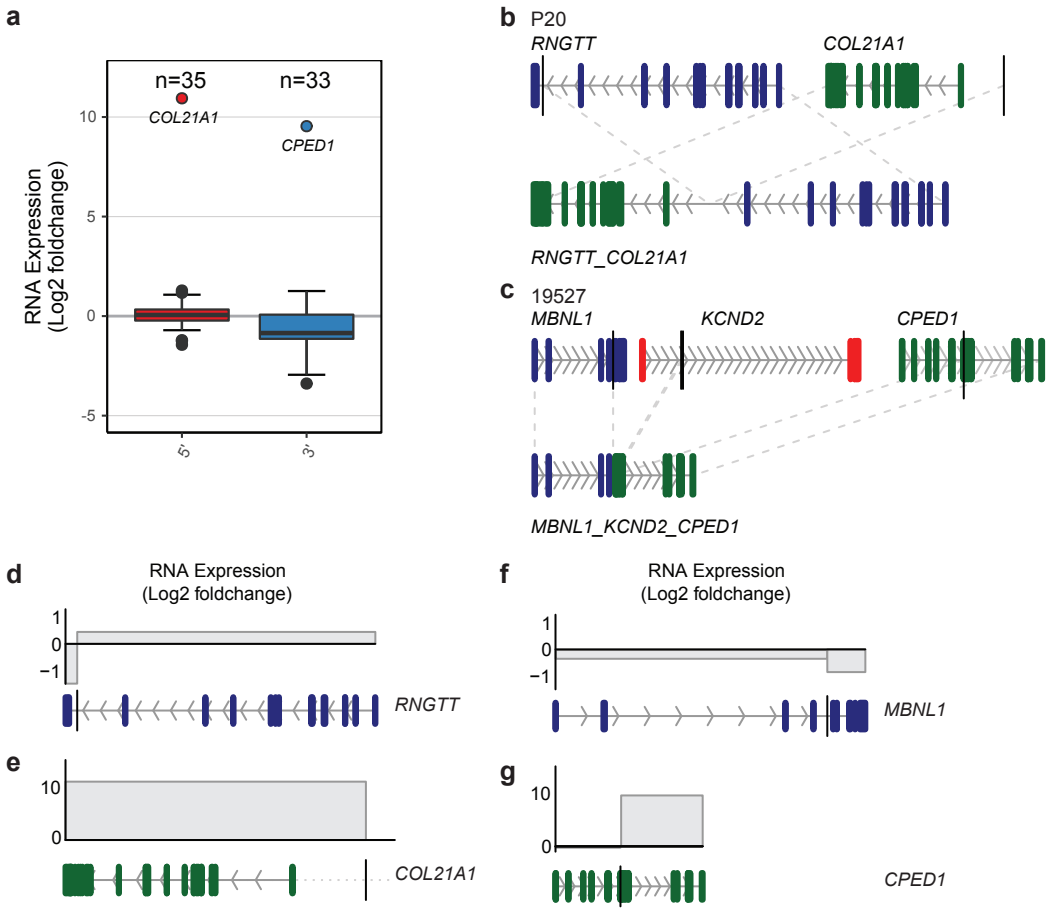



**Fig. S3** RNA expression of genes truncated by de novo germline SVs. **a** Log2 fold change expression values (compared to expression of the exons in control individuals) for 5' gene fragments and 3' gene fragments of truncated genes. The 5' fragment of *COL21A1* and the 3' fragment of *CPED1* show a strong overexpression due to a gene fusion. **b** Schematic representation of the *RNGTT_COL21A1* fusion gene caused by genomic rearrangements in individual P20. The breakpoint junctions near the *RNGTT* (ENST00000369485) and *COL21A1* (ENST00000244728) gene bodies are depicted by the vertical black lines. **c** Schematic reconstruction of the *MBNL1_KCND2_CPED1* fusion gene in individual P34. Breakpoint junctions in the truncated genes *MBNL1* (ENST00000324210), *KCND2* (ENST00000331113) and *CPED1* (ENST00000310396) are represented by the vertical black lines. **d - g** RNA log2 fold change expression values (compared to the expression in unaffected individuals) for the fragments of the truncated genes *RNGTT*, *COL21A1*, *MLBNL1* and *CPED1*.

