Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration

Colby Chiang[#], Jessie C. Jacobsen[#], Carl Ernst[#], Carrie Hanscom, Adrian Heilbut, Ian Blumenthal, Ryan E. Mills, Andrew Kirby, Amelia M. Lindgren, Skye R. Rudiger, Clive J. McLaughlan, C. Simon Bawden, Suzanne J. Reid, Richard L. M. Faull, Russell G. Snell, Ira M. Hall, Yiping Shen, Toshiro K. Ohsumi, Mark L. Borowsky, Mark J. Daly, Charles Lee, Cynthia C. Morton, Marcy E. MacDonald, James F. Gusella, and Michael E. Talkowski*

[#]equally contributing authors

*Correspondence
Michael E. Talkowski, Ph.D.
Center for Human Genetic Research
Massachusetts General Hospital, CPZN5830
185 Cambridge Street, Boston, MA 02114
talkowski@chgr.mgh.harvard.edu

SUPPLEMENTARY INFORMATION

Supplementary Bioinformatic and Statistical Methods

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]. BWA alignments were run using default parameters except disabling Smith-Waterman alignment for an unmapped mate. Novoalign and MAQ alignments were run using default parameters. Discordant paired-end alignments were filtered and clustered using BamStat, a custom program designed to obtain alignment metrics and search for anomalous mapped pairs indicative of a rearrangement[19].

For the analytical purposes of this study, we define "breakpoint" as any junction between two genomic DNA sequences that are discordant in location and/or strand orientation, allowing for more than two breakpoints in a complex rearrangement. Extra bases inserted at the breakpoint were considered to be an independent fragment if at least 100 bp aligned contiguously to the genome. Breakpoint junctions were confirmed and localized to basepair resolution by capillary sequencing, with homology included in the positions at both edges of the breakpoint. We assessed microhomology by using the EMBOSS Needle pipeline as described previously by Kidd et al. (2010). An optimal global alignment was determined between the breakpoint sequence and each of the two flanking genomic sequences using the Needle program with default mismatch and gap penalties. Then these alignments were merged into a single three-sequence alignment, from which the microhomology or extra bases were counted. Mismatches within a stretch of microhomology were not counted as contribution to the homology, and concurrent microhomology and inserted sequence was attributed when the two events occurred within 10 bp of the breakpoint. We classified the homology as type I if one copy remained at the derivative breakpoint, and type II if both copies remained, but only observed type I in this study.

To confirm the sensitivity of our methods to the homology reported in other studies, we reanalyzed our data by aligning the sequence breakpoints to the hg19 reference genome using BWA 0.5.9 Smith-Waterman (BWA-SW) alignment and the following parameters: bwa bwasw -z 100 -t 3 -H -T 1. We used the CIGAR string to determine the overlap of multiple alignments for each sequence, giving a positive number for microhomology, and a negative value for inserted bases at a breakpoint. A comparison of the homology between the Needle and BWA Smith-Waterman methods shows nearly identical distributions (Figure S2). As the BWA Smith-Waterman pipeline is much higher throughput (at the expense of allowing for concurrent microhomology and inserted bases at breakpoints), we analyzed the 1,000 Genomes data with the BWA method (Figure S4). Finally, we modified the program BreakSeq to perform a parallel analysis to those in Mills et al. (2011) for the translocations and inversions sequenced here. BreakSeq is an established approach to determine potential mechanisms of variant formation[25] and has the advantage of also allowing for the identification of the ancestral state of many of the queried variants. The current version does not work with translocations and inversions, however, and thus we modified the pipeline to use synthetically created deletions around the initial breakpoint locations in order to work with the existing framework. These synthetic variants were analyzed using the entire pipeline, and assessed for microhomology and other mechanistic signatures. From the combination of these methods we were able to consider the final breakpoint junctions and the sequence features of the initially intact chromosomes prior to their breakpoint disruption. All analytical strategies yielded consistent results.

For the transgenic models, two complementary analytical strategies were used to derive transgene internal sequence and insertion sites from the targeted capture sequencing. In the first, we performed read-pair alignments with BWA version 0.5.9 using default parameters except for disabling Smith-Waterman alignment for an unmapped mate[38]. At the time of the capture analyses, no published sheep reference genome was available. We therefore performed all alignments using a custom reference in which we combined the Baylor 4.0/bosTau4 build of the cow genome with the sequence of the transgene appended as a synthetic "chromosome". The transgenic mice were aligned using similar methodology with the R6 insert sequence appended to the mouse genome reference (NCBI37/mm9). Processing of aligned reads was done using BamStat, a custom program designed to search for anomalous mapped pairs from the alignment data [19]. Insertion sites were identified as chimeric reads in which 10 or more read pairs spanned the insertion junction, meaning one end of the read aligned to the transgene "chromosome" and the mate pair aligned to the reference genome. Similar analyses were able to distinguish junction sites resulting from rearrangements within the transgene itself.

The second method attempted to naively assemble the insertion site and transgene architecture with Velvet 1.0.18[40], using a k value of 45 and parameters based on the mean insert length and expected coverage over the transgene. We anticipated difficulty in assembling reads over the repetitive regions of the transgene (*e.g.,* the ~69-~150 unit *HTT* exon 1 CAG repeat sequence in the cDNA and genomic transgenes, respectively, for which we did not design capture probes), but otherwise expected the assembly to corroborate the insertion sites and rearrangement junctions detected from the BamStat gapped alignments.

Whole-genome jumping libraries of animal models were aligned with BWA version 0.5.9 and the transgenic sheep were mapped to the OAR2 build of the sheep genome (which became available during these experiments) with the transgene sequence appended as a synthetic "chromosome". As with the capture sequencing libraries, discordant reads were filtered and clustered for rearrangements using BamStat.

Simulation experiments were performed using custom Python scripts and BEDTools[32]. We first simulated 10,000 sets of random rearrangements in the genome, similar in size to our experimental set, to test for enrichment of annotated elements. Because many cases had multiple breakpoints, we defined a "breakpoint set" of 168 locations that included all breakpoints from 52 subjects with breakpoints less than 1000 bp apart collapsed into a single breakpoint at their midpoint, and the distribution of these "breakpoint sets" was compared to our observed distributions. Finally, to assess potential DNA structural characteristics in the vicinity, breakpoint sets were created with windows of 1bp – 500 bp spanning either side the breakpoint and 1 million simulations were performed. In all simulations, empirical probabilities were obtained by comparing the events on the tail of the distribution that exceeded our observed results.

Supplementary Results

The proximity of the 5q14.3 rearrangements between BSID42 and BSID43 was noteworthy in that the nearest breakpoints were 3.39 Mb apart, though there was no actual overlap in the disrupted segments. This region has 14 defined DECIPHER rearrangements and a number of Database of Genomic Variation rearrangements in normal individuals. The region is reasonably conserved and contains four segmental duplications of greater than 1000 bases, one of which is an 18.6 kb sequence stretch with high homology to chr7, though this is located ~52 Mb centromeric to the shattered chr7 region of BSID43. A comprehensive assessment of many additional subjects will be required to determine whether sequence features in the region might mediate formation of particularly complex rearrangements or alternatively, the proximity of the breakpoints in BSID42 and BSID43 is coincidental.

