Online/Supplementary Methods

Samples

We obtained DNA from 52 independent subjects with karyotypically balanced chromosomal rearrangements (reciprocal translocations and inversions). All rearrangements were either *de novo* or inherited from a similarly affected parent (two subjects, one of which was a CCR). Subjects were referred to various research studies including the Developmental Genome Anatomy Project (DGAP), the Autism Consortium of Boston, the Autism Genome Resource Exchange (AGRE), and individuals referred directly to a local protocol for the purposes of these experiments. All subjects provided written informed consent or de-identified DNA was obtained from the AGRE subjects and all protocols were approved by the Institutional Review Board of Partners HealthCare System. We also obtained breakpoint sequences from eight DGAP subjects for which capillary sequencing had previously resolved the junctions (n=8)[33]. Phenotypes associated with balanced SVs varied widely across DGAP subjects (see Higgins et al., 2008), but the most common phenotype across all subjects was a neurodevelopmental abnormality.

Transgenic Animals

Rearrangements were evaluated from previously reported transgenic animals created by pronuclear injection. We sequenced five sheep lines (G_0/1, G_0/2, G_0/5, G_0/4, G_0/6) generated by microinjection into pronuclei of single-celled zygotes of an 11.6 kb transgene restriction enzyme generated linear fragment, containing a full-length human *HTT* cDNA flanked by 1.1 kb of the human 5' UTR sequence (5' end) and exon 4, intron 4 and exon 5 of the bovine growth hormone (BGH) gene (3' end)[24]. R6/1 and R6/2 mice were generated by microinjection of a 1.9 kb linear genomic DNA fragment of the human *HTT* gene (derived by restriction enzyme cleavage of a phage genomic DNA clone), containing ~1kb of human 5' UTR, exon 1 and a portion of intron 1

(Fig. 3)[23]. Sheep genomic DNAs were prepared from tail tissue and the R6/1 and R6/2 mouse genomic DNAs were purchased from the Jackson laboratory (stock #2810 and #2809).

Massively parallel sequencing of breakpoints

Human Subjects: All sequencing was performed using Illumina GAIIx or HiSeq 2000 chemistry (Illumina Inc) and library formation was performed using various approaches including manufacturer protocols and customized methods, with complete details available in a previous publication[18]. In brief, we used four methods to obtain translocation and inversion breakpoint sequences. Whole-genome sequencing of paired-end 76 base pair reads was performed for two subject to 10X average nucleotide coverage across the genome. Illumina mate-pair sequencing (n=5) was performed based on the methods of Korbel et al.[34] following manufacturer's instructions to generate 76 bp paired-end sequence of fragments separated by large inserts of ~3–4 kb. Custom jumping libraries (n=25) were created for Illumina sequencing by combining aspects of the mate-pair protocols for Illumina and Applied Biosystems ABI SOLiD sequencing (Life Sciences, Inc). Briefly, 10-20 ug of DNA was sheared using a Covaris S2 with calibrated parameters and size selected to derive a tight distribution of fragments at ~3.5 kb +/- 300 bp. Cap adapters with an EcoP15I restriction site were ligated to fragment ends and circularized to an internal adapter oligo with a subject specific bar code and a single biotinylated thymine. Restriction fragment digestion was then carried out and biotinylated fragments retained by streptavidin binding. Libraries were then prepared for Illumina sequencing using off the shelf reagents from NEB (New England Biolabs, Inc) and either Illumina oligos or custom oligos designed from IDT based on published sequences[35], followed by sequencing on a single lane of a GAIIx, a single lane of a HiSeq2000, or multiplexed with other subjects on a single lane of a HiSeq2000, dependent on available yields at the time of the experiment. Targeted physical

coverage was 50X from mappable inserts and previous analyses suggested an expectation of 98%-99.8% of all sequenced fragments would contain large inserts spanning the circularization junction[18]. Breakpoints for twelve of the subjects were sequenced by a regional capture approach (CapBP)[18] for subjects in which previous cytogenetic analyses had narrowed at least one breakpoint to within a megabase. Agilent Sure-Select on-array 244K oligonucleotide 60mer probes (n=11) or in-solution 55K 120mer RNA baits (n=1) (Agilent Inc) were used. Regions were tiled at maximum possible overlap and all bases, including those flagged by RepeatMasker[36] were targeted except for single nucleotide, dinucleotide, and tri-nucleotide repeats extending over 30 bases. Samples were individually captured, quantified, and then pooled without indexing on a single lane of an Illumina flow cell. Agilent 1M aCGH array analyses were performed for BSID42 and BSID43.

