The Genomic Landscape of Balanced Cytogenetic Abnormalities Associated with Human Congenital Anomalies

Claire Redin[1,2,3], Harrison Brand[1,2,3], Ryan L. Collins[1,3,4], Tammy Kammin[5], Elyse Mitchell[6], Jennelle C. Hodge[6,7], Carrie Hanscom[1], Vamsee Pillalamarri[1], Catarina M. Seabra[1,8], Mary-Alice Abbott[9], Omar A. Abdul-Rahman[10], Erika Aberg[11], Rhett Adley[1], Sofia L. Alcaraz-Estrada[12], Fowzan S. Alkuraya[13], Yu An[1,14], Mary-Anne Anderson[15], Caroline Antolik[1], Kwame Anyane-Yeboa[16], Joan F. Atkin[17,18], Tina Bartell[19], Jonathan A. Bernstein[20], Elizabeth Beyer[21], Ernie M.H.F. Bongers[22], Eva H. Brilstra[23], Chester W. Brown[24,25], Hennie T. Brüggenwirth[26], Bert Callewaert[27], Ken Corning[28], Helen Cox[29], Edwin Cuppen[23], Benjamin B. Currall[1,5,30], Tom Cushing[31], Dezso David[32], Matthew A. Deardorff[33,34], Annelies Dheedene[27], Marc D'hooghe[35], Bert B.A. de Vries[22], Dawn L. Earl[36], Heather L. Ferguson[5], Heather Fisher[37], David R. FitzPatrick[38], Pamela Gerrol[5], Daniela Giachino[39], Joseph T. Glessner[1,2,3], Troy Gliem[6], Margo Grady[40], Brett H. Graham[41,42], Cristin Griffis[21], Karen W. Gripp[43], Andrea L. Gropman[44], Andrea Hanson-Kahn[45], David J. Harris[46,47], Mark A. Hayden[5], Ron Hochstenbach[23], Jodi D. Hoffman[48], Robert J. Hopkin[49,50], Monika W. Hubshman[51], A. Micheil Innes[52], Mira Irons[53], Melita Irving[54,55], Sandra Janssens[27], Tamison Jewett[56], John P. Johnson[57], Marjolijn C. Jongmans[22], Stephen G. Kahler[58], David A. Koolen[22], Jerome Korzelius[23], Peter M. Kroisel[59], Yves Lacassie[60], William Lawless[1], Emmanuelle Lemyre[61], Kathleen Leppig[62,63], Alex V. Levin[64], Haibo Li[65], Hong Li[65], Eric C. Liao[66,67,68], Cynthia Lim[69,70], Edward J. Lose[71], Diane Lucente[1], Michael J. Macera[72], Poornima Manavalan[1], Giorgia Mandrile[39], Carlo L. Marcelis[22], Lauren Margolin[73], Tamaron Mason[73], Diane Masser-Frye[74], Michael W. McClellan[75], Cinthya Zepeda Mendoza[5,76], Björn Menten[27], Sjors Middelkamp[23], Liya R. Mikami[77,78], Emily Moe[21], Shehla Mohammed[54], Tarja Mononen[79], Megan E. Mortenson[55,80], Graciela Moya[81], Aggie Nieuwint[82], Zehra Ordulu[5,76], Sandhya Parkash[83], Susan P. Pauker[76,84], Shahrin Pereira[5], Danielle Perrin[73], Katy Phelan[85], Raul E. Piña Aguilar[12,86], Pino J. Poddighe[82], Giulia Pregno[39], Salmo Raskin[77], Linda Reis[87], William Rhead[88], Debra Rita[89], Ivo Renkens[23], Filip Roelens[90], Jayla Ruliera[15], Patrick Rump[91], Samantha L.P. Schilit[30], Ranad Shaheen[13], Rebecca Sparkes[52], Erica Spiegel[16], Blair Stevens[92], Matthew R. Stone[1], Julia Tagoe[93], Joseph V. Thakuria[76,94], Bregje W. van Bon[22], Jiddeke van de Kamp[82], Ineke van Der Burgt[22], Ton van Essen[91], Conny M. van Ravenswaaij-Arts[91], Markus J. van Roosmalen[23], Sarah Vergult[27], Catharina M.L. Volker-Touw[23], Dorothy P. Warburton[95], Matthew J. Waterman[1,96], Susan Wiley[97,98], Anna Wilson[1], Maria de la Concepcion A. Yerena-de Vega[99], Roberto T. Zori[100], Brynn Levy[101], Han G. Brunner[22,102], Nicole de Leeuw[22], Wigard P. Kloosterman[23], Erik C. Thorland[6], Cynthia C. Morton[3,5,76,103,104], James F. Gusella[1,3,30], Michael E. Talkowski[1,2,3,*]

[1]Molecular Neurogenetics Unit and Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA;
[2]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA;
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA;
[4]Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, and Departments of Neurology and Genetics, Harvard Medical School, Boston, MA 02115, USA
[5]Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital, Boston, MA 02115, USA;
[6]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55902, USA;
[7]Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA;
[8]GABBA Program, University of Porto, Porto, Portugal;
[9]Medical Genetics, Baystate Medical Genetics, and Baystate Children's Subspecialty Center, Springfield, MA 01199, USA;
[10]Department of Pediatrics, University of Mississippi Medical Center, Jackson, MS, USA;

[11]Maritime Medical Genetics Service, IWK Health Centre, Halifax, Nova Scotia, Canada
[12]Medical Genomics Division, Centro Medico Nacional 20 de Noviembre, ISSSTE, Mexico City, Mexico
[13]King Faisal Specialist Hospital and Research Center, MBC-03 PO BOX 3354, Riyadh 11211, Saudi Arabia;
[14]Institutes of Biomedical Sciences and MOE Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai, China
[15]Center for Human Genetic Research DNA and Tissue Culture Resource, Boston, MA 02114, USA;
[16]Columbia University Medical Center, New York, NY 10032, USA;
[17]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH 43210, USA;
[18]Division of Molecular and Human Genetics, Nationwide Children's Hospital, Columbus, OH 43205, USA
[19]Sacramento Medical Center, Department of Genetics, Sacramento, CA 95815, USA;
[20]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA
[21]Children's Hospital of Wisconsin and Department of Pediatrics, Medical College of Wisconsin, City, ST ZIP, USA;
[22]Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen 6500 HB, the Netherlands;
[23]Department of Genetics, Division of Biomedical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, 3508 AB Utrecht, The Netherlands;
[24]Department of Molecular and Human Genetics, Department of Pediatrics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA;
[25]Texas Children's Hospital, 6621 Fannin, Houston, TX 77030, USA;
[26]Department of Clinical Genetics, Erasmus University Medical Centre, 3000 CA Rotterdam, The Netherlands;
[27]Center for Medical Genetics, Ghent University, De Pintelaan 185, 9000 Ghent, Belgium;
[28]Greenwood Genetic Center, Columbia, SC, 29201, USA;
[29]West Midlands Regional Clinical Genetics Unit, Birmingham Women's Hospital, Edgbaston, Birmingham B15 2TG, England, UK;
[30]Department of Genetics, Harvard Medical School, Boston, MA, USA;
[31]University of New Mexico, School of medicine, Department of pediatrics, Division of pediatric genetics, Albuquerque, NM 87131, USA;
[32]Department of Human Genetics, Organization National Institute of Health Dr Ricardo Jorge, Lisbon, Portugal
[33]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA;
[34]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA;
[35]Algemeen Ziekenhuis Sint-Jan, Brugge, Belgium;
[36]Seattle Children's, Seattle, Washington, WA 98105, USA;
[37]Mount Sinai West Hospital, New York, NY 10019, USA;
[38]Medical Research Council Human Genetics Unit, Institute of Genetic and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK;
[39]Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Italy
[40]UW Cancer Center at ProHealth Care, Waukesha, Wisconsin, USA
[41]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA;
[42]Department of Genetics, Texas Children's Hospital, Houston, TX 77054, USA;
[43]Sidney Kimmel Medical School at T. Jefferson University, Philadelphia, PA, USA;
[44]Children's National Medical Center, N.W. Washington, D.C, USA
[45]Departments of Pediatrics and Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA;
[46]Division of Genetics, Boston Children's Hospital, Boston, MA, USA;
[47]Department of Pediatrics, Harvard Medical School, Boston, MA, USA;
[48]Department of Pediatrics, Division of Genetics, Boston Medical Center, MA, USA;
[49]Cincinnati Children's Hospital Medical Center, Cincinnati, USA;
[50]Department of Pediatrics, University of Cincinnati College Medicine, Cincinnati, USA;
[51]Schneider Medical Centre, Genetics, Israel;
[52]Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[53]Academic Affairs, American Board of Medical Specialties, Chicago, IL 60654, USA;
[54]Department of Clinical Genetics, Guy's and St Thomas' NHS Foundation Trust, London, UK;
[55]Division of medical and Molecular Genetics, King's College London, UK;
[56]Wake Forest Baptist Medical Center, Winston Salem, NC 27157, USA;
[57]Shodair Children's Hospital, Molecular Genetics Department, Helena, MT, USA;
[58]Division of Genetics and Metabolism, Arkansas Children's Hospital, AR, USA;
[59]Institute of Human Genetics, Medical University of Graz, Graz, Austria;

[60]Department of Pediatrics LSUHSC and Children's Hospital, New Orleans, LA, USA;
[61]CHU Sainte-Justine, 3175 chemin de la Côte-Sainte-Catherine, Montréal QC, Canada;
[62]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, USA;
[63]Clinical Genetics, Group Health Cooperative, Seattle, Washington, USA;
[64]Wills Eye Hospital, Ste. 1210, 840 Walnut Street, Philadelphia, PA, USA;
[65]Center for Reproduction and Genetics, The affiliated Suzhou Hospital of Nanjing Medical University, Suzhou, Jiangsu, China;
[66]Center for Regenerative Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA;
[67]Division of Plastic and Reconstructive Surgery, Massachusetts General Hospital, Boston, MA 02114, USA;
[68]Harvard Stem Cell Institute, Cambridge, MA 02138
[69]HonorHealth/Virginia G. Piper Cancer Center, Scottsdale, AZ 85258, USA;
[70]Arkansas Children's Hospital, Little Rock, AR 72202, USA;
[71]Department of Medical Genetics, University of Alabama Hospital at Birmingham, Birmingham, AL, USA;
[72]New York-Presbyterian Hospital, Columbia University Medical Center, New York, USA;
[73]Program in Medical and Population Genetics and Genomics Platform, Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA;
[74]Department of Genetics, Rady Children's Hospital San Diego, CA, USA;
[75] Department of Obstetrics & Gynaecology, Madigan Army Medical Center, Tacoma, WA 98431, USA;
[76]Harvard Medical School, Boston, MA, USA;
[77]Group for Advanced Molecular Investigation, Graduate Program in Health Sciences, School of Medicine, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil;
[78]Centro Universitário Autônomo do Brasil (Unibrasil), Curitiba, Paraná,Brazil;
[79]Department of Clinical Genetics, Kuopio University Hospital, Finland;
[80]Novant Health Derrick L. Davis Cancer Center, Winston Salem, NC 27103, USA;
[81]GENOS Laboratory, Buenos Aires, Argentina;
[82]Department of Clinical Genetics, VU University Medical Center, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands;
[83]Department of Pediatrics, Maritime Medical Genetics Service, IWK Health Centre, Dalhousie University, Halifax, Nova Scotia, Canada;
[84]Medical Genetics, Harvard Vanguard Medical Associates, Watertown, MA 02472, USA;
[85]Hayward Genetics Program, Department of Pediatrics, Tulane University School of Medicine, New Orleans, LA, USA;
[86]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, United Kingdom
[87]Department of Pediatrics and Children's Research Institute, Medical College of Wisconsin, Milwaukee, WI, 53226, USA;
[88]Children's Hospital of Wisconsin and Departments of Pediatrics and Pathology, Medical College of Wisconsin, USA;
[89]Midwest Diagnostic Pathology, Aurora Clinical Labs, Rosemont, IL, USA;
[90]Algemeen Ziekenhuis Delta, Roeselare, Belgium;
[91]University of Groningen, University Medical Center Groningen, Department of Genetics, PO Box 30.001, 9700RB Groningen, The Netherlands;
[92]McGovern Medical School at The University of Texas Health Science Center at Houston, TX, USA;
[93]Genetic Services, Alberta Health Services, Alberta T1J 4L5, Canada;
[94]Division of Medical Genetics, Massachusetts General Hospital, Boston, MA 02114;
[95]Department of Clinical Genetics and Development, Columbia University Medical Center, New York, NY 10032, USA;
[96]Eastern Nazarene College, Department of Biology, Quincy, MA 02170, USA;
[97]Cincinnati Children's Hospital Medical Center, Division of Developmental Pediatrics at University of Cincinnati, OH, USA;
[98]Cincinnati Children's Hospital Medical Center, Division of Behavioral Pediatrics at University of Cincinnati, OH, USA;
[99]Laboratory of Genetics, Centro Medico Nacional 20 de Noviembre, ISSSTE, Mexico City, Mexico.
[100]Division of Pediatric Genetics & Metabolism, University of Florida, Florida, USA;
[101]Department of Pathology, Columbia University, New York, NY, USA;
[102]Department of Clinical Genetics, Maastricht University Medical Centre, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands;
[103]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA;
[104]Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Manchester Academic Health Science Center, Manchester, UK;