MLPA analyses confirmed the deletion in all R6/2 mice and absence of deletion in all wild-type mice (Fig. S5). All FISH results also confirmed the findings of the paired-end sequencing and subsequent PCR amplification and capillary sequencing of all breakpoints. For BSID42, BACs that spanned six of the breakpoints were available and estimated to be sufficiently separated to provide an interpretable signal. RP11-62j3 hybridizes to chr X and der(X) at the telomeric section of Xb and the telomeric section of Xc, while RP11-637d5 hybridizes to chr 5, der(5) insert 5j, and der(X) insert 5k. As shown on the metaphase chromosomes (Fig. S6a), there are two green signals on der(X) for RP11-62j3 with one signal partially overlapping with a red signal for RP11-637d5. There are two additional red signals on chr 5 and der(5). The interphase nucleus (Fig. S6b) also shows three green signals for RP11-62j3. RP11-639f3 exhibits three signals, one on chr 5 and two on der(X) at inserts 5k and 5l. RP11-698d19 hybridizes to chr 5, der(5) insert 5f, and der(X) inserts 5d, 5g and 5e. Signals for inserts 5d and 5g are indistinguishable due to the resolution of FISH analysis. The metaphase chromosomes (Fig. S6c) clearly show two green signals on der(X), one of which overlaps with the red signal from section 5e to create a yellow signal. The interphase nucleus (Fig. S6d) shows a yellow signal on chr 5, and two signals present on der(X): a yellow signal and a juxtaposed red-green signal due to the close proximity of the probes. The der(5) insert 5f was not detectable by FISH. RP11-622p22 hybridizes to chr 5, der(5) telomeric section of 5j, and der(X) insert 5k. There are three red signals for RP11-622p22 in the metaphase chromosomes (Fig. S6e). The proximity of RP11-622p22 and RP11-639f3 results in yellow signals on chr 5 and der(X). A red signal is observed on der(5). The interphase nucleus (Fig. S6f) exhibits a yellow signal on chr 5, a red signal on der(5), and a fused green-red-green signal which represents the two green RP11-639f3 and one red RP11-622p22 signals found on der(X).

BSID43 breakpoints were also validated by FISH and BACs were available that spanned ten of the breakpoints. RP11-111a20 hybridizes to chr 3, der(5) insert 3c and der(7) at the centromeric section of 3b, while RP11-262j1 hybridizes to chr 7 and der(5) at the telomeric section of 7c and 7b. Signals for inserts 7b and 7c are indistinguishable due to the resolution of FISH analysis, consequently only one signal is present on der(5). As shown on interphase nuclei, directly juxtaposed red-green signals (Fig. S7a) or a fused yellow signal (Fig. S7b) are observed for the RP11-111a20 and RP11-262j1 hybridizations on the der(5), two green signals for RP11-111a20 and one red signal for RP11-262j1. RP11-963b6 exhibits two signals, one on chr 7 and the other on der(7) at inserts 7f and 7g. RP11-257f18 hybridizes to chr 7, der(3) at the telomeric section of 7h, and der(7) insert 7i. As shown on metaphase chromosomes (Fig. S7c, d), a single yellow signal is seen on the der(7), a juxtaposed red-green signal is present on chr 7, and a red signal on der(3) for RP11-257f18. Three signals are observed for RP11-317k13 on chr

7, der(3) insert 7h and der(7) insert 7g, while RP11-887a20 hybridizes to chr 5, der(3) insert 5b and der(5) insert 5a. As shown on a metaphase cell (Fig. S7e) and on an interphase nucleus (Fig. S7f), three green signals are visualized, one of which is present in close proximity on the der(3) to a red signal also on the der(3) resulting in a yellow signal in the metaphase. Three red signals (one a doublet in the metaphase representing the replicated locus) are present. RP11-433p9 hybridizes to chr 5, der(3) insert 5b and der(5) insert 5a. RP11-257f18 hybridizes to chr 7, der(3) at the telomeric section of 7h, and der(7) insert 7i. As seen in the metaphase cell (Fig. S7g) and on interphase nuclei (Fig. S7h), three green and three red signals are present; proximity of the hybridizations of RP11-433p9 (green signal) and RP11-257f18 (red signal) results in a yellow signal. RP11-661d13 exhibits three green signals on metaphase chromosomes at chr 7, der(7) insert 7f and der(5) sections 7c and 7d (Fig. S7i, S7j); the der(5) sections 7c and 7d are individually indistinguishable due to resolution of the FISH. The ~5 kb section 7e in the der(3) is too small to be detected by FISH analysis. Three red signals are seen for RP11-80i8, which hybridizes to chr 3, der(3) insert 3a and der(7) at the telomeric section of 3b (Fig. S7i). One red and one green signal are seen on the der(5). Amplification of the junction sequences in BSID43 were performed alongside parental and control DNA with identical primers, resulting in positive signals for the proband but no amplification in either the parent or control samples at any of the breakpoint junctions (Fig. S8). A similar experiment was performed on the 81 kb inversion junction fragment of the transgenic sheep $G_0/2$, in which PCR amplification confirmed the presence of the junction in $G_0/2$ with no signal from any of the other five animals (Fig. S9).

Supplementary Discussion

Our results also point to general limitations in the interpretation of current genomics studies, including a few of our own cases. It is clear that while aCGH is sensitive to the extensive CNVs seen in cancer chromothripsis and to germline CCRs that involve large changes in dosage (Stephens et al., 2011)[6], this technology is insensitive to those shattering events that resolve to a relatively dosage-balanced state. The aCGH approach thus fails to account for a meaningful portion of the mutational spectrum that can underlie human disease. Similarly, the CapBP targeted sequencing approach performed here for 12 of the subjects would have been insensitive to other complex genomic rearrangements remote from the < 1 Mb targeted region, so we might have underestimated the frequency of CCRs. Moreover, it should be noted that these analyses and those of the previous genome-wide sequencing studies only account for the annotated portions of the human genome alignable by short-read sequences and are often blind to important events mediated by mechanisms such as NAHR. While we were able to localize breakpoints in most complex regions by our approach, there were two examples of subjects where we were unable to resolve fully all of the breakpoints due to sequence complexity or repetitive genomic regions (Supplementary Table 1).

Supplementary Figure S1: Chromosomal rearrangements of two transgenic sheep. **A**. Rearrangements in G_0/2 showing ten junctions within the transgene, two junctions between the transgene and *OAR17*, and an 81.8 kb inversion on *OAR17*. **B**. Rearrangements in G_0/6 showing junctions between the transgene and six different sheep chromosomes, including a complex interchromosomal rearrangement between *OAR7*, *OAR8*, and *OAR15*.