Transgenic Animals: We used SureSelect In-Solution target enrichment technology (Agilent, Inc) following Manufacturer's Instructions to specifically capture the transgene sequence in each animal. Briefly, we generated 120mer biotin-tagged probes tiled across the transgene sequence at 1 base pair on each strand. We systematically filtered out any probes with singleton, doubleton, and triplet repeats. We performed nine parallel PCR reactions for each adapter ligated library prior to hybridization. DNA libraries were incubated with capture library probes and appropriate blocking oligonucleotides for 24 hours on a thermal cycler at 65°C. Following incubation, libraries were amplified using a SureSelect post-capture indexing primer (reverse primer) and subject specific indexing primers from the Illumina multiplexing kit. We also sequenced customized jumping libraries for the transgenic animals using identical methods described above.

Molecular confirmation of breakpoints by orthogonal methods

All breakpoints described were confirmed by PCR amplification and capillary sequencing, which included a series of long range PCRs and capillary sequencing directly through each of the junction fragments. These amplifications were based on unique primers in the genome, either localized to different chromosomes for translocations or separated in genomic distance and in same-strand orientation for the inversions. In addition, for the R6/2 mouse and the G_0/1 sheep the entire transgene was amplified and capillary sequenced. In the R6/2 model, this analysis was repeated by amplifying the full length of the transgene in a single reaction using primers localized to the transgene/genomic DNA junction to provide product specificity. The transgene was then sequenced from this single amplification product, including all CAG repeats, confirming each junction fragment detected by the read pairs, jumping libraries, and assembly. In the R6/2 mice, a secondary multiplex-ligation probe amplification (MLPA) assay was initially used to confirm the 5 kb deletion. This validation was then performed in three independent R6/2 lines and three wild-type mice used as controls, and each of the six MLPA experiments were performed in triplicate. In addition, for the chromothripsis samples validation of the rearrangement structures presented in Figure 2 was performed using Fluorescence In Situ Hybridization (FISH). Lymphoblast cells were harvested and metaphase spreads prepared by standard protocols. BAC clones spanning breakpoint regions were selected for FISH mapping using the University of California Santa Cruz (UCSC) Genome Browser. BACs from the RP11 library were labeled with either SpectrumOrange or SpectrumGreen conjugated dUTP using a nick translation kit (Abbott Molecular, Des Plaines, IL) and labeled pairs were hybridized overnight to metaphase chromosome preparations. After 4x SSC/0.1% Tween in 2xSSC/0 .1% Tween, and phosphate-buffered detergent washes, chromosomes were counterstained with DAPI

and analyzed with a Zeiss Axioplan2 microscope (Thornwood, NY) with epi-fluorescence and Applied Imaging CytoVision software (Santa Clara, CA).

Bioinformatics and Statistical Analyses

Alignments of paired-end reads for human sequencing were performed using either MAQ[37], BWA[38], or Novoalign (Novocraft, Inc) to the hg19 reference, dependent on the time of the analysis and library type (all CapBP samples were aligned with BWA, jumping libraries were generally aligned with either MAQ or Novoalign) and SAMtools[39]. Sequencing alignments were processed with BamStat, a program developed to identify anomalously mapping read pairs, and statistical analyses were performed using R. In transgenic animals, we also attempted naïve assembly of the transgene with Velvet 1.0.18[40] to corroborate the breakpoint junctions. Microhomology was assessed using the EMBOSS Needle program as described previously[18] as well as with BWA Smith-Waterman 0.5.9[38] and a version of BreakSeq that was modified to assess translocations and inversions[25]. Simulation experiments were performed using custom Python scripts and BEDTools[33].