***Correspondence:**
Michael E. Talkowski, Ph.D.
Associate Professor
Center for Human Genetic Research
Massachusetts General Hospital, Harvard Medical School, Broad Institute
185 Cambridge St., Boston, MA 02114
talkowski@chgr.mgh.harvard.edu; talkowsk@broadinstitute.org

ABSTRACT

Despite their clinical significance, characterization of balanced chromosomal abnormalities (BCAs) has largely been restricted to cytogenetic resolution. We explored the landscape of BCAs at nucleotide resolution in 273 subjects with a spectrum of congenital anomalies. Whole-genome sequencing revised 93% of karyotypes and revealed complexity that was cryptic to karyotyping in 21% of BCAs, highlighting the limitations of conventional cytogenetic approaches. At least 33.9% of BCAs resulted in gene disruption that likely contributed to the developmental phenotype, 5.2% were associated with pathogenic genomic imbalances, and 7.3% disrupted topologically associated domains (TADs) encompassing known syndromic loci. Remarkably, 8 subjects harbored BCA breakpoints that localized to a single TAD encompassing *MEF2C*, a known driver of the 5q14.3 microdeletion syndrome, resulting in altered *MEF2C* expression by genomic rewiring. This study proposes that sequence-level resolution dramatically improves prediction of clinical outcomes for balanced rearrangements and provides insight into novel pathogenic mechanisms such as altered regulation due to changes in chromosome topology.

Keywords: Cytogenetics, structural variation, balanced chromosomal abnormality, congenital anomaly, intellectual disability, autism, translocation, inversion, chromothripsis, topologically associated domain (TAD), *MEF2C*

Balanced chromosomal abnormalities (BCA) are a class of structural variation that involve rearrangement of chromosome structure and result in a change in the orientation or localization of a genomic segment without a large concomitant gain or loss of DNA. This class of variation includes inversions, translocations, excisions/insertions, and more complex rearrangements consisting of combinations of such events. Cytogenetic studies of unselected newborns and control adult males estimate a prevalence of 0.2-0.5% for BCAs in the general population[1-3]. By contrast, an approximate five-fold increase in the prevalence of BCAs detected by karyotyping has been reported among subjects with neurodevelopmental disorders, particularly intellectual disability (1.5%)[4] and autism spectrum disorder (ASD; 1.3%)[5]. These data suggest that BCAs represent highly penetrant mutations in a meaningful fraction of subjects with associated congenital anomalies or neurodevelopmental disorders.

Delineating the breakpoints of BCAs and the genomic regions that they disrupt has long been a fertile area of novel gene discovery in human genetic research and has greatly contributed to the annotation of the morbid map of the human genome[6-8]. Despite their significance in human disease, the clinical detection of this unique class of chromosomal rearrangements still relies upon conventional cytogenetic methods such as karyotyping that are limited to microscopic resolution (~3-10 Mb, depending on the chromosome banding pattern and specimen type)[9]. The absence of gross genomic imbalances renders BCAs invisible to higher resolution techniques that currently serve as first-tier diagnostic screens for many developmental anomalies of unknown etiology: chromosomal microarray (CMA), which can detect microscopic and sub-microscopic copy-number variants (CNVs), or whole-exome sequencing (WES), which surveys single nucleotide variants within coding regions. Without access to precise breakpoint localization, clinical interpretation of *de novo* BCAs has been limited to estimates of an untoward outcome from population cytogenetic studies based solely on the presence of a rearrangement (6.1% of *de novo* reciprocal translocations, 9.4% for *de novo* inversions)[10]. We have recently shown that innovations in genomic technologies can efficiently reveal BCA breakpoints at nucleotide resolution with a cost and timeframe comparable to clinical CMA or karyotyping; however, only a limited number of BCAs has been evaluated to date[11,12,7,13-16].

In this study, we explored several fundamental but previously intractable questions regarding *de novo* BCAs associated with human developmental anomalies, such as the origins of their formation, the genomic properties of the sequences that they disrupt, and the mechanisms by which BCAs act as dominant pathogenic mutations. We evaluated 273 subjects ascertained based upon the presence of a

BCA discovered by karyotyping in a cytogenetics laboratory in a proband that presented with a developmental anomaly. We defined the genomic sequences that were altered by the breakpoints and created a framework in which we interpreted their significance based on convergent genomic datasets. This included CNV and WES data in tens of thousands of individuals, as well as prediction of long-range regulatory effects from recent studies that have established high-resolution maps of chromosomal compartmentalization in the nucleus[17,18]. Our findings indicate that formation of BCAs involves a variety of mechanisms and sequence characteristics, that the end-result often reflects substantial complexity invisible to cytogenetic assessment, that BCAs directly disrupt genes likely to contribute to early developmental abnormalities in at least one-third of subjects, and that BCAs can cause long-range regulatory changes due to alterations to the chromosome structure. These results highlight the myriad genomic features of BCAs that have been largely unexplored in conventional cytogenetic research and demonstrate mechanisms by which they contribute to abnormalities in human development.

RESULTS

Sequencing BCAs reveals cryptic complexity

In this study 273 subjects were sequenced originating from five primary referral sites that collectively represented an international consortium of over 100 clinical investigators. Subjects harbored a BCA that was detected by karyotyping and presented with varied developmental anomalies. Most of the 273 subjects were surveyed using large-insert whole-genome sequencing (liWGS or 'jumping libraries'; 83%), with the remainder of subjects being analyzed by standard short-insert WGS or targeted breakpoint sequencing (see **Online Methods; Supplementary Table 1**). Subjects were preferentially selected with confirmed *de novo* BCAs based on cytogenetic studies at the referring site or with rearrangements that segregated with a phenotypic anomaly within a family (72.5% of subjects); however, inheritance information was unavailable for one or both parents in the remaining 27.5% of subjects. Notably, subjects harboring BCAs that were inherited from an unaffected parent were excluded from this study. Of interest, 62.6% of subjects received clinical CMA screening prior to enrollment to confirm the absence of a pathogenic CNV (**Table 1**). Subjects presented with a spectrum of clinical features: congenital anomalies ranged from organ-specific disorders to multisystem abnormalities, as well as neurodevelopmental conditions such as intellectual disability or autism spectrum disorder (ASD; **Table 1**). While no specific phenotypes were prioritized for inclusion (see **Supplementary Fig. 1**), neurological defects were the most common feature in the cohort (80.2% of subjects when using digitalized phenotypes from Human Phenome Ontology [HPO][19]; **Table 1**; **Supplementary Table 2**).

Breakpoints were identified in 248 of the 273 cases (90.8%); all subsequent analyses were restricted to these 248 subjects. This success rate was consistent with expectations, as simulation of one million random breakpoints in the genome and comparison against all uniquely alignable 10 bp – 100 bp kmers suggests that 7.6% of simulated breakpoints were localized within N-masked regions or genomic segments that cannot be confidently mapped by short-read sequencing (**Supplementary Fig. 2**). Sequencing identified 876 breakpoints genome-wide (**Fig. 1a**) and revised the breakpoint localization by at least one sub-band in 93% of subjects when compared to the karyotype interpretation (breakpoint positions provided in **Supplementary Table 3**). Across all rearrangements, 26% (n=65) of BCAs were found to be complex (*i.e.,* involved three or more breakpoints), including 5.3% (n=13) that were consistent with the phenomena of chromothripsis or chromoplexy that we and others have previously defined in cancer genomes and the human germline (complex reorganization of the chromosomes involving extensive shattering and random ligation of fragments from one or more chromosomes)[20-24].

The most complex BCA involved 57 breakpoints (**Supplementary Fig. 3**). When analyses were restricted to the 230 subjects for which the karyotype suggested a simple chromosomal exchange, 48 (21%) were determined to be rearrangements with complexity that was cryptic to the karyotype, emphasizing the insights that are gained from nucleotide resolution. Across all BCAs, 80.7% resolved to less than ten kilobases of total genomic imbalance, although several cases harbored large cryptic imbalances (mostly deletions) of varied impact (**Fig. 1b, Supplementary Table 4**). Importantly, 9.3% of BCAs displayed an overall genomic imbalance greater than 1 Mb and only 12.2% had imbalances of >100 kb in this study, representing a significantly lower fraction than previous cytogenetic estimates[25]. The overall genomic imbalance associated with a BCA was larger among cases without CMA pre-screening, and 15.5%/18.8% of these subjects harbored imbalances greater than 1 Mb/100 kb, respectively (**Fig. 1b, Supplementary Table 4**). The total genomic imbalance generally increased with the number of breakpoints, though there were chromothripsis and chromoplexy events that were essentially balanced (*e.g.,* subject NIJ19 involved 13 junctions across five chromosomes that resolved to a final genomic imbalance of only 631 bases).

BCA formation is mediated by multiple molecular mechanisms

Extensive mechanistic studies have been performed on breakpoints of large CNV datasets; however, the limited scale and resolution of BCA studies have precluded similar analyses for balanced rearrangements. Using precise junction sequences from 662 breakpoints, we found that nearly half displayed signatures of blunt-end ligation (45%), presumably driven by non-homologous end joining (NHEJ) (**Fig. 1c**). A substantial fraction (29%) involved microhomology of 2-15 bp at the breakpoint junction (median: 3-bp microhomology), indicating that template-switching coupled to DNA-replication mechanisms such as microhomology-mediated break-induced replication (MMBIR) contribute to a substantial fraction of BCAs[26]. A comparable fraction (25%) of junctions harbored micro-insertions of several basepairs (1 to 375 inserted bases, median: 6-bp), consistent with NHEJ or fork stalling and template switching (FoSTeS) mechanisms (**Fig. 1c**). Finally, only nine junctions (1%) contained long stretches of homologous sequences (>100 bp) that would be consistent with homology-mediated repair. It is important, however, to note that this is almost certainly an underestimate given the limitations of short-read sequencing to capture rearrangements localized within highly homologous sequences such as segmental duplications or microsatellites. BCA breakpoint signatures from this study were also compared to 8,943 deletion breakpoints identified in 1,092 samples from the 1000 Genomes Project[27]. BCA breakpoints were enriched for blunt-end signatures while depleted for microhomology and large

homology sequences compared to deletion breakpoints (**Supplementary Fig. 4**), suggesting that they arose from distinct mechanisms.