Supplementary Figure S2: Microhomology distributions of 141 chromosomal rearrangement breakpoints from two independent algorithms. A: Microhomology profile of breakpoint sequences using the EMBOSS Needle program as described in Kidd et al., 2010. B: Microhomology profile of the same sequences derived from alignment with BWA Smith-Waterman.

Supplementary Figure S3: Microhomology profile of A) 16 breakpoints from seven BCRs in this study that were karyotyped as inversions, as compared to 109 inversion breakpoints sequenced by Kidd et al. (2010), showing an NAHR dominated distribution of the benign CNVs observed from Kidd et al. (2010 but a dearth of microhomology for the karyotypically defined and clinically interpreted as pathogenic inversions in this study compared to those of Kidd et al. and benign CNVs from the 1,000 Genomes Project.

Supplementary Figure S4: Direct comparison of A) the 141 BCRs sequenced in this study to B) 16,783 CNV breakpoints from the 1,000 Genomes Pilot 1 study using the identical BWA Smith-Waterman pipeline for both datasets.

Supplementary Figure S5. Multiplex ligation-dependent probe amplification (MLPA) deletion assessment of mouse chromosome 4 in R6/2 and non-carrier mouse cortex. Peak ratio analysis of A) non-carrier (three mice) and B) R6/2 (three mice) confirm the 5.4 kb deletion of the host genome. The probe mix contained 10 probes; 4 control probes to sequence up- or down-stream of the deletion 'C', 4 probes to the deleted region 'D', and two probes to each of the chromosome/transgene junctions 'J'. The squares indicate deleted regions (<0.75), transgene/chromosome junctions (>1.30), or regions considered to be normal (0.75 - 1.30). Each biological sample was repeated in triplicate. *The right hybridization sequence of this probe was primarily designed to mouse chromosome 4 (71.4%) due to lack of suitable sequence in the transgene at this junction, hence, to a lesser extent, it also detects the wild-type allele and therefore shows a typical mosaic pattern.

Supplementary Figure S6. FISH analyses for BSID42 as detailed in the Supplementary Results. **A,B)** RP11-62j3 (green) and RP11-637d5 (red), **C,D)** RP11-639f3 (green) and RP11-698d19 (red), **E,F)** RP11-639f3 (green) and RP11-622p22 (red)

Supplementary Figure S7. FISH analyses for BSID43 as detailed in the Supplementary Results. **A,B)** RP11-111a20 (green) and RP11-262j1 (red); **C,D)** RP11-963b6 (green) and RP11-257f18 (red); **E,F)** RP11-317k13 (green) and RP11-887a20 (red); **G,H)** RP11-433p9 (green) and RP11-257f18 (red); **I,J)** RP11-661d13 (green) and RP11-80i8 (red)

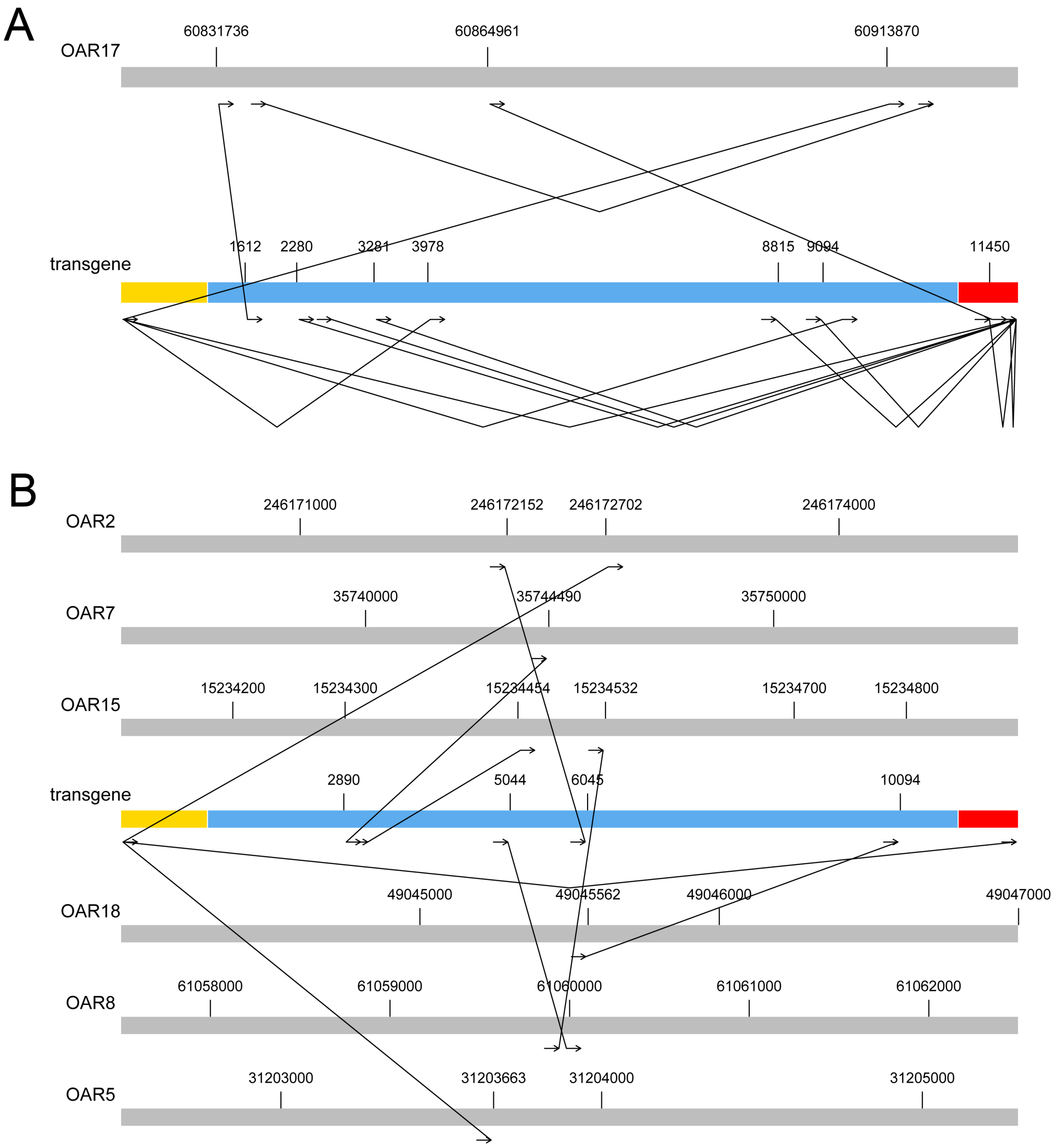
Supplementary Figure S8. A) Breakpoint amplification of each of the breakpoints of BSID43 (top) with control amplifications (bottom). No amplification was seen in control for any of the primer pairs specific to the breakpoint sequence. + = positive control primer pair. B) Breakpoint amplification in the available parent for BSID43 (top) and re-amplification of all 11 breakpoints in the proband (bottom). The amplifications were run independently between A and B with identical results in the proband.