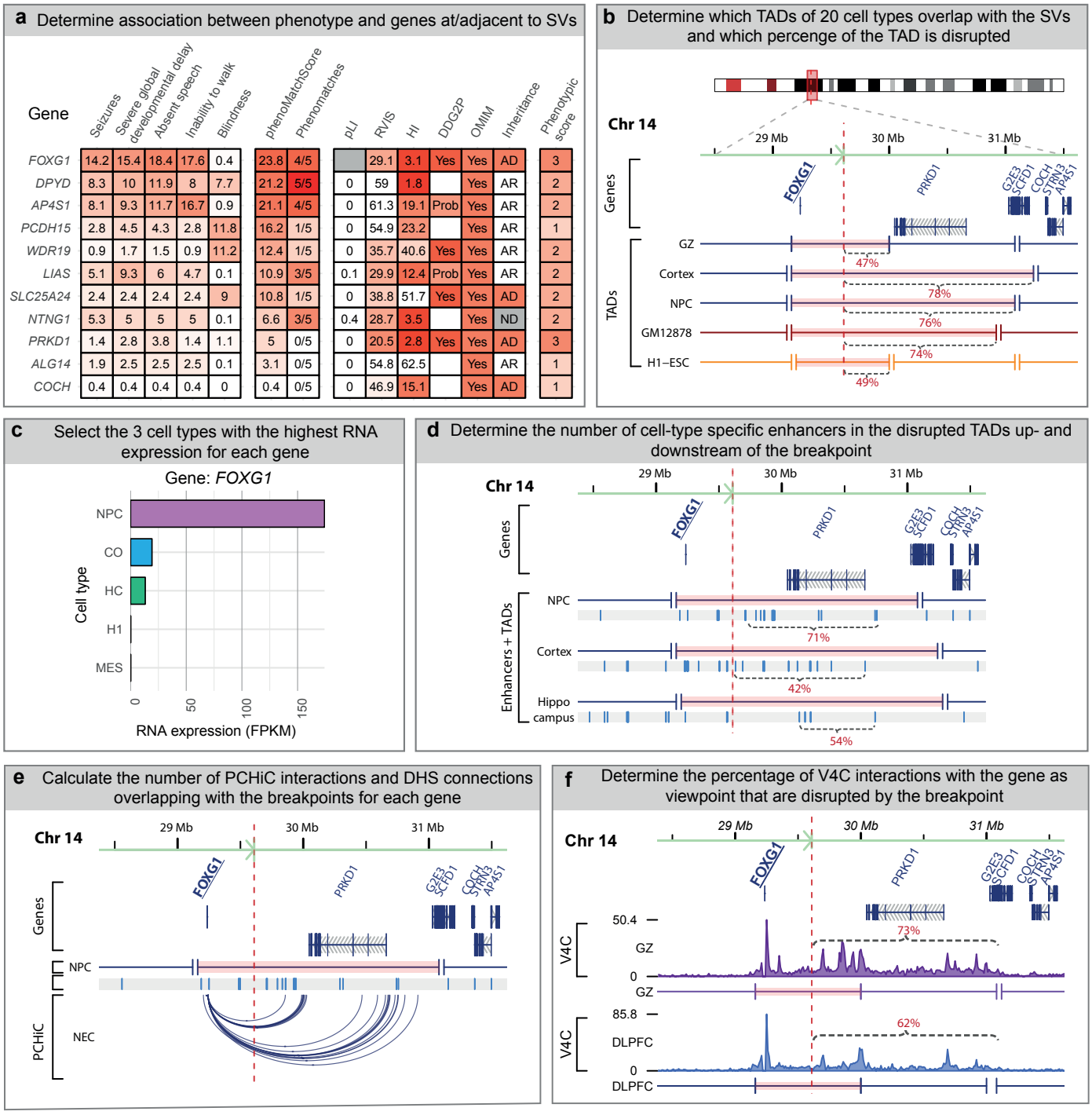



**Fig. S4** Schematic overview of computational strategy used to predict positional effects. **a** The association of a gene at or adjacent to an SV with the patient-specific phenotype is based on the Phenomatch score, the number of phenomatches, mode of inheritance and the phenotypic score. Each gene has a fixed phenotypic score (ranging from 1 to 5) based on its pLI (>0.9), RVIS (<10) and haploinsufficiency (HI, <10) scores and the presence of the gene in DDG2P and OMIM. **b** The TADs of 20 different cell types are overlapped with the SV breakpoint junctions of the individual. The TADs affected by a breakpoint are split into fragments up- and downstream of the breakpoint and the size of each fragment (relative to the size of the intact TAD) is calculated. Subsequently the genes located on each fragment are determined and each gene receives a score based on the relative size of the fragment. For example, 47% of the TAD in germinal zone (GZ) cells containing *FOXG1* is considered disrupted. **c** For each gene the 3 cell types with the highest RNA expression (FPKM: Fragments Per Kilobase Million) based on the Encode/Roadmap ChIP-seq data are selected. *(Continued on next page)*

**d** For each gene, enhancers from the three selected cell types are overlapped with the disrupted TAD fragments. The number of enhancers in the disrupted part of the TAD is compared to the number of enhancers in the TAD fragment containing the gene. This ratio is considered at the percentage of enhancers moved away from the gene (for example, the location of 71% of neural progenitor cell enhancers in the *FOXG1*TAD is changed). **e** For each gene, PCHiC interactions of 22 cell types and promoter-DHS connections are overlapped with the breakpoint junctions. The number of interactions overlapping with the junctions is divided by the total number of interactions of the gene. For example, for *FOXG1*all 107 PCHiC interactions (in multiple cell types) overlap with the breakpoint junction. **f** Virtual 4C profiles were generated for each gene and these were overlapped with the breakpoint junctions to determine the percentage of interactions that are located up- or downstream of the breakpoint junction. For *FOXG1*, 73% of the V4C interactions in dorsolateral prefrontal cortex (DLPFC) cells are considered to be disrupted.