Comparison of the observed breakpoints to 100,000 independent sets of simulated breakpoints that retained the properties of the observed dataset (see **Online Methods**) established nominal enrichment for repeat elements (P=0.015) and fragile sites (P=0.043), while no significant enrichment for the other genomic features tested (recombination hotspots, DNAse-I hypersensitive sites, or transcription factor binding sites; **Supplementary Fig. 5**). Incorporating Hi-C interaction data to explore the association between nuclear organization of the chromosomes and BCA formation revealed that pairs of loci comprising a BCA breakpoint did not stem from regions with significantly higher contact patterns in the nucleus[18]; however, pairs of BCA breakpoint loci displayed genome-wide interaction patterns that were marginally more correlated than random pairings (P=0.046; see **Supplementary Methods** and **Supplementary Fig. 6**). These results suggest that DNA fragments involved in BCA formation are more likely to be co-localized in the same or neighboring subcompartments prior to chromosomal reassembly, though at the sample sizes available they did not necessarily harbor increased direct interactions.

BCA breakpoints associated with congenital anomalies are enriched for functionally relevant loci

While protein-coding sequences represent less than 2% of the human genome, the total genic space in which a structural variation can disrupt a transcript is considerable as the cumulative coverage of transcribed regions is over 60% based on recent annotations[28]. Consistent with this expectation, 67% (589/876) of all breakpoints in this study disrupted a gene, and at least one gene was truncated in most BCAs (75%, 186/248), which did not deviate from random expectations (observed n=408 RefSeq genes, expected n=392±20, P=0.220; **Supplementary Fig. 7**). While BCA breakpoints were not enriched for gene disruptions beyond expectations, the properties of the disrupted genes deviated significantly from randomly simulated breakpoints for several key features, as described below, suggesting that the pathogenic impact of BCAs in this cohort ascertained based upon the presence of a developmental abnormality is not a consequence of their likelihood to disrupt genes but rather a reflection of the gene(s) that they alter (the list of all disrupted genes at breakpoints is provided in **Supplementary Table 5**).

We observed a significant enrichment for disruption of genes highly intolerant to truncating mutations, as defined by two independent groups (P=0.027 and P=0.001 for Residual Variation Intolerance Score

[RVIS] and probability of loss-of-function [LoF] intolerance [pLI] scores, respectively; **Fig. 2a**)[29,30]. Embryonically expressed genes (P=0.001)[31] and genes previously associated with autosomal dominant disorders (P=0.002) were also more likely to be disrupted than expected by chance, whereas no enrichment was observed for genes associated with autosomal recessive disorders (P=0.294; **Fig. 2a**)[32]. The strongest enrichment at breakpoints was detected for genes previously associated with developmental disorders (≥2 *de novo* LoF mutations [dnLoF]) as amalgamated from multiple independent datasets (P=2x10^{-5}; **Supplementary Table 6**). Significant enrichment was also observed at breakpoints for FMRP-target genes and chromatin remodeling genes[33,34], replicating the enrichment observed for genes with dnLoF in subjects with neurodevelopmental disorders (**Fig. 2b**)[35-37,7,38,31]. No enrichments were observed for *CHD8* target genes[39,40]. When further incorporating expression data of the developing brain from BrainSpan[41], truncated genes showed higher expression patterns during early developmental stages than randomly simulated datasets (**Supplementary Fig. 8**). These results appeared to be specific to early developmental anomalies; there was no significant enrichment of genes associated with schizophrenia[42,43], or gene-sets associated with complex disorders that were considered as negative controls such as type-II diabetes, cancer, or height. Given the distribution of clinical phenotypes in the cohort, we hypothesized that enrichment signals were driven by the predominance of neurological abnormalities among the subjects. We therefore performed analyses that segregated subjects with or without nervous system related conditions using HPO-reported phenotypes, and replicated most associations for the subset of cases with neurological conditions while enrichments were not significant for the smaller subset of subjects without reported nervous system abnormalities (**Supplementary Fig. 9**).

BCAs predominantly contribute to developmental anomalies by direct gene truncation

We next asked the fundamental question: "How often does a BCA represent a likely pathogenic mutation that contributes to the subject's abnormal developmental phenotype?" We sought to interpret the clinical significance of each BCA with reference to the phenotype reported in the proband and the genomic region(s) altered by the rearrangement. We built our interpretation using categories comparable to those established by ClinVar and the Deciphering Developmental Disorders consortium (DDD)[44]; however, we restricted interpretation of variants of potential clinical relevance to *Pathogenic* or *Likely Pathogenic,* as detailed below and in **Supplementary Table 7**. All other variants were interpreted as *Variant of Unknown Significance* (VUS), as we lacked sufficient clinical and functional data to interpret

variants as *Benign* or *Likely Benign*. The overall summary of the predicted impact for each mapped BCA is provided in **Supplementary Table 8**.

Pathogenic: We compared loci disrupted by BCAs to genes that had been robustly associated with dominant developmental disorders (≥3 reported cases with dnLoF in OMIM, DDD, and amalgamated large-scale sequencing studies in neurodevelopmental disorders; see **Supplementary Methods** and **Supplementary Table 6**). In total, 66 subjects (26.6%) harbored *Pathogenic* BCAs that disrupted these previously defined developmental loci either through direct gene disruption or genomic imbalance (**Fig. 2c**; **Table 2**; **Supplementary Table 9**). In the majority of these subjects (53/66), the rearrangement truncated a high confidence syndromic locus. These included known drivers of recurrent microdeletion syndromes (e.g., *SATB2, MBD5, EHMT1, NFIA, ZBTB20*)[45-49], loci associated with imprinted disorders (*SNURF-SNRPN*), and genes well-established as highly penetrant loci in developmental disorders (*e.g., CHD7, CHD8, CDKL5, CUL3*, *DYRK1A*, *GRIN2B*), as well as more recently implicated genes such as *AHDC1*, *CTNND2* and *WAC* (**Fig. 2c**; **Table 2**; **Supplementary Table 9**). Several genes were disrupted in two or more subjects in this cohort, further confirming their significant role in developmental anomalies: *AUTS2, KDM6A, MBD5, MYO6, MYT1L, PHF21A, PHIP, SNURF-SNRPN, SOX5* and *ZBTB20*. Importantly, ten subjects harbored BCAs that disrupted genes associated with dominant disorders for which the expected phenotype such as cardiovascular defects, childhood or late-onset hearing loss, neurodegenerative disorder, were not observed in the proband (**Supplementary Table 9**); in those subjects the rearrangements were likely incidental findings, but could alternatively represent pleiotropy in which disruption of the same locus can manifest in multiple distinct phenotypes. In the remaining 13 subjects with *Pathogenic* BCAs (13/66), genomic imbalances at the breakpoints either overlapped with known microdeletion/microduplication syndromes, or encompassed a gene associated with a dominant developmental disorder (*e.g.,* 12p12.1-p11.22 deletion encompassing *SOX5*; **Table 2**; **Fig. 2c**).

Likely Pathogenic: Each specific rearrangement effectively represents a private event, or an N-of-1, which is a major challenge for interpretation in genomic studies. To interpret variants as *Likely Pathogenic* when the BCA did not disrupt established developmental loci, we relied on convergent genomic evidence from large-scale datasets. The premise was that candidate genes associated with congenital anomalies or early developmental defects would show evidence of reduced reproductive fitness and intolerance to haploinsufficiency. Thirty-one subjects harbored BCAs that were considered

as *Likely Pathogenic* via direct loci disruption (**Table 2**; **Supplementary Tables 8, 10**). In 25 subjects, the rearrangement directly disrupted a gene highly intolerant to dnLoF, and in which dnLoF mutations had been previously reported in isolated cases (1 or 2 subjects, with an additional subject now represented by the BCA in our study; e.g. *CACNA2D3, ROBO2, NFIB*), some of which had strong biological support for involvement in developmental anomalies (*EP400, STXBP5, NRXN3*). Among those proposed candidate genes, several were disrupted in multiple subjects from the cohort (*NPAS3*$_{(x4)}$, *PTPRZ1*$_{(x3)}$, *SYNCRIP*$_{(x2)}$; **Table 2**, **Supplementary Tables 10-11**). Two subjects had BCAs likely associated with genomic disorders: one BCA involving a 2p21-p13.3 duplication encompassing *NRXN1,* one BCA disrupting the imprinted 11p15 region likely associated with Silver-Russel syndrome (MIM#180860). In the remaining four subjects with *Likely Pathogenic* BCAs, the rearrangement truncated genes that were established to be associated with developmental disorders yet in which only activating or missense mutations had been previously reported (*e.g., CACNA1C* and *GNB1)*[50,51], proposing a dosage sensitive model for these loci, comparable to recurrent genomic disorders. Based on these results, we interpreted that 12.5% (31/248) of subjects harbored a BCA that was likely contributing to the phenotype through the involvement of potentially novel candidate genes or disease mechanisms.

Collectively, these data suggest that 39.1% (97/248) of subjects have a phenotype that can be at least partially explained by haploinsufficiency or dosage alteration of an individual gene or locus (**Fig. 2c**; **Supplementary Tables 8-10**). Importantly, the overall diagnostic yield was significantly higher in subsets of the group, such as among those subjects who harbored confirmed *de novo* or co-segregating BCAs compared to subjects for whom inheritance was unknown (**Fig. 2d**), or among subjects who had not been screened clinically by CMA prior to enrollment (**Fig. 2e)**. Despite these substantial yields, the marked increase in the frequency of BCAs associated with birth defects compared to the general population still suggests that alternative mutational mechanisms other than gene disruption may account for the developmental defects in a fraction of subjects for which the BCAs were interpreted as VUS. We explored such potential mechanisms in this unique dataset.

Positional effects via disruption of long-range regulatory interactions

Clusters of BCA breakpoints within intergenic regions may suggest disruption of strong regulatory elements that contribute to disease manifestation via positional effects. Alternatively, this could reflect recurrent rearrangements due to fragile sites and/or recombination hotspots. To isolate genomic regions

in which an unusual number of BCA breakpoints were localized, we partitioned the genome into 1 Mb bins using a sliding window of 100 kb. Only one genomic segment, consisting of several contiguous genomic bins, achieved genome-wide significance (P=8×10^{-9}; **Fig. 3a**). This segment localized to cytogenetic band 5q14.3 and involved breakpoints from a remarkable eight independent BCAs. Of these eight BCAs, one directly disrupted *MEF2C* at 5q14.3 while the other seven BCAs mapped to intergenic regions proximal and distal to *MEF2C*. Importantly, among the seven BCAs with intergenic disruption of 5q14.3, none included a breakpoint disrupting a locus of known significance elsewhere in the genome, suggesting that an alternative mechanism to direct gene disruption was operating in the 5q14.3 region.

All BCA breakpoints from the 5q14.3 cluster overlapped with the previously described 5q14.3 microdeletion syndrome for which nearly 100 subjects have been reported, with *MEF2C* as the proposed genetic driver of the syndromic phenotypes observed (**Fig. 3b**)[52-58]. However, deletions have been reported in cases located in proximity to *MEF2C* but not encompassing this gene (**Fig. 3b**), and the presence of seven BCAs distal to *MEF2C* in this study both challenge the hypothesis that direct disruption of *MEF2C* is a necessary and sufficient cause of the syndrome. When combining three previously described subjects with reported BCAs distal to *MEF2C* with the eight subjects from our cohort, a total of 11 subjects harbor balanced rearrangement breakpoints localized to the same 1 Mb region within 5q14.3 (**Fig. 3b**)[52,57,15]. Only one BCA directly truncated *MEF2C*, while all 10 remaining BCAs were predicted to disrupt a topologically associated domain (TAD) containing *MEF2C* (**Fig. 3b**). TADs are structured chromatin domains of increased interactions that typically define a local regulatory unit bridging regulatory elements together with their target genes[59]. Their disruption by genomic rearrangements can lead to impaired gene regulation and therefore disease pathogenesis[60,61]. Correspondingly, in the four subjects that harbored BCA breakpoints up to 860 kb distal to *MEF2C* for which RNA from lymphoblastoid cell lines (LCLs) was available, analysis of *MEF2C* expression revealed a significant reduction in all four subjects compared to 16 age-matched controls divided equally by sex (**Fig. 3d**). These analyses provide compelling evidence that alteration of the TAD architecture in this genomic disorder region can disrupt normal *MEF2C* expression. When integrated with existing data, the converging clinical features suggest multiple distinct mutational mechanisms resulting in phenocopies of the 5q14.3 microdeletion syndrome: (1) direct disruption of *MEF2C* via dnLoF mutations, (2) deletions including *MEF2C*, and (3) long-range positional effects from deletions and BCAs not impacting *MEF2C* via alteration to the physical orientation of the TAD structure (**Fig. 3c**).