Supplementary Figure S9. Breakpoint amplification of an 81 kb inversion breakpoint distal to the transgene integration site. Primer pairs were specific to the sheep genome (OAR17), in same strand orientation, and separated by 81 kb of sheep sequence. No amplification was seen in any of the five independent sheep genomes studied.

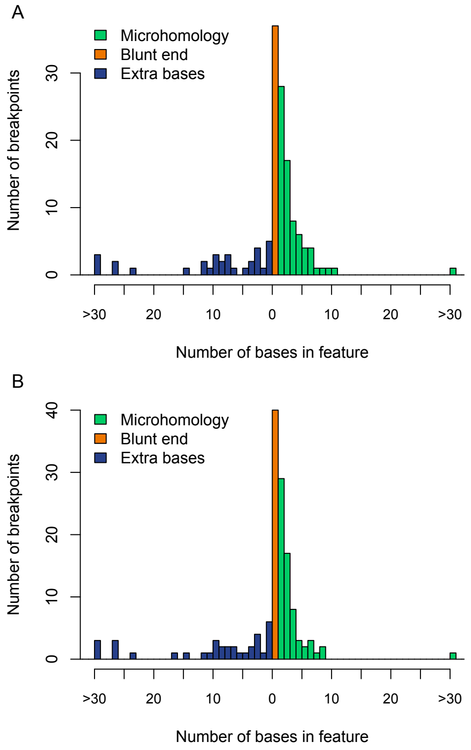
Supplementary Movie S1. Animation of chromosomal breakage and reconstitution of BSID42. Karyotypic analysis indicated a 5q to Xq balanced translocation, however sequencing revealed a 'shattering' and aberrant reorganization of localized genomic regions similar to those recently reported in cancer cells (i.e., chromothripsis).

Supplementary Movie S2. Animation of chromosomal breakage and reconstitution of BSID43. Two independent karyotypic analyses indicated a balanced reciprocal translocation between chromosomes 3q and 5q; however, sequencing revealed the shattering of chromatin from 7q and re-integration of 7q DNA shards into the junction fragments of both derivative chromosomes, resulting in no direct 3q–5q junctions.

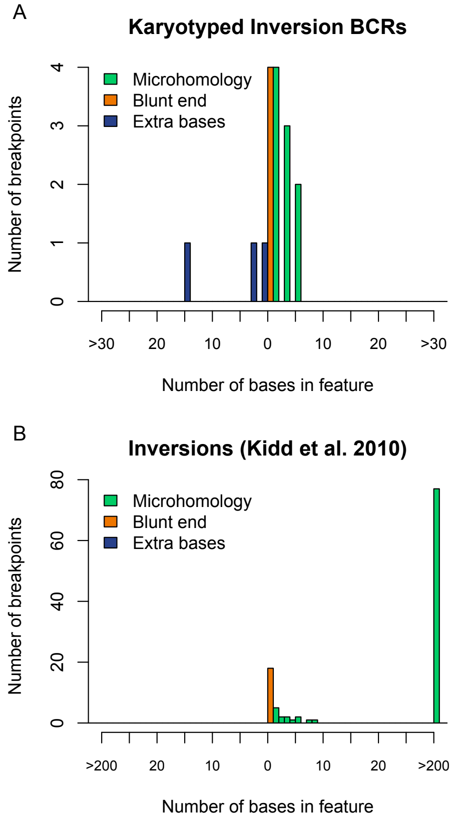
Supplementary Figure S1. Chromosomal rearrangements of two transgenic sheep



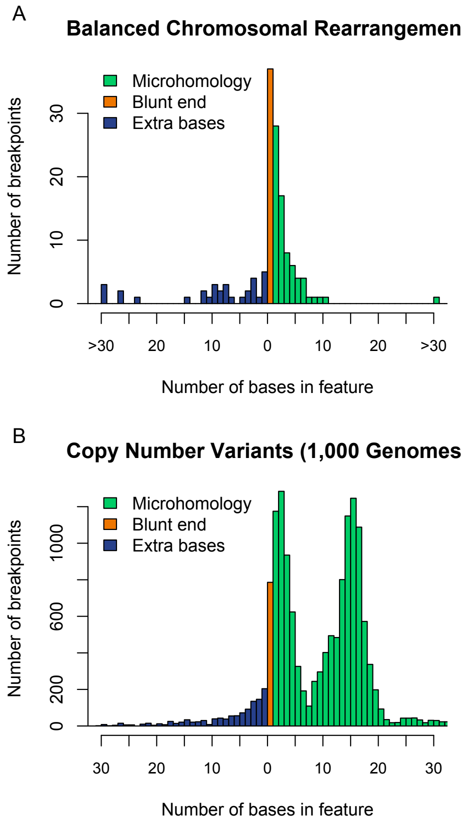
Supplementary Figure S2. Microhomology distributions of 141 chromosomal rearrangement breakpoints from two independent algorithms



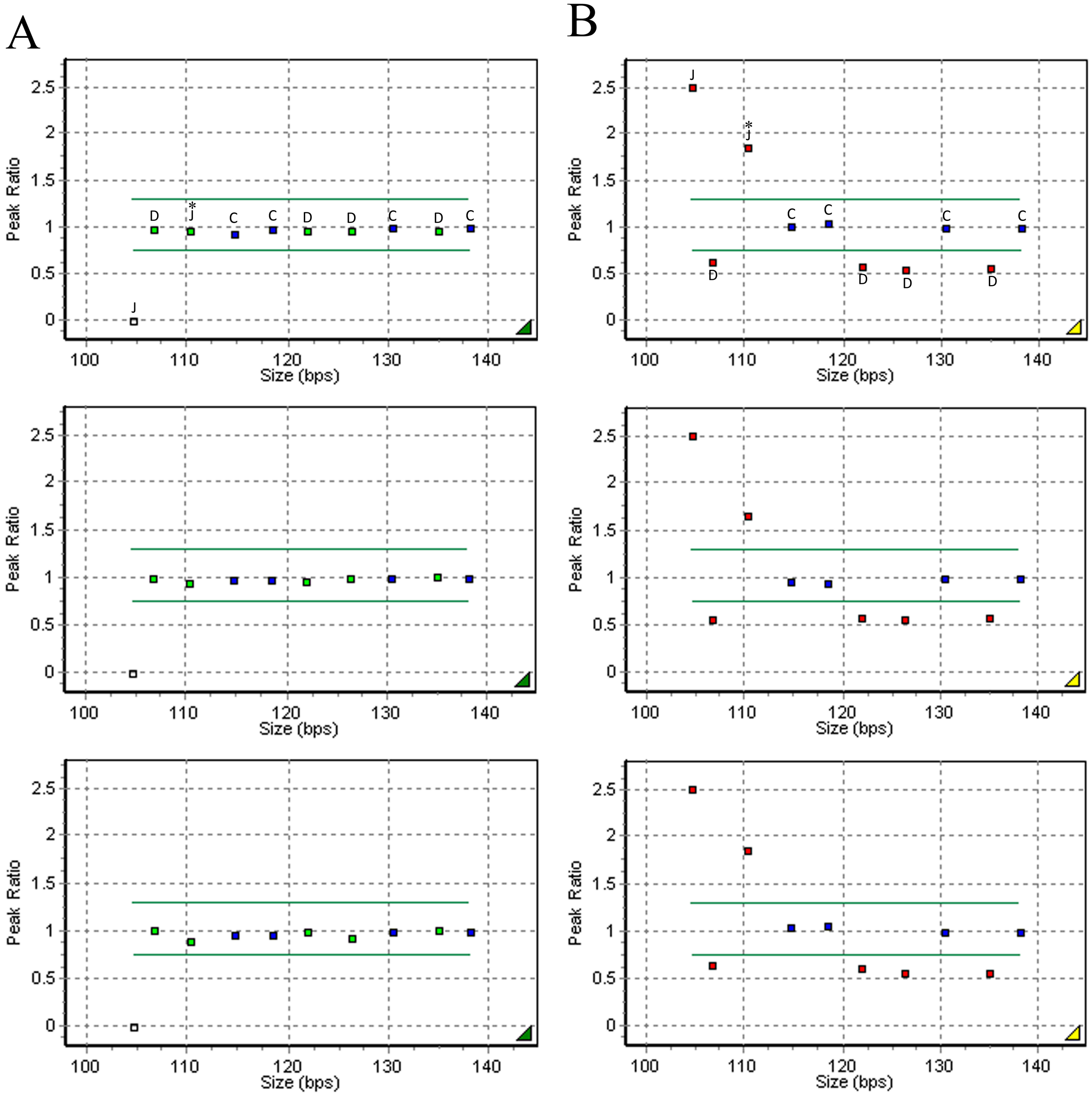
Supplementary Figure S3. Microhomology profile comparison of balanced inversions identified by karyotype to inversions in a population-based study