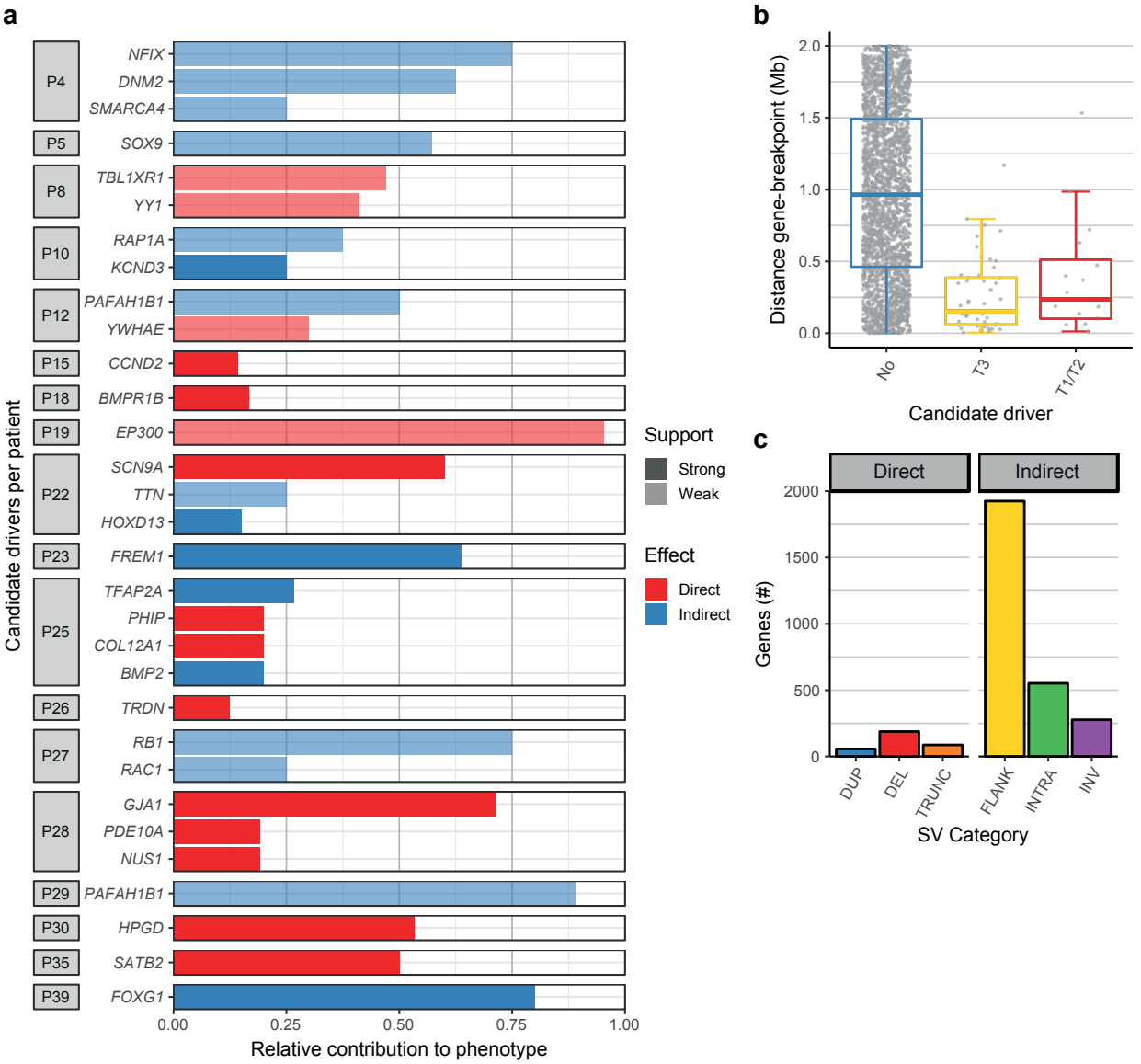



**Fig. S5** Overview of the detected candidate driver genes. **a** Relative contributions of the candidate drivers to the phenotypes of the individuals. The contributions are based on the number of Phenomatch hits (phenomatchScore > 5) of a gene with each individual HPO term assigned to an individual, e.g. a gene with a contribution of 0.75 is associated with 75% of the HPO terms of an individual. Shading indicates if there is relatively weak or strong evidence for an effect on the candidate driver. **b** Genomic distance (in base pairs) between the indirectly affected candidate driver genes (adjacent to the SVs) and the closest breakpoint junction. Most predicted candidate drivers are located within 1 Mb of a breakpoint junction. **c** Total number of analysed genes per SV category. DUP: Duplication, DEL: Deletion, TRUNC: Truncation, FLANK: Flanking region (+/- 2Mb), Intra: Intrachromosomal rearrangement, INV: Inversion.