Beyond 5q14.3, we also identified three other loci (2q33.1, 6q14.3 and 14q12) suggestive of an accumulation of BCA breakpoints, although these loci did not reach genome-wide significance (P=1x10^{-4}). Each segment contained BCA breakpoints from four independent subjects and overlapped with known microdeletion syndromes (**Fig. 3a**). At 2q33.1, one BCA disrupted *SATB2*, associated with Glass syndrome and recognized as the established driver of the 2q33.1 microdeletion syndrome[47,7], while the remaining three rearrangements were predicted to impact long-range interactions between *SATB2* and its regulatory elements, similar to the effect observed with *MEF2C* at 5q14.3 (**Supplementary Fig. 10**). In the 14q12 cluster, all BCA breakpoints were distal to *FOXG1*, which has been reported in atypical Rett syndrome[62-65]. The phenotypes associated with all four of these subjects were highly correlated based upon analyses of HPO reported terms (HPO-sim *P-value*=0.006; see **Methods** and **Supplementary Table 11**)[66,67], and were consistent with the multiple previous reports of subjects with dysregulation of *FOXG1* (**Supplementary Figure 11**)[62,68,63-65]. At 6q14.3, four BCAs were localized in proximity to *SYNCRIP*, a highly constrained gene in which dnLoF had been reported in two subjects with neurodevelopmental disorders[69]. In one subject the BCA directly disrupted *SYNCRIP*, while another subject harbored a breakpoint distal to *SYNCRIP* that was part of a cryptic 6q14.3 deletion encompassing the full gene, though the impact of the other two BCAs was unclear due to their localization to an adjacent contact domain (**Supplementary Fig. 12**). Finally, a systematic screen identified four additional subjects in which a TAD disruption could represent a positional effect on known syndromic loci associated with a developmental disorder that closely matched the subject's phenotype (*PITX2, SLC2A1, SOX9, SRCAP;* **Supplementary Fig. 13-15**). In two of these regions, LCLs were available from the corresponding subjects and expression of the proposed driver gene was significantly reduced when compared to a 16-sample control panel (*SLC2A1* and *SRCAP*, **Supplementary Fig. 13** and **14**).

Collectively, 7.3% of subjects harbored a BCA predicted to alter long-range regulatory interactions involving an established syndromic locus with comparable phenotype, recurrently involving *MEF2C*, *SATB2*, and *FOXG1* while an additional four subjects harbored a BCA that could represent long-range positional effects (two confirmed by expression studies), though our sample sizes precluded detection of a significant accumulation of breakpoints in these latter four regions. These data suggest that alterations to TAD structures likely represent a significant component of the deleterious impact of mutations associated with genomic rearrangements.

DISCUSSION

This characterization of BCAs at nucleotide resolution offers new insights into their mechanisms of formation, the properties connected to their rearrangement in the nucleus, and a substantial yield of potentially novel genes associated with human development. These results also emphasize that neither the mere presence of a BCA in a subject with developmental defects nor the number of genes it disrupts (if any) provide sufficient prognostic power, but rather that the properties of the specific genes and regions that are altered are the most informative in predicting resultant phenotypes. These data build upon recent studies on genome topology and provide further evidence that alterations to chromosome structure can lead to alternative, yet potentially predictable, pathogenic mechanisms by changing the long-range regulatory architecture of physical interactions and chromatin looping in the nucleus[66,60,61]. The yield of clinically meaningful results in this study, which ranged from 26.6% to 46.4% of the subjects evaluated, was substantial. Nonetheless, the relative enrichment from cytogenetic studies of BCAs in subjects with developmental abnormalities compared to controls suggests that there are yet additional alternative pathogenic mechanisms associated with *de novo* chromosomal rearrangements that remain to be discovered[4,5].

These data provide an initial vantage of the potential utility of emerging datasets that characterize the nuclear organization of the chromosomes. They propose novel pathogenic mechanisms by which BCAs may operate, which appear to be a consequence of the disruption of long-range interactions between regulatory elements and their target gene[66,60,61]. Structural variants can indeed easily scramble DNA topology and contact domains with potentially dramatic regulatory consequences. TADs cover a substantial fraction of the genome; therefore, the vast majority of structural variation will perturb one of those domains and cannot constitute a predictive criterion for pathogenicity *per se*. These data propose that the recurrent disruption of a TAD encompassing a high confidence locus beyond what is expected by chance, concomitant with strong phenotypic overlap between the carrier of the variant and haploinsufficiency of the locus in independent subjects, may be a first step towards highlighting putative positional effects, though definitive conclusions will still require functional validation. Expression studies in peripheral blood cells of individual subjects are a plausible, yet suboptimal, method for confirming positional effects. There is clearly a need for sensitive and specific tools to predict such positional effects caused by long-range regulatory perturbations, and to annotate further the morbid genome with more complete knowledge of these functional interactions. The fraction of BCAs in this

study that may be associated with this pathogenic mechanism is therefore just an entrée into their likely significance as a component of the unexplained genetic contribution to human birth defects.

In terms of evaluating diagnostic strategies, this study further highlights limitations of current diagnostic tools such as karyotyping or CMA in interpreting and detecting BCAs[11,13-16]. While the capability to visualize the chromosomes and detect *de novo* BCAs by traditional karyotyping represented a critical leap in genetic diagnostics, as exemplified by the seminal population cytogenetic studies performed by our late co-author, Dorothy Warburton[10], the detection of gross chromosomal abnormalities provides limited prognostic capability as to the clinical manifestation that may present in a given case. Our data demonstrate that karyotyping significantly underestimates complex rearrangements and is almost always revised by at least a sub-band. Karyotyping is also insensitive to genomic imbalances observed in the human germline that cannot be directly visualized (~5-10 Mb). By comparison, CMA is generally recommended as a first-tier diagnostic screen given its sensitivity to detect submicroscopic CNVs, yet it is blind to copy-neutral events such as those described herein. This study provides critical new insights into the fraction of BCAs that can be ascertained by CMA analyses. Compared to cytogenetic estimates suggesting that up to 40% of BCAs resolved as unbalanced rearrangements and could therefore be ascertained using CMA[25], whole-genome sequencing in this cohort suggests that, even at the resolution of 100 kb, only about 12% of BCAs involved a genomic imbalance. If we consider only the 102 subjects for whom no CMA was previously performed, this proportion increases to 18.8% at 100 kb resolution and 17.6% at 500 kb resolution, suggesting that 81.8-82.4% of BCAs in this study would be inaccessible to the resolution of most CMA platforms routinely used in clinical diagnostics. Notably, there is still benefit to an initial CMA screen, as is illustrated by the significantly lower yield of pathogenic BCAs among subjects who had been pre-screened by CMA (19-37%) compared to those who had not (41-64%; **Fig. 2e**), indicating that a fraction of pathogenic variation in these genomes was captured by the CMA prescreen either in relation to or independent of the BCA.

These data strongly argue for the implementation of technologies capable of detecting both balanced and unbalanced genomic rearrangements. This could be achieved by using a conventional cytogenetic test followed by a reflex WGS analysis when an abnormality is detected, which we have previously demonstrated can provide access to all classes of structural variation in the human genome while being accomplished in a relatively rapid timeframe[12,70]. Despite its great promise, it is important to recognize the limitations of massively parallel sequencing in routine cytogenetic practice. This study used large-

insert jumping libraries to maximize physical coverage and minimize cost per base of genome covered. Yet these analyses failed to reveal breakpoints in 9% of BCAs tested, and our simulations indicate that at large sample sizes, we would anticipate ~7-8% of breakpoints to be undetectable by short-read sequencing. At present, this result gives credence to maintaining the parallel visualization of structural changes in the genome using traditional cytogenetic approaches such as karyotyping for regions that are recalcitrant to massively parallel sequencing. As sequencing technologies and analytical capabilities improve, this component of the variant spectrum will become more tractable to genomic approaches, and the future implementation of long-read sequencing may revolutionize the capacity to survey currently inaccessible segments of the human genome[71,72].

In conclusion, these data indicate that *de novo* BCAs represent a highly penetrant mutational class in human disease, and that their delineation can provide prognostic insights not available at current cytogenetic resolution. Although encouraging, this yield does not explain all of the developmental anomalies in this cohort and suggests that additional pathogenic mechanisms await discovery. A meaningful fraction may be attributable to novel genes or regulatory alterations, but additional pathogenic mechanisms remain to be explored such as recessive modes of inheritance, gene fusions, disruption of imprinted regions, enhancer adoption[73,66], and more complex oligogenic models. Evaluation of extremely large cohorts will be required to resolve further such mechanisms, and characterization of BCAs in control populations would benefit annotation of the morbid human genome and interpretation of the biological and clinical consequences of its structural rearrangement.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGMENTS

We are infinitely grateful for the seminal work led by our co-author, Prof. Dorothy Warburton, who passed away during review of this manuscript. Dr. Warburton was a pioneer in cytogenetic research and a close colleague, mentor, and friend to so many in the cytogenetics community. We wish to thank all subjects and families who have been enrolled in this study, as well as the countless genetic counselors and clinical geneticists who contributed to the ascertainment of subjects. This study was supported by grants from the National Institutes of Health GM061354 (to M.E.T., J.F.G., C.C.M. and E.L.), the Harvard Medical School–Portugal Program in Translational and Clinical Research and Health Information to C.C.M and D.D., MH095867 and HD081256 to M.E.T., the March of Dimes 6-FY15-255 and the Desmond and Ann Heathwood MGH Research Scholars award to M.E.T, as well as fellowships from the EMBO (EMBO ALTF-183-2015), the Bettencourt-Schueller Foundation and the Philippe Foundation to C.R.

AUTHOR CONTRIBUTIONS

M.E.T, J.F.G, C.C.M, E.T., J.C.H., W.P.K., N.dL. and H.G.B designed the study. C.R., H.B., R.L.C., V.P., J.T.G., W.P.K., M.R.S. and M.J.vR. performed computational analyses. C.H., C.M.S., R.A., MA.A., C.A., E.C., B.B.C., J.K., W.L., P.M., L.M., T.M., D.P., J.R., M.J.W. and A.W. performed cellular, molecular or genomic experiments. T.K., E.M., J.C.H, MA.A, O.A.R., E.A., R.A., S.A.E., F.S.A, Y.A., K.A.Y., J.F.A., T.B., J.A.B., E.B., E.M.B., E.H.B, C.W.B., H.T.B., B.C., K.C., H.C., T.C., D.D., M.A.D., A.D., M.D., B.B.dV., D.L.E., H.L.F., H.F., D.R.F., P.G., D.G., T.G., M.G., B.H.G., C.G., K.W.G., A.L.G., A.H.K., D.J.H., M.A.H., R.H., J.D.H., R.J.H., M.W.H., A.M.I., M.I., M.I., S.J., T.J., J.P.J., M.C.J., S.G.K, D.A.K., P.M.K., Y.L., E.L., K.L., A.V.L., H.L., H.L., E.C.L., C.L., E.J.L., D.L., M.J.M., G.M., C.L.M., D.M.F., M.W.M., C.Z.M., B.M., S.M., L.R.M., E.M., S.M., T.M., M.E.M., G.M., A.N., Z.O., S.P., S.P.P., S.P., K.P., R.E.P.A., P.J.P., G.P., S.R., L.R., W.R., D.R., I.R., F.R., P.R., S.L.P.S., R.S., R.S., E.S., B.S., J.T., J.V.T., B.W.vB., J.vdK., I.vdB., T.vE., C.M.vR, S.V., C.M.M.V.T., D.P.W., S.W., M.C.A.Y., R.T.Z., B.L., H.G.B., N.d.L., W.P.K., E.C.T. and C.C.M ascertained and enrolled subjects and provided phenotypic information. C.R. and M.E.T. wrote the manuscript, which was approved by all authors.

