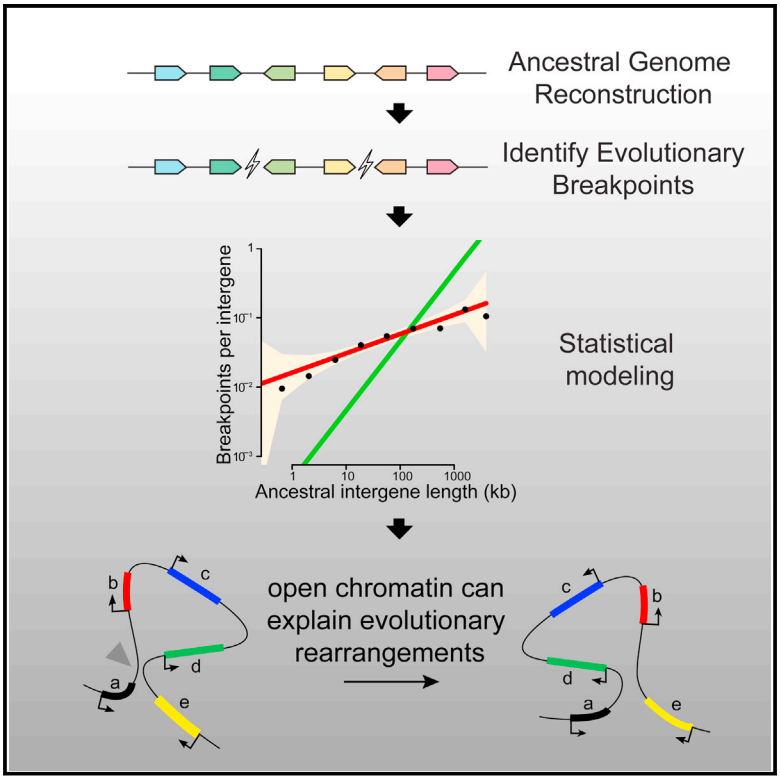
Article

Cell Reports

The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions

Graphical Abstract



Authors

Camille Berthelot, Matthieu Muffato, Judith Abecassis, Hugues Roest Crollius

Correspondence

hrc@ens.fr

In Brief

Evolutionary breakpoints occur non-randomly in genomes, for unknown reasons. Now, Berthelot et al. use ancestral genome reconstructions, statistical modeling, and computer simulations to show that 3D interactions of open chromatin in the nucleus can explain the distribution of breakpoints in five mammalian lineages, and all their known features.

Highlights

- Ancestral genome reveals 750 chromosomal breaks during the evolution of five mammals
- Breakpoint distribution strongly correlates with the length of non-coding spacers
- Statistical models point to open chromatin as the promoter of rearrangements
- Simulations support the role of open chromatin and replicate breakpoint features

Berthelot et al., 2015, Cell Reports *10*, 1913–1924
March 24, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.02.046

Cell Reports
Article

The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions

Camille Berthelot,[1,2,3,4] Matthieu Muffato,[1,2,3,4] Judith Abecassis,[1,2,3] and Hugues Roest Crollius[1,2,3,*]
[1]Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris 75005, France
[2]CNRS, UMR 8197, Paris 75005, France
[3]Inserm, U1024, Paris 75005, France
[4]Present address: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
*Correspondence: hrc@ens.fr
http://dx.doi.org/10.1016/j.celrep.2015.02.046

SUMMARY

Genomic rearrangements are a major source of evolutionary divergence in eukaryotic genomes, a cause of genetic diseases and a hallmark of tumor cell progression, yet the mechanisms underlying their occurrence and evolutionary fixation are poorly understood. Statistical associations between breakpoints and specific genomic features suggest that genomes may contain elusive "fragile regions" with a higher propensity for breakage. Here, we use ancestral genome reconstructions to demonstrate a near-perfect correlation between gene density and evolutionary rearrangement breakpoints. Simulations based on functional features in the human genome show that this pattern is best explained as the outcome of DNA breaks that occur in open chromatin regions coming into 3D contact in the nucleus. Our model explains how rearrangements reorganize the order of genes in an evolutionary neutral fashion and provides a basis for understanding the susceptibility of "fragile regions" to breakage.

INTRODUCTION

Chromosome rearrangements and their biological significance have been central to genome and genetic analyses since the early days of *Drosophila* genetics (Bridges, 1923; Sturtevant, 1925). We know today that inversions, duplications, and translocations have been a major force in the reorganization of the eukaryotic genome both during evolution, and in various genetic diseases, in particular, tumorigenesis. At the molecular level, rearrangements are thought to result from errors in double-strand break repair pathways (mainly non-homologous end joining, but also non-allelic homologous recombination) when simultaneous breaks occur in close proximity in the nucleus (Lupski and Stankiewicz, 2005; Korbel et al., 2007; Meaburn et al., 2007; Quinlan et al., 2010; Zhang et al., 2012), and, to a lesser extent, from fork stalling and template switching during replication (Shaw and Lupski, 2004; Lee et al., 2007; Kidd et al., 2010). Some rearrangements are benign, such as the human chromosome 9 pericentric inversion carried by 0.8%–2.0% of the population with no apparent functional effects (Tawn and Earl, 1992). Others have deleterious consequences, either directly by interrupting functional sequences or indirectly by physically separating regulatory elements that function in *cis* (for examples, see Keung et al., 2004; Benko et al., 2009). Little is known about the fitness effects of rearrangements in eukaryote genomes, but several studies have reported that the majority of events appear to have no functional consequences (Korbel et al., 2007; Baptista et al., 2008; Kidd et al., 2010).

