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

Colby Chiang1,6, Jessie C Jacobsen1–3,16, Carl Ernst1–3,15,16, Carrie Hanscom1, Adrian Heilbut1,
Ian Blumenthal1, Ryan E Mills3, Andrew Kirby1,5, Amelia M Lindgren4, Skye R Rudiger6,7, Clive J McLaughlan6,7,
C Simon Bawden6,7, Suzanne J Reid8,9, Richard L M Faull8,10, Russell G Snell8,9, Ira M Hall11, Yiping Shen1,12,
Toshiro K Ohsumi13, Mark L Borowsky13, Mark J Daly1,5, Charles Lee4, Cynthia C Morton4,5,14, Marcy E MacDonald1–3,5,
James F Gusella1–3,5 & Michael E Talkowski1–3,5

We defined the genetic landscape of balanced chromosomal rearrangements at nucleotide resolution by sequencing 141 breakpoints from cytogenetically interpreted translocations and inversions. We confirm that the recently described phenomenon of ‘chromothripsis’ (massive chromosomal shattering and reorganization) is not unique to cancer cells but also occurs in the germline, where it can resolve to a relatively balanced state with frequent inversions. We detected a high incidence of complex rearrangements (19.2%) and substantially less reliance on microhomology (31%) than previously observed in benign copy-number variants (CNVs). We compared these results to experimentally generated DNA breakage-repair by sequencing seven transgenic animals, revealing extensive reorganization of the transgene and host genome with similar complexity to human germline alterations. Inversion was the most common rearrangement, suggesting that a combined mechanism involving template switching and non-homologous repair mediates the formation of balanced complex rearrangements that are viable, stably replicated and transmitted unaltered to subsequent generations.

The understanding of the genetic architecture of human chromosomal rearrangements has expanded in recent years, as rapid improvements in genomics technology have spawned a growing number of mechanistic hypotheses, many of which involve some degree of homology between participant sequences1–3. In complex events, replication-based mechanisms have been proposed to have a role, including template switching from a stalled or disrupted replication fork (fork stalling and template switching, FoSteS)4 and microhomology-mediated, break-induced replication (MMBR)3,5. While it has been established that chromosomal exchanges that seem to be balanced at lower resolution can actually involve considerable complexity that may contribute to human disease in unexpected ways6,7, few studies have assessed these events at base-pair resolution. The first massively parallel sequencing of cancer cells suggested a complex rearrangement landscape6, and, more recently, an unexpected phenomenon was uncovered in cancer cells that involved massive chromosomal shattering and reorganization with frequent changes to the copy-number state across the affected region7. The authors called this phenomenon ‘chromothripsis’, hypothesizing that it was a unique feature of somatic mutation that might occur in 2–3% of all cancers8, and similar complexity was seen in several subsequent sequencing studies9,10–12.

The mutational mechanism(s) underlying such complex genomic reorganization is unknown, but an intriguing feature of cancer-related chromothripsis is a frequent transition between two copy-number

1Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. 2Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. 3Department of Neurology, Harvard Medical School, Boston, Massachusetts, USA. 4Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. 5Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. 6Molecular Biology Laboratory, Livestock and Farming Systems Division, South Australian Research and Development Institute, Roseworthy, South Australia, Australia. 7Reproductive Technology Laboratory, Livestock and Farming Systems Division, South Australian Research and Development Institute, Roseworthy, South Australia, Australia. 8The Centre for Brain Research, Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand. 9School of Biological Sciences, Faculty of Science, The University of Auckland, Auckland, New Zealand. 10Department of Anatomy with Radiology, Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand. 11Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia, USA. 12Department of Laboratory Medicine, Children’s Hospital Boston, Boston, Massachusetts, USA. 13Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. 14Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women’s Hospital, Boston, Massachusetts, USA. 15Present address: Department of Psychiatry, McGill University, Montreal, Quebec, Canada. 16These authors contributed equally to this work. Correspondence should be addressed to M.E.T. (talkowski@chgr.mgh.harvard.edu).

Received 29 September 2011; accepted 17 January 2012; published online 4 March 2012; doi:10.1038/ng.2202
states throughout the impacted genomic region. CNVs have emerged as a major component of genetic variation in humans, and two recent population-based sequencing studies found a high degree of microhomology at CNV breakpoints and suggested two predominant CNV-generating mechanisms: microhomology-mediated end joining (MHEJ) and non-homologous end joining (NHEJ). Similar events that are postulated to occur by distinct pathways, these mechanistic hypotheses may be somewhat limited, as they were restricted to events defined by DNA dosage changes in unphenotyped individuals. The rearrangement landscape of other forms of genomic rearrangement, namely constitutional balanced structural variation, such as reciprocal translocation and inversion, has not been comprehensively assessed at the sequence level. Breakpoint resolution of these structural variations is fundamental to the prediction of which rearrangement mechanisms could underlie their formation and to the understanding of the full range of mutational mechanisms involved in human structural variation.