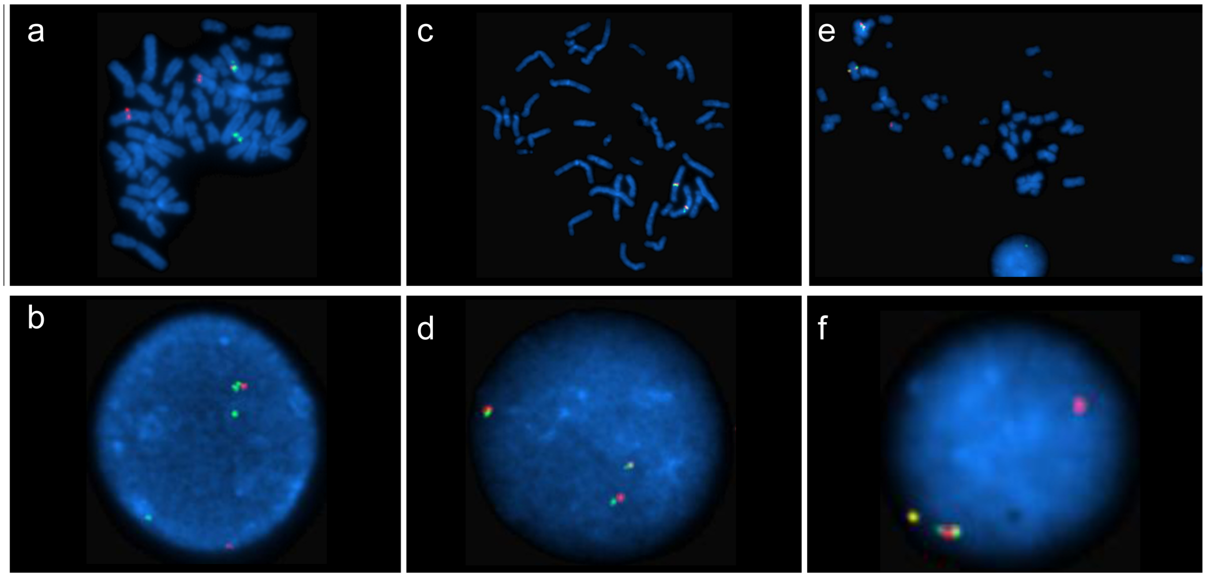
Supplementary Figure S4. Microhomology comparison of BCRs in this study to CNV breakpoints from the 1,000 Genomes Pilot 1 data



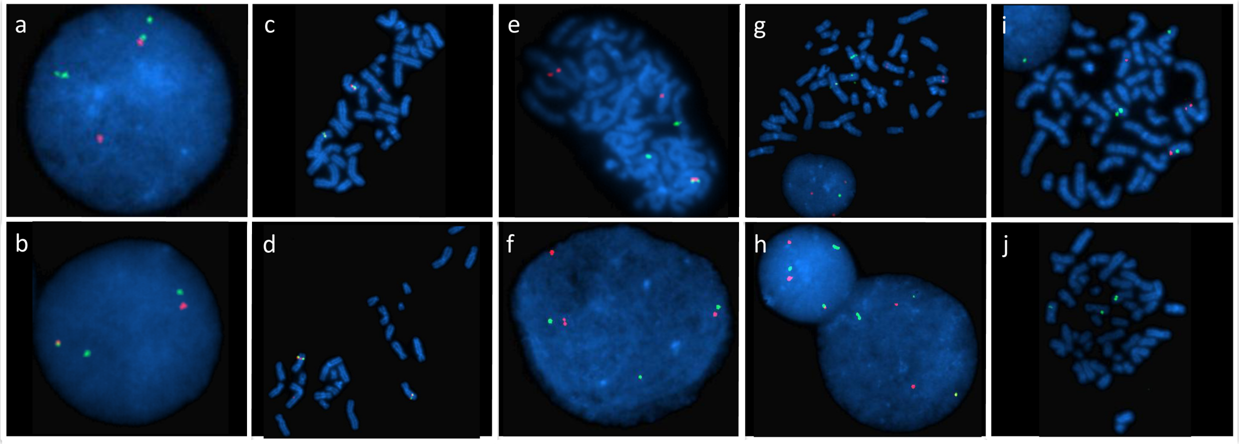
Supplementary Figure S5. Multiplex ligation-dependent probe amplification (MLPA) deletion assessment of mouse chromosome 4 in R6/2 and non-carrier mouse cortex