Correspondingly, the basis of the distribution pattern of rearrangement breakpoints in eukaryote genomes has been the subject of much debate. In 1984, Nadeau and Taylor showed that the distribution of segment lengths between consecutive breaks in the order of human and mouse genetic markers was consistent with a pure Poisson process, i.e., that the occurrence and fixation of rearrangements resulted in a random distribution of breakpoints (Nadeau and Taylor, 1984), a conclusion further supported by subsequent studies (Nadeau and Sankoff, 1998; Sankoff and Trinh, 2005). However, more recent inter-specific genomic comparisons have provided increased resolution and have revealed many closely located breakpoints that had previously been overlooked. In addition, computational approaches that identify the most likely rearrangement scenario that could theoretically transform one extant genome into another have inferred a higher frequency of closely located and sometimes indistinguishable breakpoints than would be expected on the basis of random breakpoint occurrence. This phenomenon of "breakpoint reuse" (Pevzner and Tesler, 2003; Bourque et al., 2004; Murphy et al., 2005; Alekseyev and Pevzner, 2007) leads to an excess of clustered breakpoints and has been interpreted as evidence that breakpoints are more likely to occur or become fixed in some "fragile" genomic regions. Subsequent genome-wide studies have consistently shown that both evolutionary and somatic disease-associated rearrangements are non-randomly distributed in the genome, and that rearrangements overlap more frequently than expected (Hinsch and Hannenhalli, 2006; Gordon et al., 2007; Drier et al., 2013).

The observed non-random distribution of rearrangements breakpoints has been interpreted in two distinct ways: either it directly reflects rearrangements preferentially occurring in

"fragile regions" that are more likely to undergo breakage, or the pattern of surviving rearrangements is skewed because of selective elimination of those that occur in functional regions where breakpoints are highly deleterious. There is certainly much evidence that some chromosome regions have a higher propensity to breakage, as observed in cancer genomes (Darai-Ramqvist et al., 2008) and in the finding of recurrent rearrangements associated with certain genetic diseases (Shaw and Lupski, 2004). In particular, the statistical association of rearrangement breakpoints with genomic regions characterized by high GC content, high gene density, replication origins, repeated sequences, or DNA hypomethylation suggests that structural properties may play a substantial role in the occurrence of breakage events, although the relative importance of the different factors underlying this role is debated (Ma et al., 2006; Gordon et al., 2007; Larkin et al., 2009; Lemaitre et al., 2009; Drier et al., 2013; Li et al., 2012). It should be underlined here that "fragile regions" is an operational term employed to describe regions that are prone to a higher incidence of breakage, but not necessarily weaker or fragile in the physical sense, and that a precise description of what constitutes a "fragile region" remains elusive. On the other hand, there presumably must be some selection against breakpoint rearrangements that disrupt certain functional sequences, in particular, coding genes, which are known to be under strong purifying selection and are only very rarely disrupted by breakpoints (Peng et al., 2006). In addition, there may be selective constraint on gene order to preserve clusters of co-expressed genes (Hurst et al., 2004) or the physical linkage of consecutive genes with the interdigitated conserved non-coding sequences responsible for their regulation. These conserved genomic regions include the well-described "genomic regulatory blocks" (GRBs), which exert strong constraints on local gene reorganization in some regions of eukaryotic genomes (Goode et al., 2005; Vavouri et al., 2006; Engström et al., 2007; Kikuta et al., 2007; Hufton et al., 2009; Irimia et al., 2012; Dimitrieva and Bucher, 2013) but could more generally extend to any regions containing regulatory sequences linked in *cis* to their target genes, where at least some rearrangements may be too deleterious to be tolerated by selection (Peng et al., 2006; Becker and Lenhard, 2007; Mongin et al., 2009). In short, the relative roles of mutational and selective processes in shaping the observed distribution of evolutionary breakpoints have not yet been resolved.

What are the forces shaping the rearrangement landscape of the mammalian genome? Here, we argue that this question is particularly difficult to answer if one relies solely on comparisons of rearranged and conserved regions in contemporary genomes. We show how an alternative approach, employing ancestral genome reconstructions and statistical modeling, can be used to assess the respective contribution of structural features and selective pressures to generate the pattern of rearrangement breakpoints seen in five mammalian genomes and in three yeast genomes. We find that mutational explanations alone are sufficient to describe the distribution of breakpoints in intergenic regions of the genome, with relatively weak but measurable evidence for selection to conserve synteny between genes and regulatory elements. Our results indicate that, although there is strong selective constraint on the evolution of gene sequences, changes in gene order are mostly unconstrained and occur neutrally. Strikingly, simulations show that the observed pattern of rearrangements can be accurately replicated when rearrangements occur between regions of open chromatin coming in contact because of chromosomal conformation in the nucleus, both of which have been previously suggested to play a role in rearrangements. Finally, we propose a model to explain the rearrangement process in eukaryotic genomes and suggest an explanation for the susceptibility of "fragile regions" to breakage.