Here, we provide the first high-throughput sequence-based assessment of mutational mechanisms associated with breakpoints in 52 subjects with chromosomal abnormalities that were previously defined at cytogenetic resolution as balanced (45 reciprocal translocations and 7 inversions) and were clinically assessed as likely to be pathogenic (50 arose de novo, and 2 were inherited from an affected parent). We performed a series of next-generation sequencing experiments that included either whole-genome sequencing or targeted capture of breakpoints (see Online Methods and previous descriptions), and all breakpoints were confirmed at base-pair resolution by capillary sequencing. Our results show definitively that, as in cancer, remarkably complex genomic reorganization can also occur in the human germline, but the repair process can resolve to a relatively balanced state rather than yielding extensive gains and losses of DNA. When we surveyed an experimental system of chromosomal rearrangement (in transgenic animals), we found that similar genomic reorganization can result from experimentally generated double-stranded DNA breaks (DSBs), in the absence of environmental mutagenic factors and in the presence of an abundance of homologous template. In both the human and transgenic animal breakpoints studied, the mechanism(s) that mediated rearrangements did not depend primarily on microhomology, nor were they frequently associated with large DNA dosage changes; both findings are in stark contrast to previous studies of chromothripsis in cancer and benign CNV formation. Instead, these results reveal substantial chromosomal reorganization in the germline that can be compatible with viability and is stably replicated and transmitted to subsequent generations.

**RESULTS**

**Complex genomic reorganization in the human germline**

We found cytogenetically defined and seemingly balanced structural variations to be far more complex than originally thought, detecting 141 breakpoints in 52 subjects with chromosomal abnormalities, with an average of 2.71 breakpoints per individual. Only two subjects had two derivative chromosomes with no DNA imbalance; however, most subjects did not suffer a substantial loss of genetic material (arbitrarily defined here as >1 kb of total genomic imbalance; Table 1 and Supplementary Table 1). Instead, multiple breakpoints were often reassembled in a relatively balanced manner, with microdeletions of one to several bases. We found that 19.2% of all karyotypically balanced structural variations actually met the conventional criterion for a complex chromosomal rearrangement (CCR; three or more breakpoints), which is a significantly higher frequency than the previous estimate of 2.8% from 246 de novo balanced anomalies assessed in more than 269,000 prenatal diagnoses ($P = 2.2 \times 10^{-4}$). Nine of the CCRs involved at least one inverted segment, and only two CCRs involved exchanges between more than two chromosomes.

In the ten subjects with CCRs, there was extensive rearrangement complexity. In the two most intricate cases (BSID42 and BSID43), we observed shattering and reorganization of multiple chromosomal segments similar to chromothripsis in cancer, except that these germline events resolved to a largely dosage-balanced state rather than exhibiting frequent alterations in copy number (Figs. 1 and 2, Supplementary Fig. 1 and Supplementary Table 2). In BSID42, we observed 14 different junctions between the X chromosome and chromosome 5, with extensive shattering of localized regions (Fig. 2a and Supplementary Video 1). The DNA shards at 5q14.3 integrated into the X chromosome, with frequent oscillations between inverted and same-strand orientation (Fig. 2a). In both derivative chromosomes, large segments remained fully intact between reassembled DNA fragments. Upstream of the most distal telomeric 5q-Xq junction in the derivative X chromosome, an ~871-kb segment of chromosome X (designated Xc) was excised, inverted and inserted into the strand-oscillating segment from chromosome X approximately 20 Mb from its origin on the X chromosome. Similarly, an 8-kb

---

**Table 1** Breakpoint characteristics of karyotypically balanced structural variations

| Breakpoint characteristics | All       | Simple    | Complex   | Chromothripsis |
|----------------------------|-----------|-----------|-----------|----------------|
| By subject                 |           |           |           |                |
| Total N (%)                | 52 (100.0%) | 42 (80.8) | 10 (19.2) | 2 (3.8)        |
| Balanced                   | 3.8       | 4.8       | 0.0       | 0.0            |
| Contains inversion         | 28.8      | 14.3      | 90.0      | 100.0          |
| By breakpoint              |           |           |           |                |
| Count (%)                  | 141 (100) | 83 (58.9) | 58 (41.1) | 25 (17.7)      |
| Balanced                   | 18.9      | 20.0      | 17.0      | 20.0           |
| Inverted                   | 30.5      | 14.5      | 53.4      | 48.0           |
| Base deletions             |           |           |           |                |
| <2                         | 49.6      | 51.3      | 46.8      | 40.0           |
| 2–20                       | 33.1      | 36.3      | 27.7      | 36.0           |
| 21–100                     | 7.9       | 3.8       | 14.9      | 12.0           |
| 101–1,000                  | 0.8       | 1.3       | 0.0       | 0.0            |
| >1,000                     | 8.7       | 7.5       | 10.6      | 12.0           |
| Insertions                 |           |           |           |                |
| <2                         | 80.1      | 77.1      | 84.5      | 88.0           |
| 2–20                       | 15.6      | 20.5      | 8.6       | 8.0            |
| 21–100                     | 3.5       | 2.4       | 5.2       | 4.0            |
| 101–1,000                  | 0.7       | 0.0       | 1.7       | 0.0            |
| >1,000                     | 0.0       | 0.0       | 0.0       | 0.0            |
| Microhomology              |           |           |           |                |
| <2                         | 68.8      | 68.7      | 69.0      | 60.0           |
| 2–20                       | 30.5      | 31.3      | 29.3      | 40.0           |
| >20b                       | 0.7       | 0.0       | 1.7       | 0.0            |

Percentages are derived based on the total number of events (either by subject, N = 52, or by breakpoint, N = 141). Results are presented from analyses using EMBOSS needle for direct comparison with previous findings. Data on base deletions are restricted to 127 breakpoints for which breakpoint pairs could be unambiguously determined such that the precise loss of DNA could be calculated.