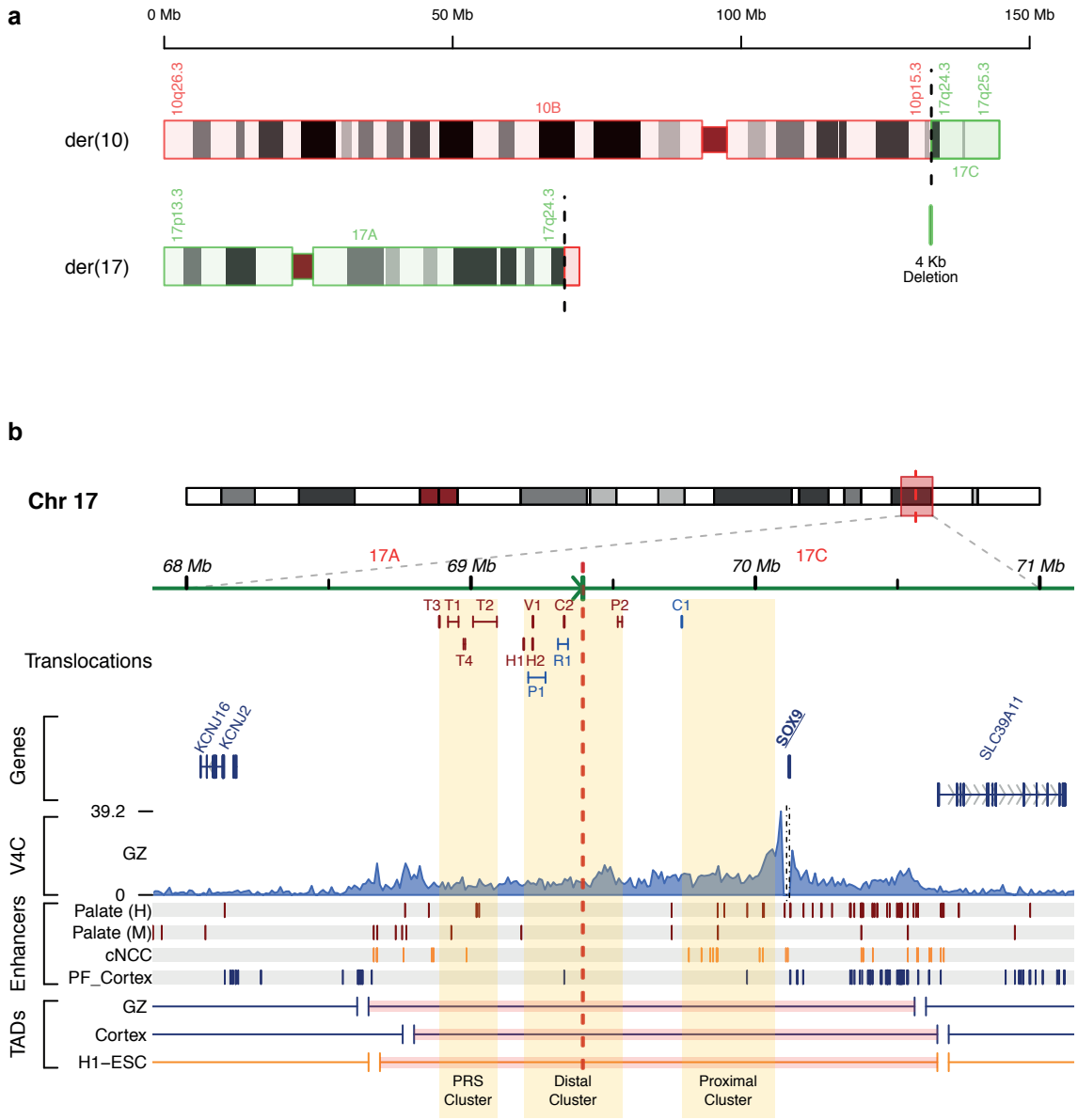



**Fig. S6** Prediction of positional effects of a translocation on SOX9 in individual P5. **a** Ideogram of the derivative chromosomes in individual P5. WGS identified a de novo translocation between chromosome 10 and 17 (46,XY,t(10;17)(p15;q24)). The breakpoint on chr17 (chr17:69395684, indicated by the vertical dotted red line) is 721 kb upstream of *SOX9*. A small 4kb fragment from chr17 is deleted (chr17:69391279-69395683). **b** Genome browser overview showing region surrounding the translocation breakpoint (red dotted line) at chromosome 17 in individual P5. The phenotype of this individual is characterized by acampomelic campomelic dysplasia and Pierre-Robin Syndrome including cleft palate, micrognathia and a long philtrum. SVs including translocations have been detected upstream of *SOX9* in individuals with various phenotypes including campomelic dysplasia. The translocations found in patients with phenotypes including cleft palate are shown in red and translocations found in patients with different phenotypes are depicted in blue. These translocations are predicted to separate *SOX9* from enhancers active in the developing palate, which may lead to the cleft palate phenotypes. Information about the other patients was obtained from the following publications: T1+T2+T3 [9]; T4 [10]; C1+C2 [11]; P1+P2 [12]; V1 [13]; R1 [14]; H1+H2 [15].