RESULTS

Identification of Evolutionary Rearrangement Breakpoints

We applied a maximum parsimony-based algorithm to reconstruct the ancestral gene order in the 95-million-year-old ancestral genome of Boreoeutheria, the last common ancestor of primates, rodents, and laurasiatherians. With 28 sequenced descendant genomes (in Ensembl v.57) and several closely branching outgroups, the Boreoeutheria ancestor is ideally placed in the mammalian tree for ancestral genome reconstruction and breakpoint analysis over many lineages (Blanchette et al., 2004; Ma et al., 2006; Chauve and Tannier, 2008; Paten et al., 2008; Jones et al., 2012). Existing reconstructions of the Boreoeutheria genome are either short stretches of ancestral sequences reconstructed at the base-pair level, with no information as to how these sequences are ordered in the genome (Paten et al., 2008), or high-level, megabase-scale reconstructions based on a few thousand genomic markers, which are informative for the evolution of the overall chromosome structure but less so for fine-scale rearrangement analysis (Ma et al., 2006; Ouangraoua et al., 2011; Jones et al., 2012). We therefore designed a graph-based parsimony algorithm to reconstruct the high-resolution order of the genes in this ancestral genome based on the gene order in all modern descendants, as described in Experimental Procedures, Figure S1A, and Supplemental Information 1. The reconstructed ancestral genome contains 18,436 gene-to-gene adjacencies, suggesting that this reconstruction is largely complete compared to a typical mammalian genome with 17,000 to 23,000 adjacencies (Supplemental Information 1). This reconstructed genome was further annotated with respect to its intergenic regions (or intergenes, the non-coding sequence between two consecutive genes), specifically, their lengths, GC content and their proportion of conserved non-coding sequence as defined by GERP (Cooper et al., 2005). These features are highly conserved in orthologous intergenes across modern Boreoeutheria genomes, which were used to estimate the ancestral state in each ancestral intergene (Figure S1A; Supplemental Information). With 18,757 gene markers, separated by intergenes with a median length of 19.5 kb, this reconstructed ancestral Boreoeutheria genome is much more resolved than previous versions based on whole-genome alignments (Ma et al., 2006; Zhao and Bourque, 2009; Ouangraoua et al., 2011; Jones et al., 2012) (Supplemental Information; Table S1).

We then identified evolutionary rearrangement breakpoints that have occurred in the human (*Homo sapiens*), mouse (*Mus musculus*), dog (*Canis familiaris*), cow (*Bos taurus*), and horse

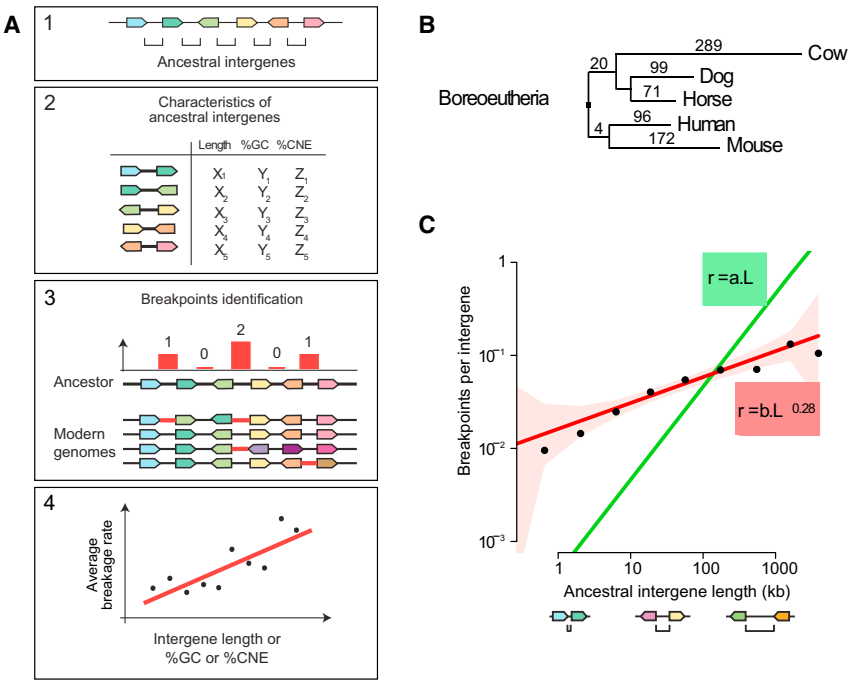


Figure 1. Evolutionary Breakage Rate Is a Log-Linear Function of Intergene Length

(A) Outline of the statistical modeling analysis. Ancestral intergenes have been annotated with three different features: length, GC%, and proportion of conserved non-coding elements (CNEs). The number of breakpoints per intergene is modeled as a function of these features using Poisson regression.

(B) Rearrangement breakpoint counts in the five mammalian genomes under study since the Boreoeutheria ancestor.

(C) In mammalian genomes, the mean number of evolutionary breakpoints per intergene is a power law of intergene length, resulting in a linear correlation after logarithmic transformation (black, observed breakage rates; red, regression equation and 95% confidence interval). The regression model is different from the expectations of the "random model" (green): small intergenes contain more breakpoints than expected, whereas large intergenes contain fewer breakpoints than expected under random breakage. Axes are in log-log scale; a and b are numerical values proportional to the total number of breakpoints and are not biologically informative.