Six of the seven karyotypically defined inversions involved just two breakpoints; one had multiple inversion breakpoints. A single inverted breakpoint junction contained 654 bp of homology. No other junctions had more than 20 bp of homology.
segment of chromosome 5 (designated 5f) was excised, inverted and reinserted in the derivative chromosome 5. Sequencing thus delineated a composite wherein shattered exchanges between chromosomes also resulted in extensive intrachromosomal reorganization. Notably, both the clinical karyotype and a 1-million feature array-comparative genomic hybridization (1M aCGH) analysis failed to detect this series of rearrangements that extensively reorganized more than 23.5 Mb of DNA (0.76% of the entire genome for this structural variation; Supplementary Table 2), as the sum of DNA imbalance from all junction fragments was just 6,357 bases.

The second subject, BSID43, had DNA available from two sources, one extracted from whole blood and another from an Epstein-Barr virus (EBV)-transformed lymphoblastoid cell line, both of which were sequenced (214.6x average physical coverage of mapped inserts; Supplementary Table 2). The fully resolved genomic reorganization involved 11 breakpoints resulting from the apparent shattering of a 3-Mb region at 7q31.31-3q31.32 and integration of fragments into the long arms of both chromosomes 3 and 5. The chromosome 3 breakpoints spanned 3.5 Mb (3q25.2-q25.31), whereas chromosome 5 had just a single breakpoint on each derivative (5q14.3), with the loss of only 16 bases between derivatives (Fig. 2b, Supplementary Figs. 2 and 3 and Supplementary Video 2). Despite the karyotype-based identification of a reciprocal translocation of chromosomes 3 and 5, there were no actual junctions connecting 3q25.2-25.31 to 5q14.3. Rather, all breakpoint junctions detected involved joining of shattered fragments from 7q31.31-q31.32, and all events were again cryptic to both karyotype and 1M aCGH diagnostics (a total of only 1,551 bases was lost from all rearrangements) (Fig. 2b). All breakpoints were identical between the two DNA sources, indicating that, once established, the rearranged chromosomal organization was stable throughout EBV transformation and subsequent cell propagation.

Eight additional subjects had CCRs with from 3–6 breakpoints each, many of which involved less than 1 kb of total genomic imbalance. All but one had inverted segments associated with a breakpoint, usually directly at the breakpoint junction, with excision, inversion and insertion events representing a common signature (Supplementary Tables 1 and 2). The only CCR event that did not have an inversion associated with a breakpoint (BSID45) contained a balanced excision of 13.5 Mb of chromosome 4, which was reinserted 5.3 Mb away from the translocation breakpoint in the derivative chromosome 13 (Supplementary Tables 1 and 2). Of the seven karyotypically defined inversions, six were simple and largely balanced events, and only one event (BSID39) was complex, containing a smaller 4.1-kb inversion nested within a larger 43.7-Mb inversion (Supplementary Tables 1 and 2).

Complex rearrangements from experimental transgenesis

Each of the previous examples of cancer chromothripsis postulated that exogenous factors, such as ionizing radiation, reactive oxygen species or chemotherapy, could prime such rearrangements by inducing multiple DSBs (reviewed in ref. 20). One previous study compared DSBR repair in the germline and following experimentally generated aphidicolin-induced CNV formation, finding that non-homologous repair was the driving mechanism in both scenarios21. The proposal that the complex reorganization of human and cancer genomes operates through aberrant repair of multiple apparently simultaneous DNA breaks prompted us to test whether a similar outcome could be experimentally induced through transgene integration, in which exogenous DNA fragments, which are themselves substrates for DSBR repair, are incorporated into the genome. We hypothesized that the process of transgene integration might provide insight into the general mechanisms involved in complex rearrangements. Notably, few such transgene integration sites have been characterized at the sequence level22. Using two sequencing approaches (targeted capture of transgenes and whole-genome jumping libraries), we analyzed seven transgenic animals: two well-established mouse lines produced by injection with a 1.9-kb fragment of human genomic DNA from 4p16.3 containing exon 1 of HTT (the gene for Huntington’s disease) (R6/1 and R6/223) and five sheep lines transgenic for HTT that were created by injection with an 11.6-kb full-length HTT cDNA (G2/1, G2/2, G2/5, G2/4 and G2/6)24 (see Online Methods and the Supplementary Note).