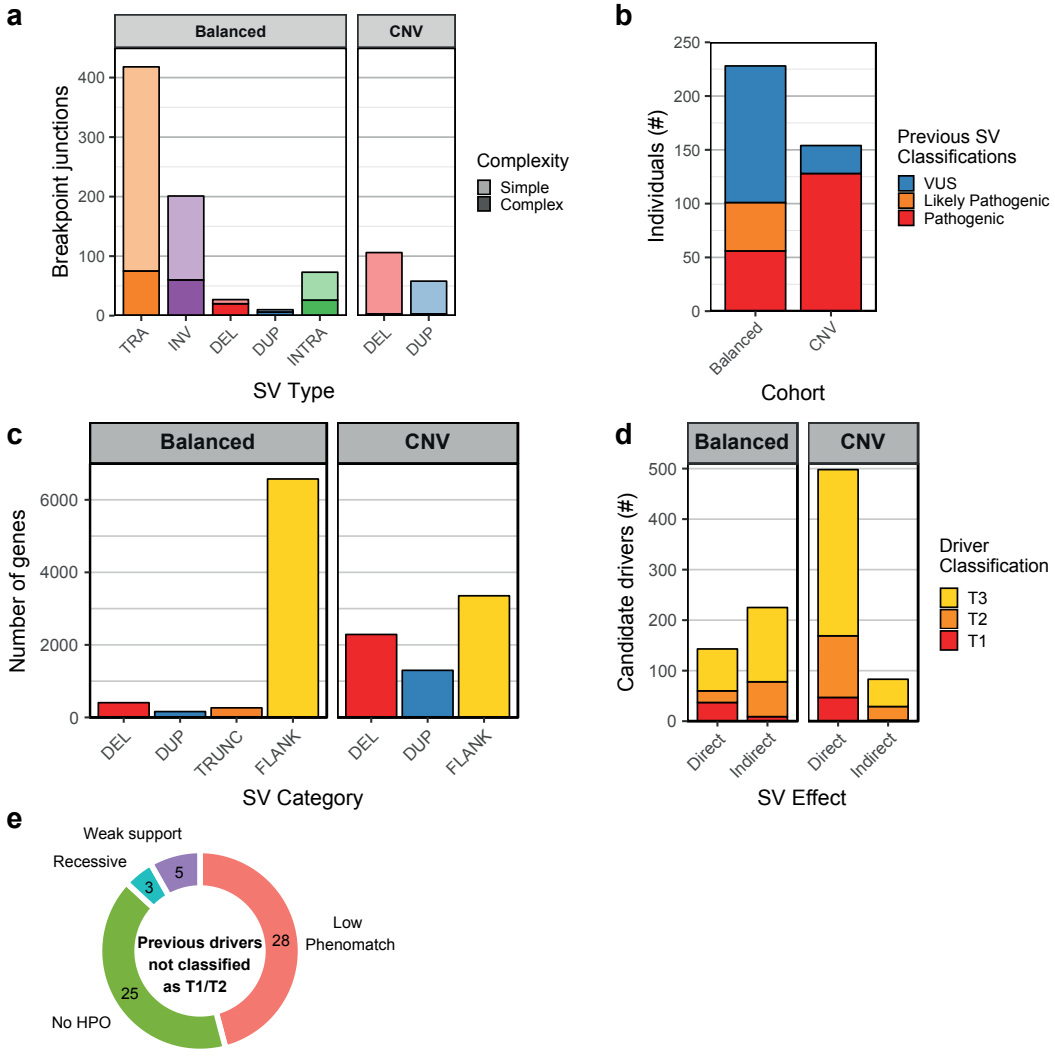



**Fig. S7** Overview of SVs and candidate drivers in two cohorts of patients with *de novo* SVs. **a** Quantification of previously identified *de novo* SVs in a cohort containing patients with mostly balanced SVs and a cohort containing patients with copy number variants (CNV). *De novo* translocations (TRA), inversions (INV) and intra-chromosomal rearrangements (INTRA) are most prevalent in the cohort of patients with balanced SVs. Some patients have complex genomic rearrangements (>3 SVs) including some deletions (DEL) or duplications (DUP). The cohort labelled as "CNV" consist of patients with relatively simple deletions and duplications (<10 Mb in size). **b** Number of patients whose *de novo* SVs were previously classified as pathogenic, likely pathogenic or variant of unknown significance (VUS) per cohort. **c** Total number of analysed genes per SV category in the two cohorts. Dup: Duplicated, Del: Deleted, Trunc: Truncated, Flank: Flanking SVs (<1 Mb). **d** Total number of predicted directly and indirectly affected candidate drivers per cohort. **e** Quantification of the genes that were previously classified as pathogenic or likely pathogenic (by Redin et al [16]), but not identified as T1 or T2 candidate driver by our approach. These classification differences may be caused by a lack of HPO terms associated with the gene, low phenomatch scores below the threshold of our method, insufficient (weak) support for an effect of an SV on the gene detected by our method or a presumed recessive mode of inheritance.