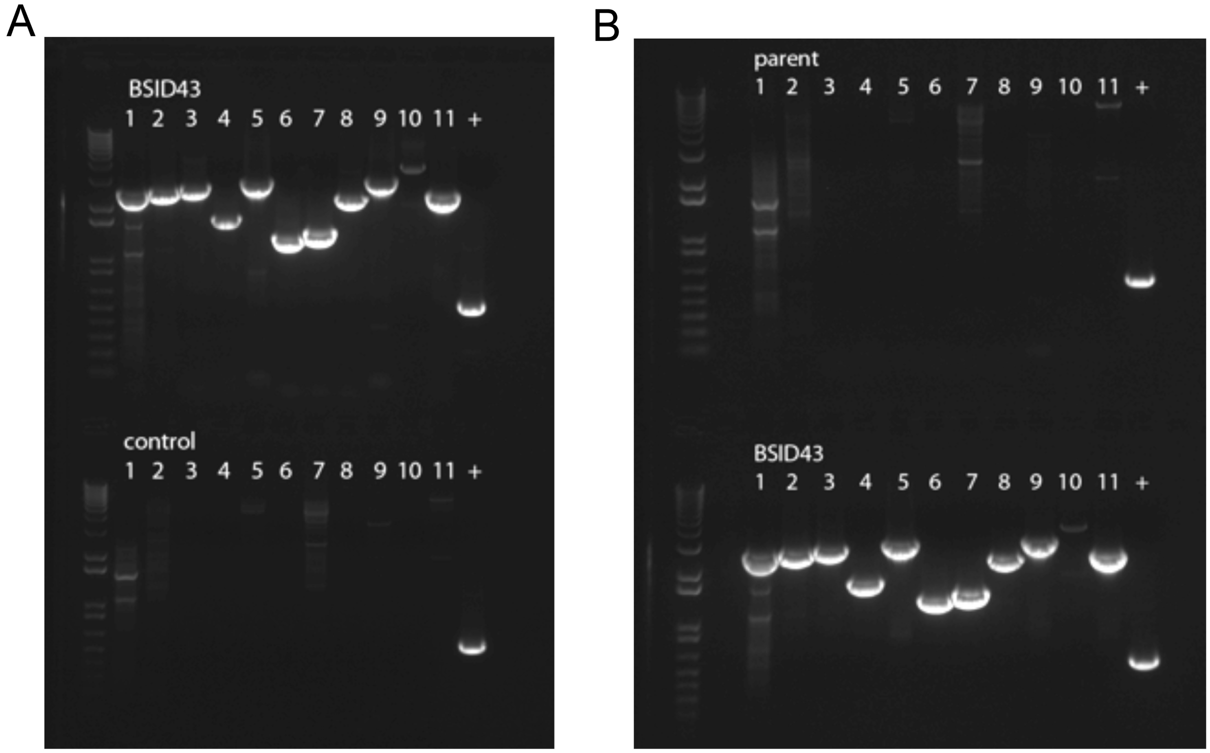
Supplementary Figure S6. FISH analyses for BSID42 chromothripsis



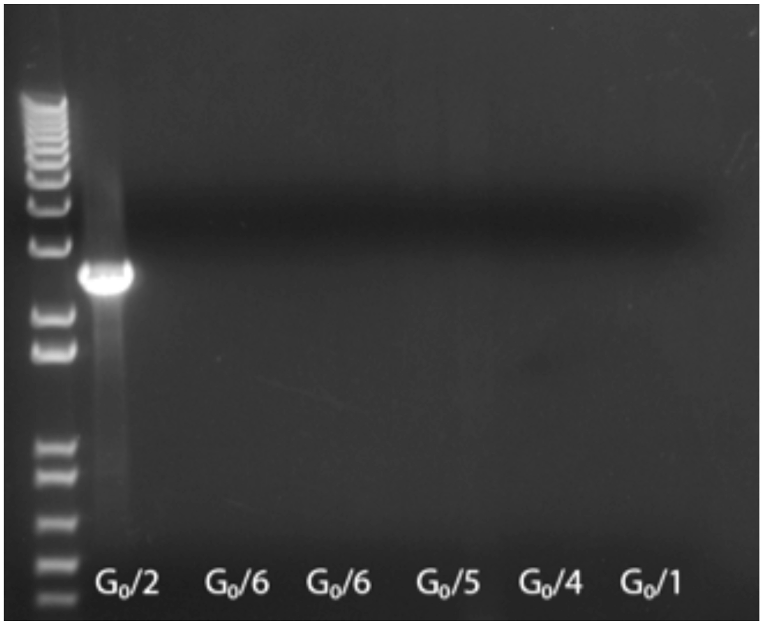
Supplementary Figure S7. FISH analyses for BSID43 chromothripsis



Supplementary Figure S8. Breakpoint amplification of each of the breakpoints of BSID43 compared to an available parent and a healthy control



Supplementary Figure S9. Breakpoint amplification of an 81 kb inversion breakpoint distal to the transgene integration site in transgenic sheep G0/2



Supplementary Table S1. Breakpoint delineation of all junction fragments.

Subject	Type	Chr	Pos	Chr	Pos	Str	Str	Microhom	Inserted Seq
BSID01	simple	chr17	68,247,919	chr3	54,949,502	-	+	0	0
	simple	chr3	54,949,512	chr17	68,247,917	+	-	5	5
BSID02	simple	chr7	47,368,688	chr8	38,292,459	-	-	0	0
	simple	chr8	38,292,460	chr7	47,368,687	-	-	0	0
BSID03	simple	chr7	88,352,066	chr12	97,860,653	+	+	2	0
	simple	chr12	97,860,651	chr7	88,352,067	+	+	-17	12
BSID04	simple	chr9	21,845,456	chr8	60,895,267	-	+	-27	27
	simple	chr8	60,895,263	chr9	21,845,448	+	-	-6	0
BSID05	simple	chrX	16,848,017	chr9	33,697,179	+	+	-2	0
	simple	chr9	33,697,171	chrX	16,848,016	+	+	0	0
BSID06	simple	chr5	138,665,752	chr1	27,926,596	-	+	2	0
	simple	chr1	27,926,596	chr5	138,665,752	-	+	2	0
BSID07	simple	chr3	77,220,642	chrY	4,892,527	+	+	0	0
	simple	chrY	4,892,525	chr3	77,220,640	+	+	0	0
BSID08	simple	chr2	203,082,134	chr8	143,062,504	+	+	-27	27
	simple	chr8	143,062,505	chr2	203,082,143	+	+	1	0
BSID09	complex	chr6	97,770,850	chr9	80,040,925	+	+	0	0
	complex	chr9	80,040,925	chr6	97,948,212	+	-	0	0
	complex	chr6	97,770,861	chr6	97,948,213	-	+	0	0
BSID10	simple	chr3	175,892,512	chr6	100,729,424	+	+	0	0
	simple	chr6	100,729,382	chr3	175,892,544	+	+	0	0
BSID11	simple	chr1	59,687,981	chr6	117,497,373	-	+	0	0
	simple	chr6	117,497,372	chr1	59,687,980	+	-	0	0
BSID12	simple	chr5	29,658,440	chr10	67,539,995	-	+	3	0
	simple	chr10	67,539,990	chr5	29,658,426	+	-	1	0
BSID13	simple	chr1	86,157,131	chr5	88,829,564	-	+	0	0
	simple	chr5	88,829,562	chr1	86,157,132	+	-	-10	10
BSID14	simple	chr5	69,973,529	chr7	94,979,567	+	-	0	0
	simple	chr5	69,973,530	chr7	94,979,568	-	+	1	0
BSID15	simple	chr22	45,671,560	chr2	149,034,432	+	+	1	1
	simple	chr2	149,034,431	chr22	45,671,558	+	+	1	0
BSID16	simple	chrX	39,741,846	chr11	126,910,404	-	+	-11	11
	simple	chr11	126,910,404	chrX	39,741,845	+	-	-3	3
BSID17	simple	chr11	46,619,320	chr9	140,661,623	+	-	-5	5
	simple	chr9	140,661,654	chr11	46,619,326	-	+	2	0
BSID18	simple	chr2	39,206,240	chr14	31,717,834	-	-	-1	1
	simple	chr2	39,206,242	chr14	31,717,833	+	+	1	0
BSID19	complex	chr6	86,488,291	chr6	85,900,540	+	-	1	0
	complex	chr6	85,897,899	chr6	93,909,993	-	+	0	0
	complex	chr6	85,897,870	chr13	80,659,609	+	+	-1	1
	complex	chr6	85,900,543	chr13	80,659,606	-	-	0	0
BSID20	simple	chr7	94,007,567	chr5	88,706,887	+	+	3	0
	simple	chr5	88,706,882	chr7	94,007,560	+	+	3	0
BSID21	complex	chr2	181,120,915	chr6	144,917,096	+	-	6	0
	complex	chr6	141,514,117	chr2	181,120,972	+	+	1	0
	complex	chr2	186,920,618	chr6	160,668,627	+	+	-1	1
	complex	chr6	160,668,625	chr2	186,920,668	+	+	0	0
BSID22[†]	complex	chr6	102,933,427	chr2	200,268,602	+	-	8	0
	complex	chr2	200,277,100	chr6	106,681,471	-	-	0	0