COMPETING FINANCIAL INTERESTS

The authors have none to declare.

SUPPLEMENTARY INFORMATION

Supplementary material is available online and contains Supplementary Methods, 12 Supplementary Tables and 77 Supplementary Figures.

ONLINE METHODS

Subject Ascertainment

Subjects were enrolled through cytogenetic reference centers including DGAP (the Developmental Genome Anatomy Project) of Brigham and Women's Hospital and Massachusetts General Hospital, Boston, MA; Mayo Clinic, Rochester, MN; University Medical Center, Utrecht, NL; Radboud University, Nijmegen Medical Center, Nijmegen, NL. Enrollment was based on the presence of a developmental anomaly and concomitant BCA (*de novo* or that segregated with the abnormal phenotype) detected by karyotyping, and exclusion of clinically significant genomic copy number imbalances using chromosomal microarray analyses (SNP array or array-CGH) when possible (171/273 tested subjects; **Supplementary Fig. 1**). In the majority of cases the BCA was confirmed to have arisen *de novo* by karyotyping (184/273) or segregated with a developmental phenotype in the family (14/273). In a subset of subjects: (1) the BCA was inherited but the phenotype of the transmitting parent was not available (3/273); (2) one parent was available and did not harbor the BCA (4/273); or (3) neither parents were available for testing (68/273). An informed consent was obtained from all subjects or their legal representative for participation in the study. All studies were approved by respective Institutional Review Boards.

Whole-genome sequencing using large-insert jumping libraries

Samples were prepared using multiple sequencing methods over several years (**Supplementary Table 1**). Most samples were sequenced using whole-genome large-insert jumping library preparation protocols for subsequent Illumina sequencing: 149 using our 2x25-bp EcoP15I protocol[11,74], 59 using a variant of our jumping library protocol in which we randomly shear circularized DNA, which enables longer reads (paired-end 50 bp, see **Supplementary Methods**) and 19 using standard Illumina mate-pair protocols. All large-insert sequencing methods allowed generation of paired-end reads with median insert size of 2.5-3.5 kb as opposed to 300 bp using conventional methods. A subset of samples were prepared with standard short-insert paired-end protocols (n=12) or targeted sequencing of the breakpoints based on previous positional cloning to narrow the breakpoint regions (n=34), as previously described[75,11,7]. Of note, 87 BCAs had been initially reported in the literature, though many had not been mapped to sequence resolution (**Supplementary Table 1**).

Digitalization and homogenization of reported phenotypes

Clinical description was converted for all 273 subjects into standardized terms using Human Phenotype Ontology (HPO; **Supplementary Table 2**)[19]. Such digitalization allowed systematic comparison of phenotypes between subjects carrying BCAs that disrupted the same gene, as well as between subjects with a disrupted gene to previously described subjects using Phenomizer[76]. HPO-sim was used to compute phenotypic similarity scores between subjects sharing the disruption of the same gene or locus compared to random expectations (**Supplementary Table 11**)[67].

BCA discovery pipeline and breakpoint inference

All computational analyses have been previously described[70,77]. In brief, reads were reverse-complemented and aligned using BWA[78]. Anomalous read-pairs in terms of insert size, mate mapping, or mate orientation were extracted using Sambamba and clustered using ReadPairCluster, our single-linkage clustering algorithm[11,79]. Anomalous read-pair clusters meeting our established thresholds of structural variation were subsequently classified based on their read-pair orientation signature into the following categories: deletions, insertions, inversions, and translocations[77]. When no clusters were found that matched the proposed karyotype, BAM files were agnostically analyzed and manually inspected for anomalous pairs or split reads. Breakpoints were successfully identified in 248 of 273 cases, leading to an overall breakpoint fine-mapping yield of 91%. All subsequent counts and yields were computed relative to mapped cases (n=248). For the remaining 25 unmapped cases, no breakpoints were identified in proximity to the karyotype interpretation following extensive analyses and visual inspection. For the majority of these latter unresolved cases, one or more breakpoints were interpreted by the karyotype to localize near centromeres heterochromatic regions, or within segmental duplications, which are recognized to be blind spots for short-read alignments. All large genomic imbalances predicted to be connected to BCA breakpoints following rearrangement reconstruction were confirmed to have aberrant depth of coverage using a custom R-script (CNView: https://github.com/RCollins13/CNView).

When additional DNA was available, precise breakpoint junctions were delineated at base-pair resolution by Sanger sequencing and final breakpoints coordinates reported; else the reported coordinates reflect the minimal breakpoint estimates based on the resolution of the jumping libraries (**Supplementary Table 3**). A total of 82.7% (725/876) of the reported breakpoints could be tested by Sanger sequencing given DNA availability, among which 662 were confirmed yielding a minimum estimate of 91.3% (662/725) sensitivity for our mapping method.

Molecular signature of BCA breakpoints

As previously described[23], we processed all Sanger sequences from validated breakpoints with the BWA Smith-Waterman algorithm (modified parameters z 100 -t 3 -H -T 1) to retrieve precise breakpoint coordinates as well as infer the associated microhomology, micro-insertions or blunt end signature. This approach was sufficiently high-throughput to enable the direct comparison of BCA breakpoints with a large set of deletion breakpoints published by Abyzov *et al.*[27], at the cost of not allowing concomitant microhomology and base insertions at breakpoints.

Monte-Carlo randomization tests

A Browser Extensible Data (BED) file containing GRCh37/hg19 genomic coordinates of all 876 breakpoints detected by WGS was used as the input. One simulation consisted of generating random coordinates based on each pair of input coordinates, conserving the size of the feature as well as the intra-chromosomal distance when several breakpoints were localized to the same chromosome in a single individual. N-masked regions were excluded from simulations for consistency as they were excluded from the initial alignment mapping. Simulations were repeated 100,000 times. The number of unique intersections between the shuffled file and a BED-file containing features of interest (gene-sets, regulatory elements, etc.) was retrieved for each simulation, and the final sets of simulations delineated the expected distribution on intersections under the null hypothesis. The observed value of intersected features in this study was compared to this expected distribution. Empirical Monte-Carlo *P-values* were indicated, and were calculated as follows: P-value = $(r + 1)/(n + 1)$, where r is the number of observations within the set of simulations that are at least as extreme as the one observed, and n is the total number of simulations[80]. References for all functional element datasets and genesets that were used to test for enrichment at breakpoints in the cohort are detailed in **Supplementary Table 12**.

BCA outcome interpretation

To build reference lists of genes associated with dominant developmental disorders we amalgamated data from multiple large-scale exome sequencing, genome sequencing, or CNV studies investigating developmental (*e.g.* DDD consortium) and neurodevelopmental disorders (mostly intellectual disability, autism, and epilepsy cohorts; see **Supplementary Methods** and **Supplementary Table 6** for detailed references). We then built our interpretation using standard categories comparable to those established by ClinVar and the Deciphering Developmental Disorders consortium (DDD)[44], as detailed below and in **Supplementary Table 7**.

Pathogenic: Confirmed Loci associated with developmental disorders. Any gene with three or more *de novo* LoF mutations (frameshift, nonsense or splice mutation, CNV, or BCA) reported from independent cases in those amalgamated studies or in OMIM was considered as high confidence for a particular phenotype, and any BCA impacting one of those loci was therefore considered to be *Pathogenic* (**Supplementary Table 9**).

Likely Pathogenic: Novel candidate genes or mechanisms. To evaluate the impact of the remaining BCAs and the genes they likely impacted, we relied on convergent genomic evidence from other large-scale datasets to prioritize which gene would most likely contribute to the subject's phenotype. Multiple BCAs were considered to be *Likely Pathogenic*, based on various evidences (**Supplementary Table 10**):

(1) Disruption of a likely risk factor: Disruption of one copy of a gene in which one or two dnLoF mutations had been previously reported **and** which demonstrated significant constraint (top 10% of constrained genes)[29,30]

(2) Novel mechanisms: Disruption of a gene established as associated with dominant developmental disorders yet with a distinct mutation type (*e.g.* activating or missense mutations while we reported LoF)

(3) Disruption of long-range interactions: BCA breakpoints located in the vicinity of a gene associated with dominant developmental disorders in a subject with a consistent phenotype, and predicted to impact long-range regulatory interactions.

VUS: All BCAs impacting genes not fitting in any of the above-mentioned categories were considered as VUS.

Predicted disruption of contact domains by BCAs

Topological associated domains (TADs) and predicted loops for lymphoblastoid cells were retrieved from Dixon *et al.* and Rao et *al.*[59,18], and genes contained within a domain for which at least one of its insulating boundaries was disrupted by a BCA were assessed. Only genes that had been previously robustly associated with dominant developmental disorders (*i.e.*, with dnLoF reported in three or more subjects) were considered for potential positional effects. A detailed comparison of the reported phenotypes in the corresponding subjects to phenotypes associated with disrupted genes in the literature was performed. For subjects identified with a BCA of plausible positional effect, the region was visualized using Juicebox[18,81] (**Supplementary Fig. 10-15**). Heatmaps represent observed intrachromosomal interactions in GM12878 lymphoblastoid cells in a specific window; previously reported contact domains (regions of increased contact, not necessarily materializing as loops) and loops

(sites of increased focal contacts indicating the presence of a loop) were indicated[59,18], as well as the RefSeq genes located in the region.

Measuring gene expression from lymphoblasts

In subjects for whom the BCA was suspected to result in positional effects and for whom LCLs derived from blood were available, gene expression was investigated by quantitative RT-PCR. LCLs were not tested for mycoplasma contamination. Total RNA was extracted from LCLs using TRIzol® (Invitrogen) followed by RNeasy Mini Kit (Qiagen) column purification. cDNA was synthetized from 750 ng of extracted RNA using SuperScript® II Reverse Transcriptase (ThermoFisher Scientific with oligo(dT), random hexamers, and RNase inhibitor. Quantitative RT-PCR was performed for mRNA expression of genes of interest in the following subjects (*MEF2C*: DGAP131, DGAP191, DGAP218, DGAP222; *SATB2*: DGAP237; *SLC2A1*: DGAP170; *SRCAP*: DGAP134) using custom designed primers (see **Supplementary Methods**). *ACTB, GAPDH* and *POLR2A* were each used as independent endogenous controls. Custom designed primers (0.75 μM final), cDNA (1:100 final) and nuclease-free water were added to the LightCycler® 480 SYBR Green I Master Mix (Roche) for a final 10 μL reaction volume. A LightCycler® 480 (Roche) was used for data acquisition. Values of each individual (subject or control) were obtained in three technical replicates. Results of technical replicates for each gene of interest were normalized against the average of the three endogenous controls (*ACTB, GAPDH* and *POLR2A*). Normalized expression levels were set in relation to eight age and sex-matched controls for the genes of interest *SATB2, SLC2A1* and *SRCAP*, or 16 (eight males, eight females) age-matched controls for the gene of interest *MEF2C,* using the ΔΔCt method. Results are expressed as fold-change relative to the averaged control individuals. The significance of differential gene expression from a subject in comparison to controls was tested using a two-sided Wilcoxon Mann-Whitney test. All qRT-PCR results were independently replicated twice in the laboratory.