(*Equus caballus*) lineages, each an essentially independent lineage within the Boreoeutheria clade (Experimental Procedures; Supplemental Information). Previous observations have shown that evolutionary breakpoints occur only very rarely within genes because of high selective pressure to maintain gene structure (Lemaitre et al., 2009; Mongin et al., 2009). Therefore, we considered genes as rearrangement-free markers, and intergenes as potential breakpoint regions. We inferred that a breakpoint must have occurred in an ancestral intergene if, in the modern genome, the ancestral genes are no longer adjacent and have new neighbors. By this criterion, we identified a total of 751 breakpoints, 20 of which correspond to independent breakpoint reuse in different lineages (Figures 1A and 1B). The magnitude of these figures is in agreement with previous reports (Ma et al., 2006; Larkin et al., 2009). The breakpoints largely overlap with a previously published, independent data set of 433 breakpoint regions (Larkin et al., 2009) but also reveal previously unidentified breakpoints at a higher resolution (Supplemental Information). Additionally, the identified breakpoints show the typical characteristics of rearrangement breakpoints; i.e., they occur in GC-rich, gene-dense regions possessing lower proportions of conserved non-coding sequence (mean GC content of 44.5% versus that expected at random, 40.7%; mean intergene length of 179 kb versus that expected at random, 882 kb; mean proportion of conserved sequence of 2.4% versus that expected at random: 4.4%; all $p < 2.10^{-16}$, Wilcoxon's rank-sum test).

Breakpoint Frequency Is a Power Law of Ancestral Gene Density

Ancestral genome reconstructions provide a picture of the founding genome of a group before rearrangements occurred in the different, divergent lineages. We used multiple Poisson regression to assess whether the distribution of rearrangement breakpoints in the intergenes of the ancestral genome can be accurately explained based on simple features of these intergenes (Figure 1A; Experimental Procedures; Supplemental Information). Poisson regression is a generalized linear modeling approach that can be used to model the number of occurrences of a rare event (here, rearrangement breakpoints) in intervals (intergenes) according to features of these intergenes (length, GC content, proportion of constrained sequence, etc.). The resulting models report which characteristics are significantly associated with variations in the frequencies of rare events. In addition, by using goodness-of-fit statistics, they describe how accurately these characteristics account for variations in these frequencies. The null hypothesis, which corresponds to the classical Random Breakage Model (Nadeau and Taylor, 1984; Pevzner and Tesler, 2003), is that breakpoint density is uniform in intergenic regions across the genome (genes themselves being under strong selection and "unbreakable"). Under this hypothesis, the average number of breakpoints per intergene (breakage rate) should increase in proportion to intergene length and therefore follow a classical Poisson distribution.

To test the null hypothesis, we constructed a regression model describing the breakage rate as a function of intergene length. We find that a very high positive correlation exists between breakage rates and ancestral intergene lengths, but this correlation does not match the predictions of the Random Breakage Model (Figure 1C). Breakpoint events per intergene increase as a power law of intergene length rather than a proportionality law. This results in a striking linear relationship in

Table 1. Coefficients and Statistics of Poisson Regression Models Describing the Average Number of Breakpoints per Intergene as a Function of Intergene Length, %GC, and %CNE

	Coefficients					Goodness of Fit		
	Simple Regression	Stepwise Regression	P(> |z|)	Null Deviance (df)	Residual Deviance (df)	χ^2 p value	Stepwise χ^2 p value	Pseudo R^2
Model 1: length only								
Intergene length	0.28	—	$< 2.10^{-16}$[a]	167.3 (10)	12.4 (9)	0.19[a]	—	0.93[a]
Model 2: length + %GC								
Intergene length	0.26	0.27	$< 2.10^{-16}$[a]	137.8 (28)	25.7 (27)	0.53[a]	—	0.81[a]
%GC	—	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82
Model 3: length + %CNE								
Intergene length	0.28	0.30	$< 2.10^{-16}$[a]	179.2 (19)	26.3 (18)	0.09[a]	—	0.85[a]
%CNE	—	−4.55	0.01[a]	179.2 (19)	20.7 (17)	0.24[a]	0.02[a]	0.88[a]
Simulation: 3D contacts in open chromatin								
Intergene length	0.28	—	$< 2.10^{-16}$	253.8 (14)	29.6 (13)	0.005	—	0.88

A parameter significantly affecting the breakage rate has a regression coefficient statistically different from 0 (P(> |z|) < 0.05). The goodness of fit of each model is assessed by a χ^2 test on the residual deviance and degrees of freedom (i.e., likelihood ratio test): a non-significant p value means that the residual deviance may be attributed to statistical noise. The effect of an additional parameter on the fit is assessed by a χ^2 test on the difference in residual deviances and degrees of freedom with and without the parameter: a significant p value means that the fit is significantly better with the additional parameter. The pseudo R^2 corresponds to McFadden's pseudo R^2 (proportion of null deviance explained by the model).
For methods, see the Supplemental Information.
[a]Values indicative of an improvement in the model.

log-log scale that corresponds to the following equation (r, breakage rate; L, intergene length):

$$r = 2.4.10^{-3} \times L^{0.28}.$$

Strikingly, 93% of variation in breakpoint occurrence is explained by intergene length with statistical noise accounting for residual variability (McFadden's pseudo $R^2 = 0.93$; likelihood ratio test: p = 0.19; Table 1). As previously reported (Ma et al., 2006; Larkin et al., 2009; Lemaitre et al., 2009), small intergenes (i.e., regions of high gene density) contain more breakpoints than expected, whereas large intergenes (i.e., regions of low gene density) contain fewer breakpoints than expected under random breakage. However, our finding that there is a power-law relationship between intergene length and breakpoint density cannot be readily explained. We checked for a potential confounding effect of GC content, which is strongly correlated with gene density in mammalian genomes, by constructing a second regression model describing breakage rate as a function of both intergene length and GC content. Consistent with previous observations (Lemaitre et al., 2009), the ancestral GC content has no influence on breakpoint occurrence (Table 1).