	complex	chr2	200,306,110	chr6	107,030,770	-	-	2	0
	complex	chr6	107,170,881	chr6	107,168,553	-	+	656	0
	complex	chr6	107,030,770	chr6	102,554,425	-	+	-1	1
	complex	chr6	106,704,624	chr6	102,933,469	-	+	1	0
BSID23§	simple	chr5	24,272,302	chr5	88,400,843	+	-	1	0
	simple	chr5	88,400,848	chr5	24,272,298	-	+	-15	15
BSID24	simple	chr3	156,276,251	chr14	21,899,864	+	+	1	0
	simple	chr14	21,899,864	chr3	156,276,247	+	+	2	0
BSID25	simple	chr3	118,271,920	chr18	52,909,158	+	+	-8	8
	simple	chr18	52,909,152	chr3	118,271,923	+	+	1	0
BSID26*	simple	chr6	57,575,919	chr5	21,573,437	+	-	2	0
BSID27§	simple	chr3	189,669,179	chr3	111,406,164	+	-	0	0
	simple	chr3	111,406,165	chr3	189,669,189	-	+	-1	1
BSID28	simple	chr11	45,965,069	chr19	8,030,126	-	-	1	0
	simple	chr11	45,965,064	chr19	8,030,133	+	+	0	0
BSID29††	simple	chr6	146,998,514	chr11	98,571,927	+	+	1	0
	simple	chr11	97,119,538	chr6	146,998,534	+	+	1	0
BSID30	simple	chr11	15,825,269	chr2	8,247,757	+	+	3	0
	simple	chr2	8,247,756	chr11	15,825,273	+	+	0	0
BSID31	simple	chr18	13,529,461	chr17	5,456,019	+	+	2	0
	simple	chr17	5,456,021	chr18	13,529,462	+	+	2	0
BSID32	complex	chr2	186,033,110	chr11	87,665,403	+	+	0	0
	complex	chr11	87,665,449	chr2	186,039,460	+	-	-3	3
	complex	chr2	186,039,111	chr2	186,033,142	-	+	8	0
BSID33§	simple	chr2	171,827,243	chr2	32,310,440	+	-	5	0
	simple	chr2	171,827,243	chr2	32,310,711	-	+	0	0
BSID34	simple	chr14	43,931,475	chr10	11,411,948	+	-	4	0
	simple	chr10	14,750,351	chr14	43,931,475	-	+	1	0
BSID35§	simple	chr7	157,577,842	chr7	69,685,967	+	-	0	0
	simple	chr7	69,685,977	chr7	157,577,842	-	+	-3	3
BSID36†	simple	chr15	25,415,771	chr9	41,418,249	-	+	2	0
	simple	chr9	41,418,231	chr15	25,495,990	+	-	-4	4
BSID37§	simple	chr12	82,319,028	chr12	13,955,818	+	-	3	0
	simple	chr12	13,955,816	chr12	82,319,026	-	+	3	0
BSID38	simple	chr2	148,852,132	chr10	117,238,824	-	-	-7	7
	simple	chr2	148,659,854	chr10	117,345,237	+	+	-8	8
BSID39†§	complex	chr6	166,069,041	chr6	116,841,827	+	-	0	0
	complex	chr6	166,069,041	chr6	122,333,156	-	+	3	0
	complex	chr6	130,848,186	chr6	130,852,295	+	-	1	0
	complex	chr6	130,848,186	chr6	130,852,295	-	+	1	0
BSID40	simple	chrX	18,456,504	chr19	6,384,984	-	-	1	0
	simple	chrX	18,456,503	chr19	6,384,980	+	+	2	0
BSID41†**	complex	chr17	2,081,621	chrX	123,461,047	-	-	-67	67
	complex	chrX	123,295,172	chrX	105,764,010	+	-	-10	10
	complex	chrX	105,521,517	chr17	123,376,856	+	+	-90	90
	complex	chrX	122,982,217	chr17	3,823,146	-	+	-10	10
	complex	chr17	3,552,677	chr2	61,889,744	+	+	-125	125
BSID42	thripsis	chrX	110,368,673	chr5	90,466,028	+	+	7	0
	thripsis	chr5	90,502,468	chrX	130,567,294	+	+	2	0
	thripsis	chr5	90,711,091	chr5	90,892,029	-	+	2	0
	thripsis	chr5	90,892,144	chrX	131,442,883	+	+	1	0
	thripsis	chrX	110,368,663	chr5	90,719,199	-	+	0	0
	thripsis	chrX	131,438,029	chr5	92,483,810	+	+	4	0