TABLES

Table 1. Overview of clinical phenotypes for all 273 subjects

	Affected subjects	Frequency in cohort
Gender		
Male	159	58.2%
Female	114	41.8%
Co-Segregation		
De novo	184	67.4%
Unknown	75	27.5%
Inherited, segregating	14	5.1%
array-CGH analyses		
Normal	139	50.9%
VUS	32	11.7%
Not Performed	102	37.4%
Abdomen defects	54	19.8%
Cardiovascular defects	41	15.0%
Eye defects	54	19.8%
Hearing defects	52	19/0%
Genitourinary defects	50	18%
Growth defects	64	23%
Head/Neck/Craniofacial	140	51%
Integument defects	50	18.3%
Limb defects	57	20.9%
Musculature defects	71	26.0%
Neurological defects	219	80.2%
Behavior disorders	51	18.7%
Developmental delay	159	58.2%
Epilepsy	51	18.7%
Hypotonia	41	15.0%
ASD/autistic features	31	11.4%
High functioning ASD	4	1.5%
Respiratory defects	30	11.0%
Skeletal defects	116	42.4%

Clinical description was converted for all 273 subjects into standardized terms using Human Phenotype Ontology (HPO)[19], which allowed systematic association with broad phenotypic categories for each enrolled subject.

583 **Table 2.** Genes and loci disrupted by BCAs and likely associated with developmental disorders

Pathogenic	
Genomic imbalances at breakpoints	2q24.3 deletion (*SCN9A*); 4q34 deletion; 6q13-q14.1 deletion (*PHIP*)[a]; 6q14.1 deletion (*TBX18*)[b]; 6q22.1-22.31 deletion (*GJA1*); 10p15.3-p14 deletion (*GATA3*); 11p14.2 deletion; 12p12.1-p11.22 deletion (*SOX5, PTHLH*); 13q14.2 deletion; 14q12-q21.1 deletion (*NFKBIA, NKX2-1*)[c]; 18p11.32-p11.22 deletion[d]; 19q12-q13.11 deletion; Xq25 duplication
Gene disruption	*AHDC1; AUTS2*$_{(x2)}$*; CAMTA1; CDKL5; CHD7; CHD8; CTNND2; CUL3; DYRK1A; EFTUD2; EHMT1; FGFR1; FOXP1; FOXP2; GRIN2B; IL1RAPL1; KAT6B; KDM6A*$_{(x2)}$*; MBD5*$_{(x3)}$*; MEF2C; MTAP; MYT1L*$_{(x2)}$*; MYO6*[e]*; NALCN; NF1A; NFIX; NODAL; NOTCH2; NR2F1; NR5A1; NRXN1; NSD1; PAK3; PDE10A; PHF21A*$_{(x2)}$[d]*; PHIP*[e]*; SATB2; SCN1A; SMS; SNRPN-SNURF*$_{(x3)}$*; SOX5*$_{(x2)}$*; SPAST; TCF12; TCF4; WAC; ZBTB20*$_{(x2)}$
Likely Pathogenic	
Genomic imbalances at breakpoints	2p21-p13.3 duplication (*NRXN1*)
Gene disruption	*ARIH1; BBX; CACNA2D3; CACNA1C; CADPS2*[f] *CDK6*$_{(x2)}$*; CELSR1; EP400*[g]*; GNB1; GRM1*[h]*; KCND2; MDN1; NFIB; NPAS3*$_{(x4)}$[c,i]*; NRXN3; PRPF40A; PSD3*[j]*; PTPRZ1*$_{(x3)}$[a,f]*; ROBO2; SHROOM4*[g]*; SPTBN1; SYNCRIP*$_{(x2)}$[b,j]*; STXBP5*[h]*; UPF2;* 11p15 region
Positional effect	*FOXG1*$_{(x4)}$[i]*; MEF2C*$_{(x7)}$*; PITX2; SATB2*$_{(x3)}$[j]*; SLC2A1; SOX9; SRGAP*

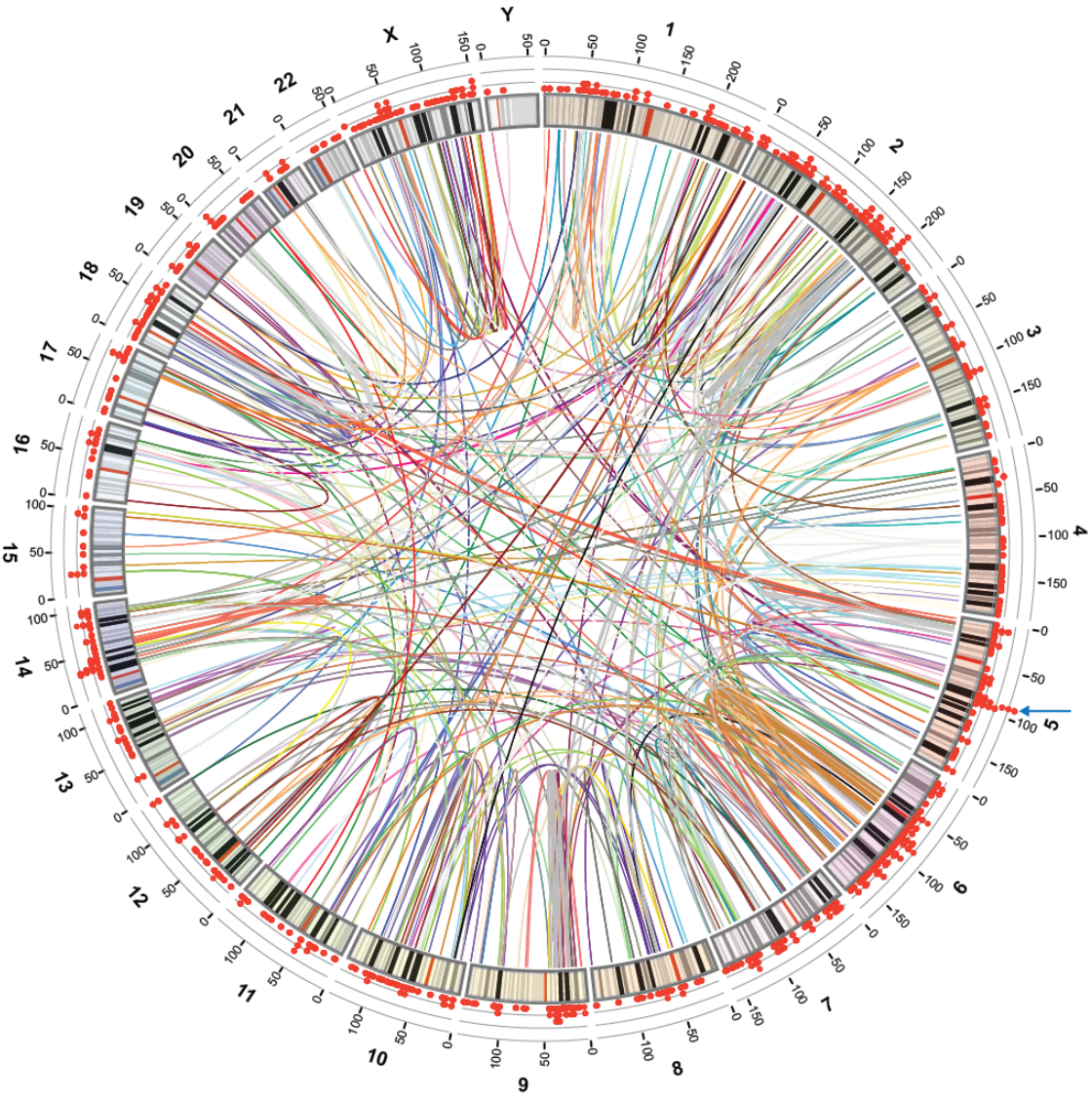
612 Details on BCA interpretation are provided in **Methods** and **Supplementary Table 7**. Genes that have
613 been associated to dominant developmental disorders and encompassed by genomic imbalances at
614 breakpoints are indicated in brackets; lower-scripts indicate when a gene was disrupted by a BCA in

multiple subjects; upper-scripts report subjects with a BCA disrupting multiple genes/loci that may each contribute to their developmental phenotype and to distinct clinical features; [a]: Subject DGAP133; [b]: Subject DGAP317, [c]: subject DGAP002, [d]: subject DGAP316, [e]: subject NIJ2, [f]: subject DGAP168, [g]: subject DGAP172, [h]: DGPA196; [i]: DGAP246; [j]: DGAP237.

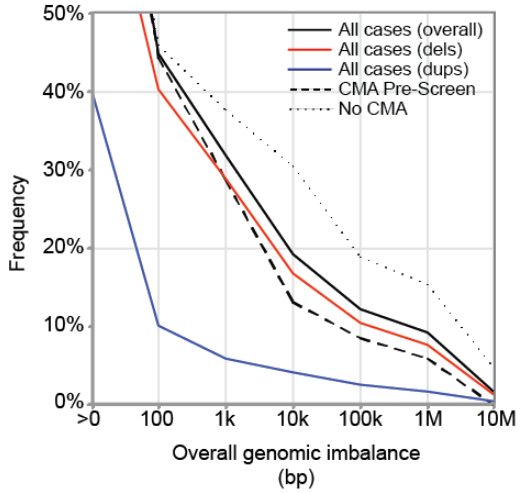
FIGURES

Figure 1. Characterization of BCAs detected by karyotyping at nucleotide resolution

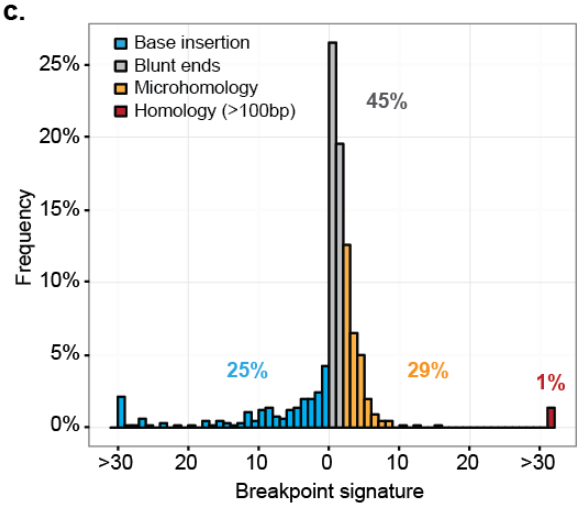
a.



b.

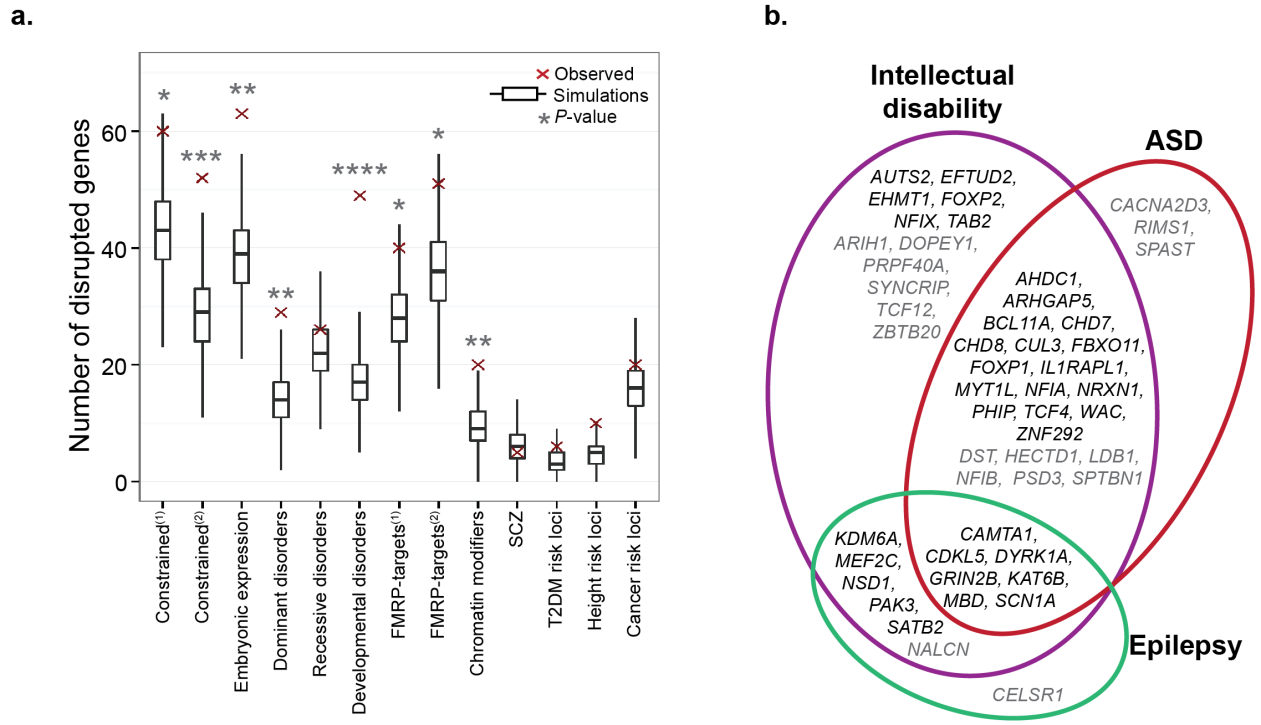


c.