Transgene integration sites were identified to base-pair resolution in all seven animal lines, revealing multiple complex rearrangements
in both the internal structure of the final transgene (deletions, duplications, inversions and excision events followed by inversion and insertion) and in the host genomic DNA. Inverted segments within the transgene were a common feature, showing frequent oscillations in strand orientation at junction fragments, inverted insertions of many fragments and an apparent induction of intrachromosomal inverted excision and insertion events in both derivative chromosomes. (b) Two independent karyotype-based analyses indicated a balanced reciprocal translocation between chromosomes 3q and 5q; however, sequencing revealed the shattering of chromatin from 7q and reintegration of 7q shards into the junction fragments of both derivative chromosomes (der(3), der(5)), resulting in no direct 3q-5q junction. There were four different inverted excision and insertion events, including intrachromosomal excision, inversion and reinsertion at 7q with a co-occurring inverted insertion of an intact 3.5-Mb segment of 3q, all of which involved only 1,551 bases of total DNA imbalance. All positions are based on hg19. (See also Supplementary Videos 1 and 2.)
Figure 4  Breakpoint sequence signatures from balanced structural variations and copy-number variation from independent population-based studies. Histograms of nucleotide distribution for each sample. Orange, breakpoints with no homology or inserted sequences (blunt ligation); green bars, sequence microhomology between breakpoints; blue bars, the number of inserted nucleotides at the breakpoints. (a) Breakpoint sequence distribution from karyotypically balanced and presumably pathogenic rearrangements sequenced in this study. (b) Sequence distribution of 545 breakpoints from six tumors (two chronic lymphocytic leukemias (CLLs), one colorectal cancer, one thyroid cancer, one renal cancer and one small cell lung cancer) localized by paired-end sequencing and confirmed with PCR as described5. (c) Sequence distribution of 315 deletions captured and sequenced in control individuals using a previously described protocol14. (d) Breakpoint signatures of 16,783 deletion breakpoints from the 1,000 Genomes Project Pilot 1 analyzed with the same pipeline as the karyotypically balanced rearrangements in a, showing an identical distribution of breakpoint sequences to those seen previously15.

Mechanistic signatures from rearrangement breakpoints
We discovered a considerable disparity between breakpoint features of de novo and presumably pathogenic translocations or inversions and those observed in population-based CNV studies, with important mechanistic implications. Only 30.5% of the 141 de novo breakpoints we identified were flanked by microhomology, suggesting that most events were unlikely to have arisen by MHEJ, homology-mediated repair or microhomology-mediated replication-based mechanisms, such as MMBIR15 (Fig. 4a). This microhomology distribution resembled the profile of six tumors that had undergone chromothripsis (40.5% of breakpoints contained microhomology; Fig. 4b)9. In contrast, recent studies have shown that 70–80% of all CNVs in unphenotyped individuals contained microhomology14,15 (Fig. 4c,d).

The previous CNV estimates were almost exclusively based on the identification of deletions, as these events were most readily detected by previous CNV capture14 and low-depth sequencing15 analyses; neither study evaluated balanced events. Another study surveyed structural variations at fosmid resolution and suggested that non-allelic homologous recombination (NAHR) was the dominant mechanism mediating large inversion formation in population-based samples (69% had flanking homologous segments of >200 bp)25. In our study of potentially pathogenic balanced structural variations, we found that only 1 of the 43 total inverted segments contained greater than 200 bp of homology (654 bp) (Supplementary Fig. 7). This one instance was not a karyotypically defined large inversion but rather a small 2.3-kb inversion associated with a CCR (Supplementary Table 1). Collectively, the microhomology profile of the karyotypically defined inversions mirrored the breakpoints of the translocations, 69% of which contained <2 bp of microhomology. The disparity between our findings and the previously analyzed variants25 was not biased by sequencing methodology (NAHR is a notoriously difficult mechanism to delineate by next-generation sequencing), as we localized to base-pair resolution each of the seven karyotypically defined de novo inversions defined herein. None of these karyotypically defined and presumably pathogenic inversions contained long stretches of homology at the breakpoint.

We surveyed breakpoint profiles in the transgenic animals with experimentally induced DSB repair and found compelling evidence for mechanisms more similar to those involved in de novo balanced structural variations than in benign CNVs, as 78% of the host genomic DNA integration sites contained no microhomology. To test the sensitivity of the paired-end sequencing results, we performed capillary sequencing validation of the entire transgene sequence in two animals (R6/2 and Gp/1, excluding 121 bp of a hairpin formation in Gp/1), which confirmed all of the rearrangement junctions detected by paired-end sequencing that are shown in Figure 3. We also found that 76.9% of the internal transgene junctions had less than 2 bp of homology, despite the presence of abundant homologous template for homology-mediated mechanisms.

We further explored this marked difference in the frequency of microhomology between de novo balanced structural variations tested here and CNVs in previous studies. We initially looked for methodological-based effects in our data from karyotypically balanced structural variations by using three independent methods to assess microhomology, including (i) high-throughput alignment using Burrows-Wheeler Aligner–Smith-Waterman (BWA-SW; Supplementary Fig. 8), with which 26.2% of breakpoints contained two or more bases of microhomology, (ii) the European Molecular Biology Open Software Suite (EMBOSS) needle program25 that allows concurrent inserted template sequences with microhomology, with which we identified...
microhomology in seven additional individuals (31.2% total), an overall number that was not statistically different from results with the BWA-SW procedure \((P = 0.43)\), and (iii) a previously used revised version of BreakSeq\(^3,5,26\) that we customized to analyze balanced structural variations, which gave nearly identical results to those from the BWA-SW and EMBOSS methods (R.M., unpublished data; see the Supplementary Note). We therefore considered both the final breakpoint junctions and the sequence features of the initially intact chromosomes before their disruption in our analyses, with consistent results from the three methods. When we compared these data to the breakpoint homology from 16,783 CNV breakpoints assembled for the 1000 Genomes Project Pilot 1 study that were analyzed using identical methods, we found highly significant differences in sequence microhomology with both the BWA-SW analysis \((\chi^2 = 244.6, 1 \text{ degree of freedom}; P = 2.6 \times 10^{-35})\) (Supplementary Fig. 9) and the EMBOSS method \((\chi^2 = 201.6, 1 \text{ degree of freedom}; P = 1.26 \times 10^{-45})\) (Fig. 3). The EMBOSS results are reported here, as they represent the most conservative interpretation of the distinction between previously published population-based studies and our results in karyotypically defined balanced rearrangements.