	thripsis	chr5	90,494,025	chr5	90,892,027	+	-	0	0
	thripsis	chr5	90,494,024	chr5	90,502,468	-	+	1	0
	thripsis	chr5	90,691,787	chr5	90,886,192	+	-	2	0
	thripsis	chr5	92,932,823	chr5	90,693,163	-	+	0	0
	thripsis	chr5	92,483,805	chr5	90,719,148	+	-	0	0
	thripsis	chr5	90,892,142	chr5	90,466,010	-	-	2	0
	thripsis	chr5	90,886,193	chr5	92,932,823	-	-	1	0
	thripsis	chrX	130,567,294	chr5	90,711,016	+	-	3	0
BSID43	thripsis	chr3	155,627,600	chr7	120,004,011	+	+	-12	12
	thripsis	chr7	119,889,673	chr3	155,627,601	+	+	0	0
	thripsis	chr3	152,159,406	chr7	119,999,922	+	+	6	0
	thripsis	chr7	120,765,008	chr3	152,159,406	+	+	-24	24
	thripsis	chr7	119,646,830	chr7	120,254,239	+	-	0	0
	thripsis	chr7	119,889,690	chr7	119,646,829	-	+	2	0
	thripsis	chr7	119,995,774	chr7	119,999,919	+	-	0	0
	thripsis	chr7	122,606,556	chr7	120,254,240	-	+	0	0
	thripsis	chr5	87,073,891	chr7	119,995,778	+	+	-7	2
	thripsis	chr7	122,606,486	chr5	87,073,908	+	+	0	0
	thripsis	chr7	120,004,010	chr7	120,766,450	+	+	0	0
BSID44	simple	chr7	128,114,028	chr19	32,861,768	+	+	1	0
	simple	chr19	29,247,464	chr7	128,114,023	+	+	6	0
BSID45	complex	chr4	151,513,246	chr4	156,815,165	+	+	1	0
	complex	chr4	138,022,086	chr4	156,815,165	-	-	-1	1
	complex	chr4	138,022,087	chr13	71,842,887	+	+	2	0
	complex	chr13	71,842,874	chr4	151,513,239	+	+	1	0
BSID46	simple	chr3	71,072,118	chr10	62,133,175	+	-	1	0
	simple	chr3	71,072,125	chr10	62,133,183	-	+	1	0
BSID47§	simple	chr3	114,735,958	chr3	80,706,876	+	-	0	0
	simple	chr3	80,706,874	chr3	114,735,959	-	+	1	0
BSID48	simple	chr10	107,714,387	chrX	51,707,815	-	+	-27	8
	simple	chrX	51,703,306	chr10	107,711,581	+	-	-9	9
BSID49	simple	chr9	102,425,452	chr16	26,393,002	+	-	0	0
	simple	chr16	26,393,002	chr9	102,425,452	-	+	0	0
BSID50	simple	chrX	36,118,091	chr18	31,709,058	+	-	-9	9
	simple	chr18	31,709,058	chrX	36,118,098	-	+	0	0
BSID51	simple	chr9	123,045,928	chr10	33,186,334	+	-	-3	3
	simple	chr9	123,045,929	chr10	33,186,326	-	+	1	0
BSID52	simple	chr11	35,077,326	chr12	23,711,192	+	+	4	0
	simple	chr12	23,711,182	chr11	35,077,343	+	+	-4	4

Positions and strand orientations of breakpoints from 52 subjects. All positions are hg19.
Type = rearrangement class, simple: exactly two breakpoints, complex: three or more breakpoints, thripsis: germline chromothripsis.
Chr = chromosome involved.
Pos = genomic position on the preceding chromosome.
Str = strand orientation for the two junction sequences.
Microhomology calculations based on EMBOSS procedure (Figure S2a).
Negative homology scores indicate inserted sequence, and zero homology indicates blunt end joining.
**Only a single derivative chromosome successfully amplified.*
***Possible chromothripsis subject but all junction fragments required to fully reconstruct the complete genomic reorganization could were not delineated.*
§*Karyotyped as an inversion.*

[†] Precise DNA imbalance (gain or loss of DNA) could not be unambiguously determined for at least one breakpoint junction, by example for a CCR in which all junction fragments were not detected, so the complete reorganization and associated imbalance remain ambiguous.

[††] Targeted capture was performed based on previous cytogenetic analyses narrowing the breakpoint region to less than a megabase, however the sequencing discovered one of the derivative breakpoints was outside the targeted region and the two breakpoints were separated by over 1.4 Mb. Analysis of read depth within the region did not indicate a deletion, and a follow-up aCGH experiment confirmed no genomic imbalance at array resolution in the region, suggesting this subject contains a CCR and the fragment has been integrated into a different genomic position, or some small genomic fragment integrated into the captured region from a locus 1.4 Mb telomeric to the breakpoint. However, we could not unambiguously confirm either possibility from the targeted capture and thus could not calculate the resultant genomic imbalance beyond array resolution.

Supplementary Table S2. Comparison of sequence motifs that could result in non-B DNA conformations flanking random and experimental rearrangement breakpoints

Flanking Bases	A-Phased	Direct	Gquad	Inverted	Mirror	Zdna
1,000,000 Randomly Simulated Breakpoints						
500bp	15.21%	11.33%	5.22%	14.16%	17.05%	3.85%
200bp	6.90%	5.27%	2.41%	7.01%	8.40%	1.69%
100bp	3.85%	3.13%	1.36%	4.46%	5.28%	0.95%
30bp	1.67%	1.56%	0.63%	2.63%	2.99%	0.40%
2bp	0.75%	0.88%	0.29%	1.81%	2.07%	0.17%
Experimental Breakpoint Sequencing						
500bp	9.52%	8.93%	5.36%	15.48%	17.26%	3.57%
200bp	2.98%	4.17%	2.98%	10.12%	8.93%	1.19%
100bp	2.38%	2.98%	2.38%	7.14%	4.17%	0.60%
30bp	2.38%	1.19%	1.79%	3.57%	2.98%	0.00%
2bp	0.60%	1.19%	0.60%	2.38%	1.19%	0.00%

One million randomly simulated breakpoints were compared to experimental balanced SV breakpoint sequencing. No statistically significant differences were detected between our experimental sequencing set and the randomly simulated breakpoints.
See (Bacolla and Wells, 2009; Wells, 2007 for details).
A-Phased = A-Phased repeats
Direct: Direct repeat (10-50 bases repeated within a 5 base spacer)
Gquad = G-Quadruplex forming repeats (4 or more G-tracts (3-7 G's) separated by 1-7 base spacers.
Inverted = inverted repeat (10-100 bases with reverse complelent within 100 bases, can result in cruciform formation).
Mirror = Mirror repeat (10-100 bases mirrored within 100 bases)
Zdna = Z-DNA repeat (G followed by C or T for at least 10 bases)

Supplementary Table S3. Sequencing metrics for two chromothripsis samples

ID	Aligned Reads	Median Insert	Ave Cov	Breaks	Chrs	Inverted Junctions	Intra Chr	Inter Chr	Total DNA Rearranged	Total DNA Lost
BSID42	84.0M	2,887.3	80.0X	14	2	8	8	6	23,541,033	6,357
BSID43*	260.9M	2,667.5	214.6X	11	3	4	5	6	6,427,936	1,551

Ave Cov = average genomic coverage of mapped inserts between paired reads.
Chrs = chromosomes involved in all rearrangements for each subject.
Intra Chr = number of intrachromosomal rearrangements.
Inter Chr = number of interchromosomal rearrangements.
Total DNA rearranged = total number of bases involved in chromosomal reorganization (i.e. displaced from the reference position) from the sum of all rearrangements.
Total DNA lost = total number of bases deleted from the sum of all rearrangements.
**Multiple libraries for BCID43 were created and sequenced independently using DNA extracted from whole blood as well as DNA extracted from EBV-transformed lymphoblastoid cell lines. All rearrangement junctions were identical between sequencing lanes and data combined here.*