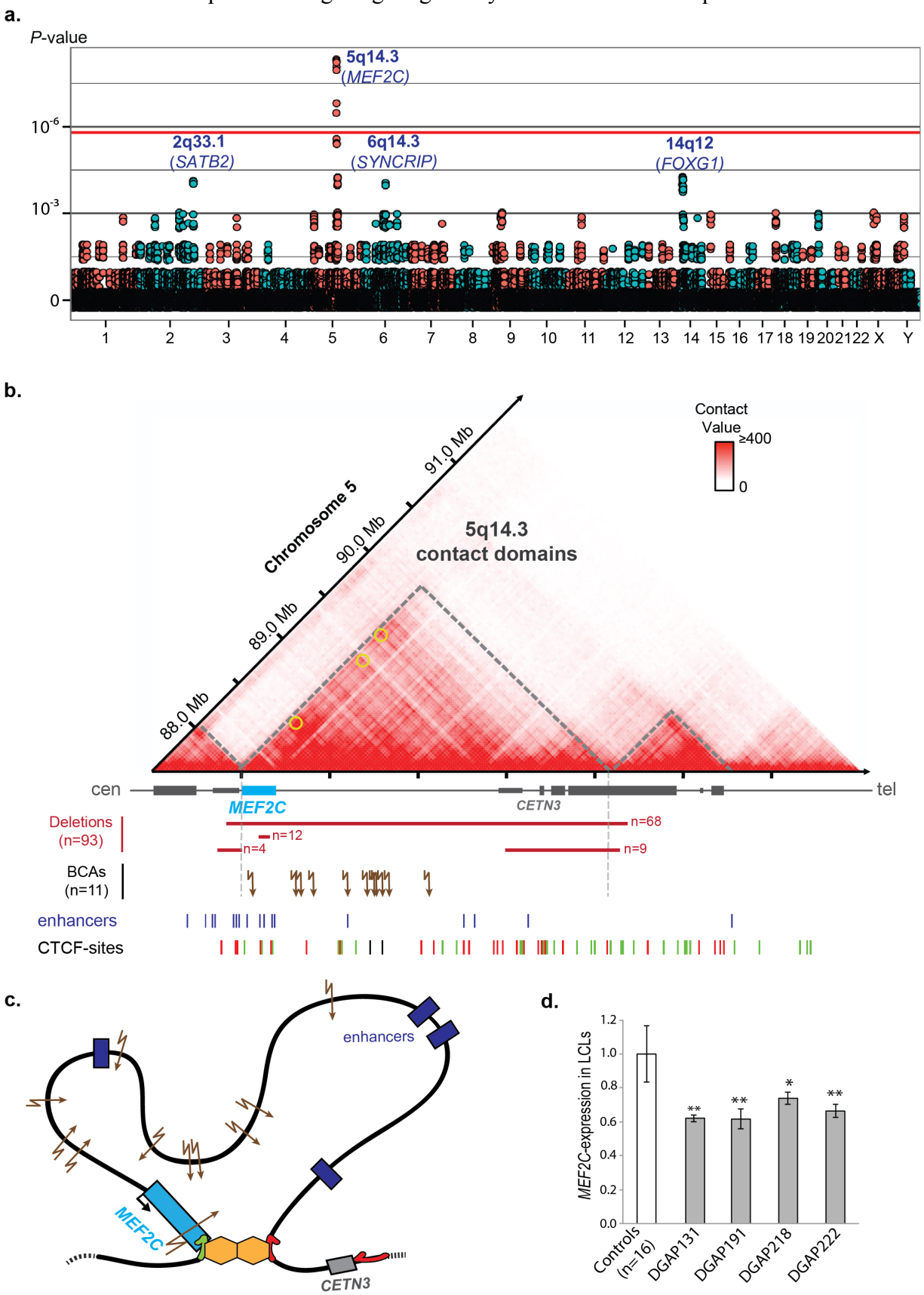
a. Circos plot of all BCA breakpoints identified in the cohort by whole-genome sequencing and their distribution across all chromosomes[82]. One color is used per BCA to represent all rearrangement breakpoints in each individual subject. The scatter plot on the outside ring denotes breakpoint density per 1-Mb bin across the genome, with a blue arrow displaying the largest clustering of breakpoints at the 5q14.3 cytoband; **b**. Scatter plot summarizing the overall genomic imbalance associated with all fully reconstructed BCAs at varying size thresholds. Curves represent the fraction of cases with final genomic imbalances greater than the corresponding size provided (see details in **Supplementary Table 4**). Solid lines denote the final genomic imbalances for all BCAs, and are further delineated by deletions (red) or duplications (blue) emphasizing that cryptic imbalances connected to breakpoints are predominantly copy-losses. The final genomic imbalances among fully mapped BCAs is also split between cases that have been pre-screened by chromosomal microarray (CMA; dashed line) versus cases without CMA data (dotted line); **c**. Sequence signatures of BCA breakpoints. Histogram representing nucleotide signatures at the junction of 662 Sanger-validated breakpoints: inserted nucleotides (blue), blunt ends (grey), microhomology (orange), or longer stretches of homology (red).

Figure 2. *De novo* BCAs associated with congenital anomalies disrupt functionally relevant loci



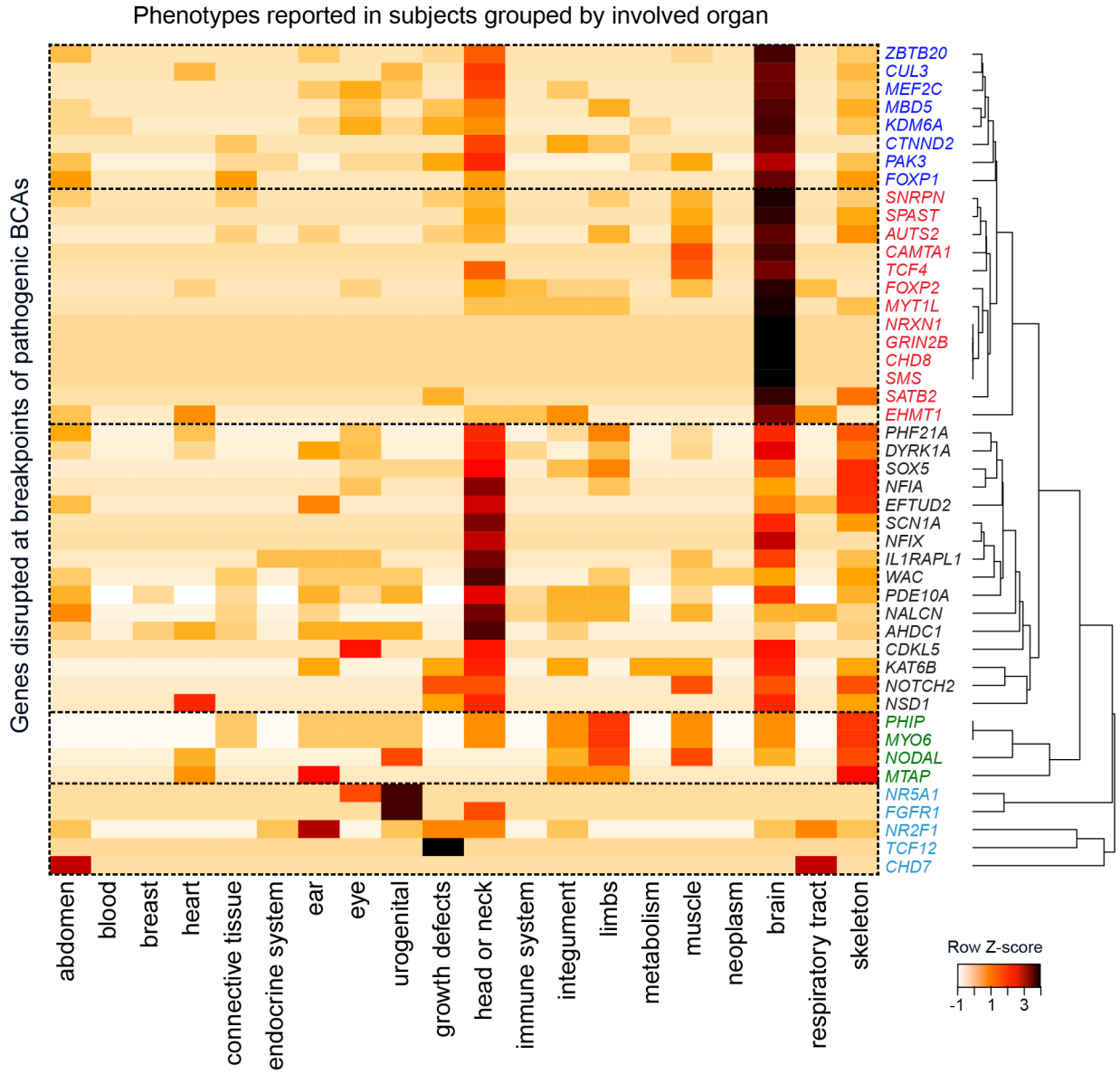
a. Genes localized to BCA breakpoints in subjects with congenital anomalies were significantly enriched for constrained genes (1: Petrovski *et al.*[29] [*P*=0.027], and 2: Samocha *et al.*[30] [*P*=0.0009]), embryonically-expressed genes (*P*=0.001), genes previously associated with autosomal dominant disorders (*P*=0.002), developmental disorders (*P*=0.00002), FMRP-target genes (1: Ascano *et al.*[34], and [*P*=0.036], and 2: Darnell *et al.*[33] [*P*=0.031]), and genes involved in chromatin remodeling *(P*=0.007). Each boxplot represents the expected distribution (median, first and third quartiles) based on total intersections between 100,000 sets of simulated breakpoints and a particular gene-set; red diamonds indicate the observed intersection values against the expected distribution. Empirical Monte-Carlo *P-values* are indicated. *P-values* thresholds were denoted by: * = P≤0.05, ** = P≤0.01, *** = P≤0.001, **** = P≤0.0001; **b**. Venn diagram showing the detailed overlap of disrupted genes that had been previously associated with three neurodevelopmental phenotypes (intellectual disability, ASD, and epilepsy) in amalgamated exome and CNV studies. In black: high-confidence genes (3 or more *de novo* LoF mutations reported), in grey: low-confidence genes (two *de novo* LoF mutations). **c-e**) Pie charts illustrating diagnostic yields associated with the overall cohort and multiple subgroups of BCAs. Clinical interpretation was restricted to *Pathogenic, Likely Pathogenic*, or *Variant of Unknown Significance (VUS)*, as described in the text. **c**. Diagnostic yield associated with 248 mapped BCAs from subjects with congenital or developmental anomalies; **d**. The overall diagnostic yield was significantly higher among BCAs that were confirmed to be *de novo* or segregated with the developmental phenotype (n=198, 186 mapped) compared to the yield from BCAs of unknown segregation status (n=75, 62 mapped); **e.** The overall diagnostic yield associated with BCAs in which large pathogenic CNVs had been excluded by a CMA pre-screen (n=171, 160 mapped) was lower compared to the yield from BCAs that had not been previously screened by CMA (n=102, 88 mapped).