This persistent deficiency in microhomology in microbes with de novo structural variations led us to question the extent to which any mechanism other than random NHEJ contributed to formation of chromosomal rearrangements in these individuals. We generated 1,000 simulated chimeric sequences of random breakpoints in the genome and analyzed the microhomology at these random junctions relative to the breakpoints in our subjects. The overall distributions of microhomology were significantly different \((\chi^2 = 24.5, 2 \text{ degrees of freedom}; P = 4.88 \times 10^{-6})\). When we examined these data, we found close concordance between the experimental set and random breakpoints for exactly 2 bp of homology \((12.1\% \text{ compared to } 10.7\%, \text{ respectively}; P = 0.66)\). Of the remaining breakpoints (having \(<2 \text{ bp or }>2 \text{ bp of homology})\), there was a significant enrichment of microhomology in the experimental set compared to the random breakpoints \((\text{experimental versus simulated for } <2 \text{ bp, } 68.8\% \text{ versus } 82.3\%; P = 3.6 \times 10^{-4} \text{ and for } >2 \text{ bp, } 19.1\% \text{ versus } 7.0\%; P = 1.38 \times 10^{-5})\) (Fig. 5a), indicating nonrandom reliance on a microhomology-mediated repair mechanism in the formation of these balanced structural variations. In simulations of 10,000 datasets of equal size to our experimental set, we found a marginal enrichment of long interspersed elements (LINEs) at the breakpoints \((\text{empirical } P = 0.085)\) but not short interspersed elements (SINEs) (Fig. 5b,c), and with 1 million simulations of windows ranging from 2–500 bp surrounding random genomic breakpoints, we found no significant enrichment in the frequency of repetitive DNA motifs that could fold into non-B DNA structure\(^{27,28}\) (Supplementary Table 3).

**DISCUSSION**

This sequence-level evaluation of the genetic architecture of karyotypically balanced, potentially pathogenic chromosomal rearrangements reveals several notable and underappreciated features of such events. (i) Karyotypically balanced structural variations are rarely truly balanced at the primary sequence level. (ii) Human germline rearrangements can be extraordinarily complex, including localized deconstruction of chromosomal segments into many small fragments that can be rejoined and resolved in an aberrant but largely dosage-balanced manner that is compatible with life. (iii) Local inversions and what we refer to as ‘inverted translocations’ (inversions associated with translocation breakpoints) are a commonly observed signature of CCRs. (iv) The breakpoint characteristics of potentially pathogenic karyotypically balanced structural variations are markedly different from those of seemingly benign CNVs. These studies also provide the first high-throughput sequencing assessment of the complexity resulting from the integration of linear exogenous DNA in transgenic animals, revealing resolved breakpoints that in many ways resemble those of the human germline and cancer chromothripsis\(^9,11,12\). These similarities suggest that the mechanisms involved in establishing transgene integration in model organisms may overlap with those that produce complex genomic DNA rearrangement in humans.

Cytogenetic studies previously established that a small subset of translocations has multiple chromosomal breakpoints rather than simple reciprocal exchanges between chromosomes\(^6,29\). By sequencing karyotypically balanced structural variations, we find a substantially greater proportion of CCRs than previously estimated. Sequence-level resolution enabled us to expand the definition of CCR to include local inversions at translocation breakpoints (inverted translocations), which seem to be an underappreciated feature of CCRs. These rearrangements were one of the most pervasive characteristics of germline chromothripsis, with frequent switching of strand orientation occurring between integrated fragments. This is consistent with cancer chromothripsis\(^9\), as well as with the finding that 4.5% of sequenced CNVs contained inserted sequence from the opposite strand in close proximity to the breakpoint\(^{14}\) and the identification of small insertions at breakpoints of constitutional CNVs\(^{21}\). We further observed that inverted excision and insertion events in CCRs were frequently accompanied by additional genomic rearrangements that can be interspersed with large, fully intact segments, suggesting an association between local microinversion and aberrant repair of DNA breaks.