Selective Pressure to Maintain Synteny between Regulatory Elements and Genes Is Marginal

We then tested whether breakage probability was influenced by the ancestral density of conserved non-coding elements (CNEs), deduced from the conserved non-coding elements detected by GERP (Cooper et al., 2005) across boreoeutherian mammals (Supplemental Information). CNEs are putative regulatory elements that have been conserved over long evolutionary time. It has been proposed that strong selection may act against rearrangements that disrupt synteny between such regulatory elements and their target genes (Kikuta et al., 2007; Hufton et al., 2009; Mongin et al., 2009). Evidence for such constraints exists for several highly regulated genes and their long-distance enhancers, resulting in so-called "genomic regulatory blocks" (GRBs) (Engström et al., 2007; Kikuta et al., 2007; Irimia et al., 2012; Dimitrieva and Bucher, 2013). However, beyond a few specific examples, it is not known whether such constraints are widespread and have a significant impact on the distribution of breakpoints at the scale of entire genomes. If selective constraints to preserve *cis*-regulatory interactions are pervasive genome-wide, we would expect regions with high CNE density to be particularly resistant to rearrangements. To test this hypothesis, we constructed a third regression model describing breakage rate as a function of both ancestral intergene length and ancestral CNE density. Consistent with this prediction, we observe that ancestral intergenes with high CNE content have been disrupted by significantly fewer breakpoints than intergenes of similar length with lower CNE content (Table 1; Figure S10A). However, this difference is marginal and taking into account CNE content improves the fit of the model by only three percentage points of explained deviance (McFadden's pseudo $R^2 = 0.88$) compared to a model built on intergene length only (McFadden's pseudo $R^2 = 0.85$) (Table 1). Selective pressure to preserve synteny between genes and conserved regulatory elements thus exists, but its overall influence on the genome-wide breakpoint distribution is small and probably restricted to a few specific regions of the genome. Interestingly, conserved non-coding elements are not the cause of the genome-wide power law relationship observed between intergene length and breakpoint numbers: this relationship remains even when CNE content

is included in the regression model (Table 1), and the regression equation then becomes (*CNE* = % of intergenic length included in CNEs):

$$r = 2.0.10^{-3} \times L^{0.28} \times e^{-4.55.CNE}.$$

These results show that selection on syntenic relationships between functional genes and their associated regulatory elements is not the main cause of the non-random distribution of evolutionary breakpoints, unlike previously hypothesized (Becker and Lenhard, 2007; Engström et al., 2007; Mongin et al., 2009). The near-perfect correlation between intergene length and number of breakpoints reported here, in fact, suggests that outside of genes, rearrangements are neutral random events but their probability of occurrence at particular sites is biased by structural or functional genomic properties of those sites.

Breakpoints Distribution Is Not an Artifact Caused by Closely Located Inversion Breakpoints

A plausible explanation for the surprisingly strong correlation between intergene length and breakpoint frequency may be because most rearrangements are inversions involving two synchronized, potentially dependent breakpoints (Ma et al., 2006; Zhao and Bourque, 2009) and are not independent events as assumed by the classical Random Breakage Model and the Poisson distribution. If inversions are typically short, many of them may occur within an intergene without disrupting the gene order and would then be missed by our gene-based detection method. This effect would be particularly strong in gene-poor regions, where intergenes are large, and could potentially result in a distribution of breakpoints similar to the one we described above.

To control for this, we tested whether the observed distribution of breakpoints can be approximated by realistic simulations of inversions in the human genome. The true distribution of inversions lengths in mammalian genomes is unknown; however, rearrangements have been shown to occur between regions in close 3D proximity in the nucleus in different contexts (Branco and Pombo, 2006; Véron et al., 2011; Zhang et al., 2012), suggesting that contact probability is a good proxy for rearrangement probability. We used the Hi-C map of the human genome to estimate the contact and inversion probability between any two regions of a chromosome (Lieberman-Aiden et al., 2009). We then sampled pairs of breakpoints according to this probability (Figure 2A; Experimental Procedures; Supplemental Information). In line with the distance-dependent nature of DNA contacts in the nucleus, the simulations realistically produce a large number of short rearrangements, many of which would be undetectable to us because they do not encompass a gene and do not modify the gene order (Figure 2B). But even when restricted to detectable breakpoints alone, the observed breakage probability as a function of intergene length does not deviate greatly from the expectations of the Random Breakage model and does not reflect the observations of the real data (Figure 2C). This control suggests that the power law found with true breakpoints is not explained by the dependency between inversion breakpoints alone. We therefore examine next which genomic feature(s) or mechanism(s) may result in such a distribution.