Figure 3. Recurrent disruption of long-range regulatory interactions at the 5q14.3 locus



a. Manhattan plot showing the distribution of all BCA breakpoints in the cohort across each 1-Mb bin of the human genome. P-values were computed by comparing observed to expected cluster sizes after 100,000 Monte Carlo randomizations, and corrected for the total number of windows interrogated. Corrected *P*-values associated with each cluster of breakpoints coming from independent BCAs are indicated. One cluster containing breakpoints from eight independent cases at 5q14.3 achieved genome-wide significance demarcated by the red line (5q14.3 maximum *P*-value=7.7×10^{-9}), while three other regions provided nominal evidence of an unusual cluster of breakpoints (*P*-value = 1×10^{-4}), as highlighted. **b**. The 5q14.3 cluster of eight breakpoints overlaps with a region associated with the 5q14.3-q15 microdeletion syndrome. Multiple pathogenic mechanisms appear to converge on a similar phenotypic consequence: multi-genic deletions that encompass *MEF2C* along with one or both TAD boundaries (n=68), *MEF2C*-intragenic deletions (n=12) or LoF mutations, deletions that do not encompass *MEF2C* but overlap one TAD boundary (n=13), and BCA breakpoints distal to or truncating *MEF2C* (breakpoints from the eight subjects reported in this study along with three previously reported subjects)[52,57,15]. Overlapping Hi-C data from LCLs suggest that the topology of the MEF2C-contact domain is lost in subjects carrying BCAs[18], leading to altered expression of *MEF2C*. Multiple brain-expressed enhancers are located in the region distal to *MEF2C*[83], and three loops involving *MEF2C* have been observed in the region (yellow circles)[18]. Forward (green) and reverse (red) CTCF binding sites are shown, several of which overlap with *MEF2C*-associated loop and domain boundaries; **c**. A proposed model of the chromatin folding in the region defining a regulatory unit for *MEF2C*: a loop is formed anchored at bidirectional CTCF binding sites resulting in distal enhancers being bridged in close proximity to *MEF2C* promoter regulating *MEF2C* expression; **d**. Significantly decreased expression was observed in LCLs from subjects harboring BCAs that disrupt the *MEF2C*-associated TAD when compared to age-matched controls, suggesting regulatory changes via a positional effect that disrupts the *MEF2C* TAD based on real-time qRT-PCR compared to mean expression value from 16 age-matched controls using three technical replicates and normalized against the average of three endogenous controls (*ACTB, GAPDH* and *POLR2A*). Differential gene expression was tested using a Wilcoxon Mann-Whitney test (* $P<0.05$, ** $P<0.01$).

Figure 4. Correlations between phenotypes and genes disrupted in subjects harboring pathogenic BCAs



Heatmap summarizing the correlation between disrupted genes at breakpoints of pathogenic BCAs and phenotypes reported in subjects from this study (**Supplementary Table 2**). For each gene, the phenotypes reported in the corresponding subject were digitalized using HPO terms and grouped together under broad HPO categories[19]. One tile represents the normalized count of HPO terms belonging to each broad category reported in the subject(s). The generated matrix of counts of HPO-terms per category for each gene was normalized per gene, and genes were clustered together when sharing similarly affected organs. Five groups are delineated based on clustering: 1- genes associated with severe nervous system and craniofacial abnormalities (dark blue); 2- genes connected to severe neurological phenotypes (red); 3- genes associated with craniofacial abnormalities and moderate neurological symptoms (black); 4- genes associated with skeletal and limb abnormalities, and with milder neurological involvement (green); 5- genes without neurological involvement (light blue).

REFERENCES

1. Jacobs, P.A., Melville, M., Ratcliffe, S., Keay, A.J. & Syme, J. A cytogenetic survey of 11,680 newborn infants. *Ann. Hum. Genet.* **37**, 359-376 (1974).
2. Nielsen, J. & Wohlert, M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Hum. Genet.* **87**, 81-83 (1991).
3. Ravel, C., Berthaut, I., Bresson, J.L., Siffroi, J.P. & Genetics Commission of the French Federation of, C. Prevalence of chromosomal abnormalities in phenotypically normal and fertile adult males: large-scale survey of over 10,000 sperm donor karyotypes. *Hum. Reprod.* **21**, 1484-1489 (2006).
4. Funderburk, S.J., Spence, M.A. & Sparkes, R.S. Mental retardation associated with "balanced" chromosome rearrangements. *Am. J. Hum. Genet.* **29**, 136-141 (1977).
5. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477-488 (2008).
6. McKusick, V.A. & Amberger, J.S. The morbid anatomy of the human genome: chromosomal location of mutations causing disease. *J. Med. Genet.* **30**, 1-26 (1993).
7. Talkowski, M.E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525-537 (2012).
8. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125-138 (2013).
9. Warburton, D. Current techniques in chromosome analysis. *Pediatr. Clin. North Am.* **27**, 753-769 (1980).
10. Warburton, D. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* **49**, 995-1013 (1991).
11. Talkowski, M.E. *et al.* Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am. J. Hum. Genet.* **88**, 469-481 (2011).
12. Talkowski, M.E. *et al.* Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N. Engl. J. Med.* **367**, 2226-2232 (2012).
13. Schluth-Bolard, C. *et al.* Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. *J. Med. Genet.* **50**, 144-150 (2013).
14. Utami, K.H. *et al.* Detection of chromosomal breakpoints in patients with developmental delay and speech disorders. *PLoS One* **9**, e90852 (2014).
15. Vergult, S. *et al.* Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. *Eur. J. Hum. Genet.* **22**, 652-659 (2014).
16. Tabet, A.C. *et al.* Complex nature of apparently balanced chromosomal rearrangements in patients with autism spectrum disorder. *Mol. Autism* **6**, 19 (2015).
17. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

18. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
19. Kohler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966-974 (2014).
20. Kloosterman, W.P. *et al.* Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20**, 1916-1924 (2011).
21. Meyerson, M. & Pellman, D. Cancer genomes evolve by pulverizing single chromosomes. *Cell* **144**, 9-10 (2011).
22. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
23. Chiang, C. *et al.* Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44**, 390-397, S391 (2012).
24. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).
25. De Gregori, M. *et al.* Cryptic deletions are a common finding in "balanced" reciprocal and complex chromosome rearrangements: a study of 59 patients. *J. Med. Genet.* **44**, 750-762 (2007).
26. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849-853 (2009).
27. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6**, 7256 (2015).
28. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
29. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709 (2013).
30. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944-950 (2014).
31. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
32. Berg, J.S. *et al.* An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**, 36-44 (2013).
33. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011).
34. Ascano, M., Jr. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382-386 (2012).
35. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).
36. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).
37. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).
38. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
39. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4468-4477 (2014).

40. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun* **6**, 6404 (2015).
41. Hawrylycz, M.J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391-399 (2012).
42. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
43. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190 (2014).
44. Landrum, M.J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862-868 (2016).
45. Kleefstra, T. *et al.* Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. *Am. J. Hum. Genet.* **79**, 370-377 (2006).
46. Lu, W. *et al.* NFIA haploinsufficiency is associated with a CNS malformation syndrome and urinary tract defects. *PLoS Genet* **3**, e80 (2007).
47. Rosenfeld, J.A. *et al.* Small deletions of SATB2 cause some of the clinical features of the 2q33.1 microdeletion syndrome. *PLoS One* **4**, e6568 (2009).
48. Talkowski, M.E. *et al.* Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.* **89**, 551-563 (2011).
49. Rasmussen, M.B. *et al.* Neurodevelopmental disorders associated with dosage imbalance of ZBTB20 correlate with the morbidity spectrum of ZBTB20 candidate target genes. *J. Med. Genet.* **51**, 605-613 (2014).
50. Splawski, I. *et al.* Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8089-8096; discussion 8086-8088 (2005).
51. Petrovski, S. *et al.* Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability, Hypotonia, and Seizures. *Am. J. Hum. Genet.* **98**, 1001-1010 (2016).
52. Floris, C. *et al.* Two patients with balanced translocations and autistic disorder: CSMD3 as a candidate gene for autism found in their common 8q23 breakpoint area. *Eur. J. Hum. Genet.* **16**, 696-704 (2008).
53. Cardoso, C. *et al.* Periventricular heterotopia, mental retardation, and epilepsy associated with 5q14.3-q15 deletion. *Neurology* **72**, 784-792 (2009).
54. Engels, H. *et al.* A novel microdeletion syndrome involving 5q14.3-q15: clinical and molecular cytogenetic characterization of three patients. *Eur. J. Hum. Genet.* **17**, 1592-1599 (2009).
55. Le Meur, N. *et al.* MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J. Med. Genet.* **47**, 22-29 (2010).
56. Zweier, M. *et al.* Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. *Hum. Mutat.* **31**, 722-733 (2010).
57. Saitsu, H. *et al.* De novo 5q14.3 translocation 121.5-kb upstream of MEF2C in a patient with severe intellectual disability and early-onset epileptic encephalopathy. *Am. J. Med. Genet. A* **155A**, 2879-2884 (2011).
58. Zweier, M. & Rauch, A. TheMEF2C-Related and 5q14.3q15 Microdeletion Syndrome. *Mol. Syndromol.* **2**, 164-170 (2011).

59. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
60. Lupianez, D.G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025 (2015).
61. Lupianez, D.G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* **32**, 225-237 (2016).
62. Mencarelli, M.A. *et al.* 14q12 Microdeletion syndrome and congenital variant of Rett syndrome. *Eur. J. Med. Genet.* **52**, 148-152 (2009).
63. Ellaway, C.J. *et al.* 14q12 microdeletions excluding FOXG1 give rise to a congenital variant Rett syndrome-like phenotype. *Eur. J. Hum. Genet.* **21**, 522-527 (2013).
64. Perche, O. *et al.* Dysregulation of FOXG1 pathway in a 14q12 microdeletion case. *Am. J. Med. Genet. A* **161A**, 3072-3077 (2013).
65. Takagi, M. *et al.* A 2.0 Mb microdeletion in proximal chromosome 14q12, involving regulatory elements of FOXG1, with the coding region of FOXG1 being unaffected, results in severe developmental delay, microcephaly, and hypoplasia of the corpus callosum. *Eur. J. Med. Genet.* **56**, 526-528 (2013).
66. Ibn-Salem, J. *et al.* Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* **15**, 423 (2014).
67. Deng, Y., Gao, L., Wang, B. & Guo, X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One* **10**, e0115692 (2015).
68. Brunetti-Pierri, N. *et al.* Duplications of FOXG1 in 14q12 are associated with developmental epilepsy, mental retardation, and severe speech impairment. *Eur. J. Hum. Genet.* **19**, 102-107 (2011).
69. McDermott, S.M. *et al.* Drosophila Syncrip modulates the expression of mRNAs encoding key synaptic proteins required for morphology at the neuromuscular junction. *RNA* **20**, 1593-1606 (2014).
70. Brand, H. *et al.* Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am. J. Hum. Genet.* **95**, 454-461 (2014).
71. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688-696 (2014).
72. Chaisson, M.J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611 (2015).
73. Lettice, L.A. *et al.* Enhancer-adoption as a mechanism of human developmental disease. *Hum. Mutat.* **32**, 1492-1499 (2011).
74. Hanscom, C. & Talkowski, M. Design of large-insert jumping libraries for structural variant detection using illumina sequencing. *Curr Protoc Hum Genet* **80**, 7 22 21-29 (2014).
75. Higgins, A.W. *et al.* Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *Am. J. Hum. Genet.* **82**, 712-722 (2008).
76. Kohler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457-464 (2009).
77. Brand, H. *et al.* Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. *Am. J. Hum. Genet.* **97**, 170-176 (2015).
78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

79. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
80. North, B.V., Curtis, D. & Sham, P.C. A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.* **71**, 439-441 (2002).
81. Durand, N.C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101 (2016).
82. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).
83. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).