Our findings confirm and further characterize the phenomenon of chromothripsis resulting from highly aberrant DSB repair; however, they foster further uncertainty about the cause of the initial damaging event and the mechanism of repair. Our results strengthen the previously described conclusion\(^9\) that these events occur in a one-off
chromosomal catastrophe rather than by progressive accumulation, as we find the same phenomenon in the human germline and in the first generation of transgenic sheep. We also confirm in humans and transgenic animals that, once resolved, the events can be viable, allowing them to be replicated during mitosis with high fidelity and stably transmitted to subsequent generations. The cause of chromothripsis thus does not seem to be specific to aberrant repair pathways in cancer, nor is it likely to be driven by other suggested mechanisms, such as disruption of tumor suppressor genes, formation of double minutes or telomere disruption from end-to-end chromosome fusion\(^9\). Although cancer and germline chromothripsis seem to differ in the presence of extensive copy-number change, dosage alterations may be favored in cancer cells due to dysregulation of growth pathways via loss of tumor suppressor genes. Frequent dosage alterations likely also occur in the germline, as one recent study found an unexpectedly high number of segmental imbalances in cleavage stage embryos (~70%)\(^{30}\). At least a portion of these complex events are viable, as evidenced in a study published during the review of this manuscript that selected for human cases with multiple copy-number imbalances on a single chromosome\(^{24}\). It is possible that the co-occurrence of dosage alteration with substantial interchromosomal exchanges results in a non-viable embryo or fetus, eliminating such cases before birth. Thus, the relatively balanced dosage of individuals with CCRs may simply reflect a selection for viability. With the increasing application of aCGH in research and clinical diagnostics, additional human cases with extensive copy-number alterations will undoubtedly emerge, and it will be interesting to determine whether they have the high frequency of concomitant chromosomal reorganization seen here and in cancer chromothripsis.

It is unknown if the NAHR-dominated mechanism of large inversions observed previously by fosmid sequencing\(^{25}\) is a characteristic of recurrent events, as inversions were not assessed in other population-based sequencing studies. One recent study of recurrent unbalanced translocations found predictable homology at the recurrent breakpoints\(^{32}\). In contrast, accumulating evidence indicates that formation of both common and rare CNVs is predominantly driven by non-homologous repair mechanisms that are reliant on much smaller degrees of homology between segments than NAHR\(^{14,15,25}\). Our results are in line with those supporting non-homologous repair as the favored mechanism for deletions and the formation of the non-recurring balanced rearrangements sequenced here\(^{15}\); however, the level of microhomology associated with breakpoint signatures of these de novo balanced rearrangements was distinctly lower than reported in the previous studies. There was also a dearth of sequence homology at the integration sites in transgenic animals and in the chromothripsis seen in cancer cells, as determined by our analysis of previous data\(^8\). The distinguishing feature of all three groups (balanced structural variations, cancer cells and transgenic animals) is that they presumably involve randomly selected breakpoints, which increased with the complexity of the structural variation.

These analyses provide an initial sequence-based survey of karyotypically balanced chromosomal rearrangements that, like the previous findings of chromothripsis in cancer, set the stage for extensive further investigation. Yet, our results already have a potentially wide-ranging impact, as they indicate that data from traditional and current clinical diagnostic methods, including chromosomal microarrays, can be insufficient to understand the true nature of genomic disruption in affected individuals. They also raise questions about whether the mechanisms proposed in CNV studies of unphenotyped populations can be fully generalized to other structural genomic alterations, suggesting instead that more complex mechanistic hypotheses are required to explain how the genome can dynamically remodel itself in an unexpectedly extensive manner to produce a stable and viable complex constitutional rearrangement that can contribute to altered human development and disease.

**METHODS**

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

**ACKNOWLEDGMENTS**

We thank all study participants, their families and the referring clinicians for their invaluable contributions. We thank T. Gillis, M.A. Anderson, W. Varney, J. Buliera, T. Henderson, C. Zhang, L. Griffin, The Genomics Core of the Massachusetts General Hospital Center for Human Genetic Research, the Genome Sequencing Platform at the Broad Institute and the Partners Center for Personalized Genomic Medicine for technical assistance. We thank M. Sun for performing the multiplex ligation-dependent probe amplification (MLPA) experiments. We also thank the Partners Research Computing at Massachusetts General Hospital. This work was funded by grants from the US National Institutes of Health (GM061354, HD065286, MH087123, NS32765 and NS16367), the Simons Foundation Autism Research Initiative, Autism Speaks and the CHDI Foundation. J.C.J. was the recipient of the Philip Wrightson Fellowship of the Neurological Foundation of New Zealand, C.E. received a Bissy Fellowship from the Canadian Institute for Health Research (CIHR).

**AUTHOR CONTRIBUTIONS**

M.E.T. and J.F.G. wrote the manuscript, which was edited by all coauthors. C.C., J.C.J., C.E., M.I.D., M.E.M., J.F.G. and M.E.T. designed the experiments. J.C.J., C.E., C.H. and A.M.L. performed the molecular studies. C.C., R.E.M., M.L.B., T.K.O., M.I.D., I.M.H. and M.E.T. designed the bioinformatic and statistical analyses. C.C., A.H., A.K., I.B., T.K.O., I.M.H., R.E.M. and M.E.T. performed the analyses. A.M.L. and C.C.M. performed the FISH analyses. J.C.J., C.C.M., A.M.L., S.R.R., C.J.M., C.S.B., S.J.R., R.L.M.F., R.G.S., Y.S., M.I.D., M.E.M. and J.E.G. provided the subjects and transgenic animals, and C.L. provided the clinical microarray data for the human chromothripsis samples.
Assessment of 2q23.1 microdeletion syndrome implicates the genomic complexity of primary human prostate cancer. balanced chromosome rearrangements in prenatal mapping copy number variation by population-scale genome © 2012 Nature America, Inc. All rights reserved.