Breakpoints Density Is Reminiscent of the Density of Open Chromatin in Modern Genomes

Repeated elements (Ovcharenko et al., 2005; Ma et al., 2006; Carbone et al., 2009; Zhao and Bourque, 2009), recombination (Larkin et al., 2009; Völker et al., 2010), replication origins (Di Rienzi et al., 2009; Lemaitre et al., 2009), topological chromatin domain limits (Dixon et al., 2012), and open chromatin (Lemaitre et al., 2009; Véron et al., 2011) have all been suggested to influence rearrangements. If one of these feature(s) causes rearrangements, we would expect it to be distributed similarly to breakpoint density in modern genomes, i.e., be denser in short intergenes and less frequent in large intergenes compared to the uniform density expected by chance (Supplemental Information 5). In contrast, we find that repeated elements and recombination frequencies are distributed radically differently from breakpoints (Figure 2D), eliminating them as potential candidates to explain the breakpoint pattern. Both replication origins and limits of topological domains are more frequent in short intergenes, thereby correlating with breakpoint density. Replication origins and limits of topological domains are both partly conserved in mammals (Ryba et al., 2010; Dixon et al., 2012). Therefore, we mapped these inherited features to the ancestral Boreoeutheria genome but found that breakpoints have not significantly co-occurred with either replication origins or topological domain boundaries (Supplemental Information). The density of open chromatin, however, is similar to the pattern of breakpoints with the proportion of DNA in an open state decreasing as intergene size increases (Figure 2D). These results suggest DNA accessibility as a plausible candidate for the primary determinant of rearrangement probability.

Breakpoints Simulated between Open Regions in Contact in the Nucleus Reproduce the Evolutionary Distribution

Previous reports have hypothesized a role for chromatin structure in the occurrence of rearrangements (Drier et al., 2013; Li et al., 2012; Roukos et al., 2013). Whether this alone can explain the genome-wide distribution of rearrangements, however, is not known. Because this model would readily explain our observations (Figure 2B), we tested this hypothesis by simulating inversions in the human genome according to contact probability as described above, except that rearrangements were allowed to occur only between open chromatin regions, using chromatin state profiles for different cell types published by the ENCODE consortium (ENCODE Project Consortium, 2012) (Experimental Procedures; Supplemental Information). Under this model, the simulated average number of breakpoints per intergene closely reproduces the relationship with intergene length observed in real data (Figure 2F). Specifically, when regression is performed, simulated breakage rates increase as a function of intergene length and follow a power law with the same coefficients as seen with real breakpoints—a result that is highly unlikely to arise by chance (Table 1). This result is not a coincidental finding, because simulations performed using open chromatin profiles from four different cell types result in strikingly similar average

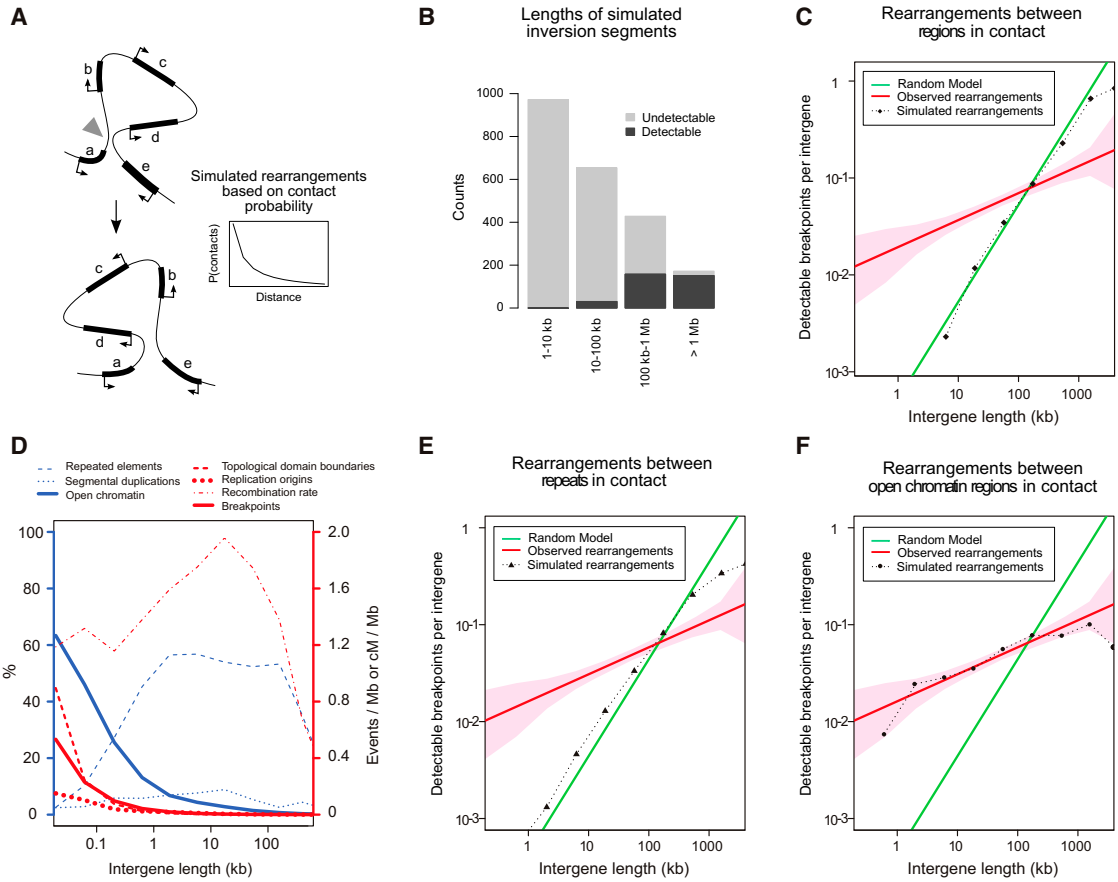


Figure 2. Simulated Rearrangement Breakpoints between Open Chromatin Regions that Are in 3D Contact in the Nucleus Reproduce the Distribution of True Evolutionary Breakpoints

(A) Inversions are simulated in the human genome (gray arrow) based on the probability of 3D DNA contacts experimentally derived from Hi-C studies (right inset).