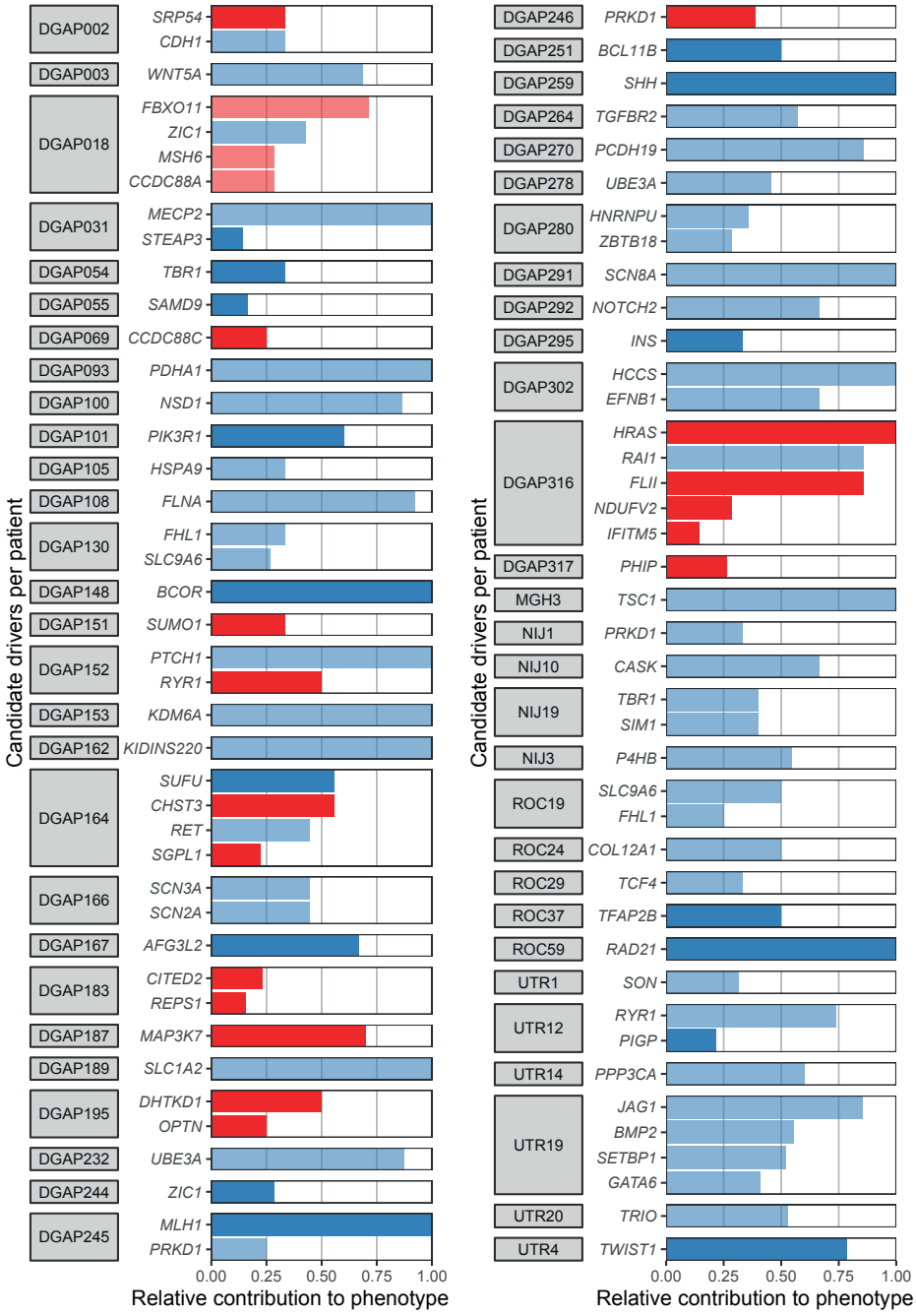



**Fig. S8** Predicted contributions of candidate drivers to the phenotypes of patients with balanced structural variants of unknown significance. T1/T2 candidate drivers were detected in 31 patients whose *de novo* SVs were previously classified as VUS by Redin et al [16]. The contributions to the phenotypes are based on the number of Phenomatch hits (phenomatchScore > 5) of the gene with each individual HPO term used to describe the phenotype of a patient. Shading indicates if there is relatively weak or strong evidence for an effect on the candidate driver.