1. Iliaakis, G. et al. Mechanisms of DNA double strand break repair and chromosome aberration formation. Cytogenet. Genome Res. 104, 14–20 (2004).
2. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. 1, e49 (2005).
3. Zang, F., Carvalho, C.M. & Lupski, J.R. Complex human chromosomal and genomic rearrangements. Trends Genet. 25, 298–307 (2009).
4. Lee, J.A., Carvalho, C.M. & Lupski, J.R.A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131, 1235–1247 (2007).
5. Hastings, P.J., Ira, G. & Lupski, J.R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 5, e1000327 (2009).
6. 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).
7. Giardino, D. et al. De novo balanced chromosome rearrangements in prenatal diagnosis. Prenat. Diagn. 29, 257–265 (2009).
8. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 40, 722–729 (2008).
9. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144, 27–40 (2011).
10. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. Nature 470, 214–220 (2011).
11. Magrangeas, F., Astbury, C. & Wells, R.D. Non-B DNA conformations, mutagenesis and disease. Trends Biochem. Sci. 32, 271–278 (2007).
12. Hochward, C. & Wells, R.D. Non-B DNA conformations as determinants of mutagenesis and human disease. Trends Biochem. Sci. 28, 273–285 (2003).
13. Abt, M. et al. Comparison of constitutional and replication stress–induced genome structural variation by SNP array and mate-pair sequencing. Genetics 187, 675–683 (2011).
14. Bates, C., Astbury, C. & Wells, R.D. Non-B DNA conformations as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. Am. J. Hum. Genet. 89, 551–563 (2011).
15. Mills, R.E. et al. Mapping copy number variation by population-scale genome sequencing. Nature 470, 59–65 (2011).
16. Lieber, M.R. The mechanism of human nonhomologous DNA end joining. J. Biol. Chem. 283, 1–5 (2008).
17. Berardi, N., Cheng, A., Huang, N. & Stark, J.M. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. PLoS Genet. 4, e1000110 (2008).
18. 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).
19. 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).
20. Khranna, K.K. & Jackson, S.P. DNA double-strand breaks: signaling, repair and the cancer connection. Nat. Genet. 27, 247–254 (2001).
21. Arlt, M.F. et al. Comparison of constitutional and replication stress–induced genome structural variation by SNP array and mate-pair sequencing. Genetics 187, 675–683 (2011).
22. Wörtel, H., Littke, K.C. & Chartrand, P. Illegitimate DNA integration in mammalian cells. Genome Res. 10, 1791–1799 (2003).
23. Jacobson, J.C. et al. An ovine transgenic Huntington’s disease model. Hum. Mol. Genet. 19, 1873–1882 (2010).
24. Kidd, J.M. et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143, 837–847 (2010).
25. Lam, H.Y. et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat. Biotechnol. 28, 47–55 (2010).
26. Wells, R.D. Non-B DNA conformations, mutagenesis and disease. Trends Biochem. Sci. 32, 271–278 (2007).
27. Bacolla, A. & Wells, R.D. Non-B DNA conformations as determinants of mutagenesis and human disease. Mol. Carcinog. 48, 273–285 (2009).
28. Abt, M. et al. Delineation of complex chromosomal rearrangements: evidence for increased complexity. Hum. Genet. 114, 448–457 (2004).
29. Vanneste, E. et al. What next for preimplantation genetic screening? High mitotic chromosome instability rate provides the biological basis for the low success rate. Hum. Reprod. 24, 2679–2682 (2009).
30. Liu, P. et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 146, 889–903 (2011).
31. Ou, Z. et al. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. Genome Res. 21, 33–46 (2011).
**ONLINE 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 whom had a CCR). Subjects were referred to various research studies, including the Developmental Genome Anatomy Project (DGAP), the Autism Consortium of Boston and the Autism Genome Resource Exchange (AGRE), and individuals were also referred directly to a local protocol for the purposes of these experiments. All subjects provided written informed consent or, in the case of the AGRE subjects, de-identified DNA was obtained, and all protocols were approved by the Institutional Review Board of Partners HealthCare System. We also obtained breakpoint sequences from eight DGAP subjects for whose sequencing had previously resolved the junctions (n = 8)\(^33\). Phenotypes associated with balanced structural variations varied widely across DGAP subjects (see ref. 33), 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 (G1/1, G2/2, G4/4, G6/5 and G6/6) generated by microinjection into pronuclei of single-celled zygotes of an 11.6-kb restriction enzyme–generated linear transgene fragment containing full-length human HTT cDNA flanked by 1.1 kb of its 5′ UTR sequence followed by exon 4, intron 4 and exon 5 of the BGH gene\(^34\). 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 1–2 kb of the 5′ UTR, exon 1 and a portion of intron 1 (Fig. 3)\(^33\). Sheep genomic DNA was prepared from tail tissue, and R6/1 and R6/2 mouse genomic DNA was purchased from the Jackson Laboratory (2010 and 2809).