(B) Length distribution of simulated inversions (average over 100 iterations). Because the breakpoint detection method in this study is based on gene order, inversions that do not encompass genes cannot be detected. Simulations that produce a number of detectable breakpoints equal to real breakpoint data also produce a large number of short, undetectable rearrangements that do not affect gene order.

(C) Simulated rearrangements based on the probability of 3D contact alone result in a distribution of detectable breakpoints similar to the random model expectation and do not appropriately reproduce the observation of real data (green, random distribution; red, observed distribution of breakpoints and 95% confidence interval as in Figure 1C; dotted line and diamonds: simulated breakpoints).

(D) Genomic features associated with rearrangements are expected to follow the same distribution trend as breakpoints, i.e., be denser in small intergenes than in large ones. This is the case for open chromatin, replication origins, and topological domains boundaries. Blue and red curves refer to blue (left) and red (right) axes, respectively. Values on the right axis should be multiplied by 10^{-4} for breakpoints and replication origins, and by 10^{-3} for topological domains boundaries.

(E) Simulated rearrangements between repeated sequences in 3D contact (dotted line and triangles) do not follow the distribution of real data breakpoints (red line), but rather the expectations of the random distribution control (green).

(F) Rearrangements between open chromatin regions in 3D contact (dotted line and circles) result in a distribution of detectable breakpoints similar to real breakpoints (red line; shaded area: 95% CI of the distribution of real breakpoints), showing that this mechanism would appropriately explain the biased occurrence of rearrangement breakpoints in mammalian genomes.

breakage patterns (Figure S10C), which reflect the higher-order properties of open chromatin regions and, notably, their increased density around genic sequences (Thurman et al., 2012). Conversely, this breakpoint pattern is not reproduced when the same simulations are performed with other interspersed features such as transposable elements instead of open chromatin regions (Figure 2E).

Taken together, these results suggest a simple model in which chromosomal rearrangements occur in a biased manner due to misrepaired double-strand breaks between active chromatin domains in physical contact in the nucleus and then are mostly evolutionarily neutral if they do not directly disrupt a functional sequence. Remarkably, our model of randomly generated breakpoints simulated solely on the basis of open chromatin profiles and 3D contact probability replicates the known genomic properties of rearrangement breakpoints. Similar to real breakpoints, breakpoints simulated according to this model show associations with higher gene density, CpG islands, segmental duplications and repeats (Figure 3A) (Murphy et al., 2005; Ovcharenko et al., 2005; Ma et al., 2006; Carbone et al.,

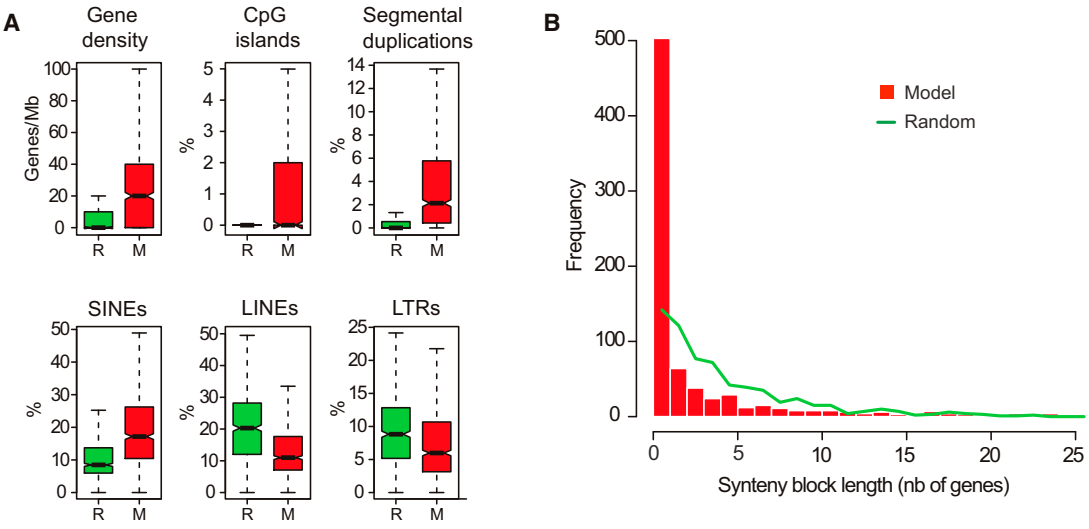


Figure 3. Simulations of Rearrangements between Open Chromatin Regions in Contact in the Nucleus Result in Detectable Breakpoints Exhibiting the Known Features of Real Evolutionary Breakpoints

(A) Simulated breakpoint regions display high gene density, CpG island, segmental duplication and SINEs content, and low LINEs and LTR content. Red, breakpoint regions under the model; green, random control. Wilcoxon's rank-sum tests: gene density, $p = 8.05\ 10^{-158}$; CpG islands, $p = 3.3\ 10^{-18}$; segmental duplications, $p = 4.9\ 10^{-114}$; SINEs, $p = 6.8\ 10^{-124}$; LINEs, $p = 1.5\ 10^{-111}$; LTRs, $p = 8.8\ 10^{-38}$. Percentages correspond to the proportion of the rearranged intergene covered by each feature.