# Supplemental references

1. Brewer MH, Chaudhry R, Qi J, Kidambi A, Drew AP, Menezes MP, et al. Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3. PLoS Genet. 2016;12:e1006177.

2. Haines B, Hughes J, Corbett M, Shaw M, Innes J, Patel L, et al. Interchromosomal insertional translocation at Xq26.3 alters SOX3 expression in an individual with XX male sex reversal. J Clin Endocrinol Metab. 2015;100:E815–20.

3. Bunyan DJ, Robinson DO, Tyers AG, Huang S, Maloney VK, Grand FH, et al. X-Linked Dominant Congenital Ptosis Cosegregating with an Interstitial Insertion of a Chromosome 1p21.3 Fragment into a Quasipalindromic Sequence in Xq27.1. OJGen. 2014;04:415–25.

4. DeStefano GM, Fantauzzo KA, Petukhova L, Kurban M, Tadin-Strapps M, Levy B, et al. Position effect on FGF13 associated with X-linked congenital generalized hypertrichosis. Proc Natl Acad Sci U S A. 2013;110:7790–5.

5. Zhu H, Shang D, Sun M, Choi S, Liu Q, Hao J, et al. X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. Am J Hum Genet. 2011;88:819–26.

6. Attanasio C, Nord AS, Zhu Y, Blow MJ, Li Z, Liberton DK, et al. Fine tuning of craniofacial morphology by distant-acting enhancers. Science. 2013;342:1241006.

7. Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, et al. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell. 2015;163:68–83.

8. Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. Cell Rep. 2018;23:1581–97.

9. Benko S, Fantes JA, Amiel J, Kleinjan D-J, Thomas S, Ramsay J, et al. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet. 2009;41:359–64.

10. Jakobsen LP, Ullmann R, Christensen SB, Jensen KE, Mølsted K, Henriksen KF, et al. Pierre Robin sequence may be caused by dysregulation of SOX9 and KCNJ2. J Med Genet. 2007;44:381–6.

11. Leipoldt M, Erdel M, Bien-Willner GA, Smyk M, Theurl M, Yatsenko SA, et al. Two novel translocation breakpoints upstream of SOX9 define borders of the proximal and distal breakpoint cluster region in campomelic dysplasia. Clin Genet. 2007;71:67–75.

12. Fonseca ACS, Bonaldi A, Bertola DR, Kim CA, Otto PA, Vianna-Morgante AM. The clinical impact of chromosomal rearrangements with breakpoints upstream of the SOX9gene: two novel de novo balanced translocations associated with acampomelic campomelic dysplasia. BMC Med Genet [Internet]. 2013;14. Available from: http://dx.doi.org/10.1186/1471-2350-14-50

13. Velagaleti GVN, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, et al. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. Am J Hum Genet. 2005;76:652–62.

14. Refai O, Friedman A, Terry L, Jewett T, Pearlman A, Perle MA, et al. De novo 12;17 translocation upstream of SOX9 resulting in 46,XX testicular disorder of sex development. Am J Med Genet A. 2010;152A:422–6.

15. Hill-Harfe KL, Kaplan L, Stalker HJ, Zori RT, Pop R, Scherer G, et al. Fine mapping of chromosome 17 translocation breakpoints > or = 900 Kb upstream of SOX9 in acampomelic campomelic dysplasia and a mild, familial skeletal dysplasia. Am J Hum Genet. 2005;76:663–71.

16. Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat Genet. 2017;49:36–45.

17. Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013;154:185–96.

18. Monti R, Barozzi I, Osterwalder M, Lee E, Kato M, Garvin TH, et al. Limb-Enhancer Genie: An accessible resource of accurate enhancer predictions in the developing limb. PLoS Comput Biol. 2017;13:e1005720.