**Massively parallel sequencing of breakpoints.** For human subjects, all sequencing was performed using an Illumina Genome Analyzer IIx or with HiSeq 2000 chemistry (Illumina), and library formation was performed using various approaches, including manufacturer protocols and customized methods, with complete details as described\(^18\). In brief, we used four methods to obtain translocation and inversion breakpoint sequences. Whole-genome sequencing of paired-end 76-bp reads was performed for two subjects to 10× average nucleotide coverage across the genome. Illumina mate-pair sequencing (N = 5 subjects) was performed according to previously described methods\(^24\) and following the manufacturer’s instructions to generate 76-bp paired-end sequences of fragments separated by large inserts of 3–4 kb. Custom jumping libraries (N = 25 subjects) were created for Illumina sequencing by combining aspects of the mate-pair protocols for Illumina and Applied Biosystems ABI SOLID sequencing (Life Sciences). Briefly, 10–20 µg of DNA was sheared using a Covaris S2 sonicator with calibrated parameters and with size selection to derive a tight distribution of fragments of 3.35 kb ± 300 bp. Cap adapters with an EcoP15I restriction site were ligated to fragment ends and circularized to an internal adaptor oligonucleotide with a subject-specific barcode and a bionitinated thymidine. Restriction fragment digestion was then carried out, and bionitinated fragments were retained by streptavidin binding. Libraries were then prepared for Illumina sequencing using reagents from NEB and either Illumina oligonucleotides or custom oligonucleotides designed from IDT based on published sequences\(^35\). Sequencing was performed on a single lane of a Genome Analyzer IIx or a single lane of a HiSeq2000, or samples were multiplexed with those from other subjects on a single lane of a HiSeq2000, depending on available yields at the time of the experiment. Targeted physical coverage was 50× 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\(^25\). Breakpoints for 12 of the subjects were sequenced by a regional capture approach (CapBP)\(^17\) for subjects for whom previous cytogenetic analyses had narrowed at least one breakpoint to within a 1-Mb region. Agilent SureSelect On-Array 244K oligonucleotide 60-mer probes (N = 11 subjects) or In-Solution 55K 120-mer RNA baits (Agilent; N = 1) were used. Regions were tiled at maximum possible overlap, and all bases, including those flagged by RepeatMasker\(^6\), were targeted except for single-nucleotide, dinucleotide and trinucleotide 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 subjects BSID42 and BSID43.

For transgenic animals, we used SureSelect In-Solution target enrichment technology (Agilent) following the manufacturer’s instructions to specifically capture the transgene sequence in each animal. Briefly, we generated 120-mer biotin-tagged probes tiled across the transgene sequence at 1 bp on each strand. We systematically filtered out any probes with singleton, doublet or triplet repeats. We performed nine parallel PCR reactions for each adaptor-ligated library before hybridization. DNA libraries were incubated with capture library probes and appropriate blocking oligonucleotides for 24 h on a thermal cycle at 65°C. After this incubation, libraries were amplified using a SureSelect post-capture indexing primer (reverse primer) and subject-specific indexing primers from the Illumina Multiplex Adaptor kit. We also sequenced customized jumping libraries for the transgenic animals using methods identical to those described above.

**Molecular confirmation of breakpoints by orthogonal methods.** All breakpoints described were confirmed by PCR amplification and capillary sequencing, which involved long-range PCR amplifications and capillary sequencing directly through each of the junction fragments. PCR amplifications were performed using 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 G1/1 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′-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 FISH. Lymphoblast cells were harvested and metaphase spreads prepared by standard protocols. BAC clones spanning breakpoint regions were selected for FISH mapping using the UCSC Genome Browser. BACs from the RP11 library were labeled with either SpectrumOrange- or SpectrumGreen-conjugated dUTP using a nick translation kit (Abbott Molecular), and labeled pairs were hybridized overnight to metaphase chromosome preparations. After four washes with SSC with 0.1% Tween, 2 washes with saline–sodium citrate (buffer) with 0.1% Tween, chromosomes were counterstained with DAPI and analyzed with a Zeiss Axioplan2 microscope with epifluorescence and Applied Imaging CytoVision software.

**Bioinformatics and statistical analyses.** Alignments of paired-end reads for human sequencing were performed using either MAQ\(^37\), BWA\(^38\) or Novoalign (Novocraft) to the hg19 reference, dependent on the time of the analysis and library type (all CapBP samples were aligned with BWA, and jumping libraries were generally aligned with either MAQ (Novoalign) and SAMTools\(^39\). Sequencing alignments were processed with Bambus, a program developed to identify anomalously mapping read pairs, and statistical analyses were performed using R. In transgenic animals, we also attempted naive assembly of the transgene with Velvet 1.0.18 (ref. 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 (ref. 38) and a version of BreakSeq that was modified to assess translocations and inversions\(^22\). Simulation experiments were performed using custom Python scripts and BEDTools\(^41\). Full breakpoint and homology sequences, as well as local IDs for all samples (AGRE or DGAP), are available in Supplementary Table 4.

33. 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).

34. Korbel, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420–426 (2007).

35. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59 (2008).

36. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 16, 210–214 (2000).

37. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858 (2008).

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).

39. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

40. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829 (2008).

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