(B) The distribution of synteny block lengths defined by simulated rearrangements between open chromatin regions in contact in the nucleus (in red) displays a high number of very short synteny blocks. The distribution of synteny blocks expected under a random distribution of breakpoints is superimposed (green line).

2009; Larkin et al., 2009). The simulated breakpoints also partition the genome into synteny blocks in which there is a striking excess of short blocks (Figure 3B). Notably, this finding formed the basis for the initial hypothesis that the genome contains regions of higher breakage probability, or "fragile" regions (Pevzner and Tesler, 2003; Zhao et al., 2004) and was previously observed in real genomic comparisons. The excess observed here is similar in magnitude to the excess predicted by these early studies.

Last, the simulated breakpoints were compared to regions of functional relevance in the human genome. We calculated the predicted rate of disruption for 241 ultra-conserved genomic regulatory blocks (uGRBs), corresponding to arrays of non-coding elements and genes, mostly involved in development, found in conserved order between human and chicken and thought to be under selective constraint (Dimitrieva and Bucher, 2013), as well as 2,996 topologically associated domains (TADs) corresponding to highly self-interacting chromatin structures that are thought to be important for gene expression and regulation (Dixon et al., 2012). We found that the predicted rate of uGRB disruption is consistent with observations, with ten human uGRBs predicted to be disrupted by large inversions in mouse, similar to the number seen in real data (odds ratio = 1.03; p = 1, Fisher's exact test) (Dimitrieva and Bucher, 2013). The model also predicts the existence of an average of 19 small inversions within GRBs that do not affect the organization of their conserved elements. Conversely, we found that 19% of simulated rearrangements are expected to displace a TAD boundary, affecting 12% of all TADs during the evolution of five different mammals. This is, in fact, a much lower rate of TAD disruption than observed in real data, since as high as 25%–50% of TADs boundaries have changed between human and mouse (Dixon et al., 2012). This result is in line the authors' observations that TAD boundaries are fairly flexible and largely change due to transposition of repeated elements carrying CTCF binding sites, rather than rearrangements.

Yeast Genomes Display a Characteristic Breakpoint Pattern Similar to Mammals

To assess whether our findings might extend more widely across eukaryotic genomes, we reconstructed the ancestral genome of the last common ancestor of *Kluyveromyces* and *Lachancea* yeasts (Supplemental Information). We identified 505 rearrangement breakpoints since the ancestor in three mostly independent lineages, *Lachancea kluyveri*, *Lachancea waltii*, and *Kluyveromyces lactis*, a finding consistent with a previous analysis in these genomes (Figure 4A) (Gordon et al., 2009). We found that breakage rates in yeasts follow a very similar correlation with ancestral intergene lengths as seen in mammals, suggesting that a similar occurrence mechanism could be at work in yeast genomes (Figure 4B; Supplemental Information).

DISCUSSION

We still miss a general model of the dynamics of genome organization, and, nearly a century after the discovery of the first chromosome rearrangement (Bridges, 1923), we still cannot explain the biased distribution of these rearrangement in genomes. Previous studies aimed at characterizing breakpoint regions have frequently reported that breakpoints statistically

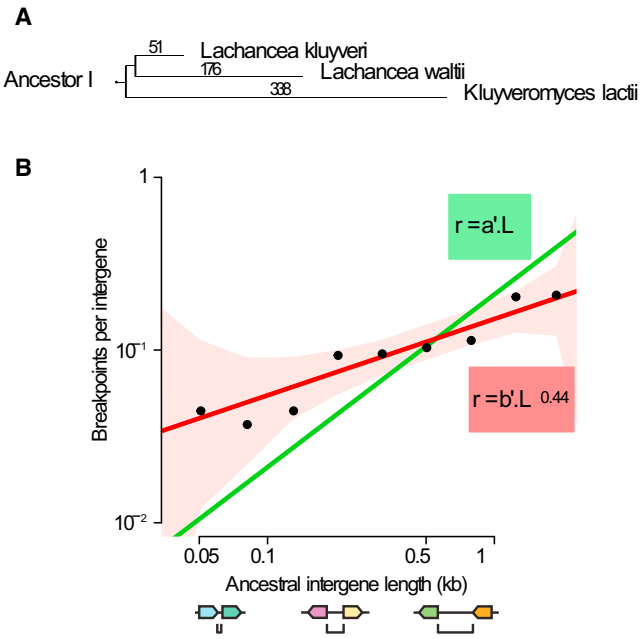


Figure 4. Rearrangement Breakpoints in Yeast Genomes Display a Similar Distribution to Those of Mammalian Genomes

(A) Rearrangement breakpoint counts in the three yeast genomes under study since ancestor I.

(B) Similarly to mammals, breakage rates follow a power law of intergene length in yeast genomes. Black: observed breakage rates; red: regression equation and 95% confidence interval; green: random model. Axes are in log-log scale.

associate with features including regions of high gene density, high GC content, and high repeat content. However, these studies could not distinguish between true determinants of breakage and secondary correlations, or disentangle mutational effects related to structural fragility from selective effects (i.e., purifying selection against chromosomal breakage). Here, we show that the distribution of rearrangements can be accurately explained as misrepaired breaks between open chromatin regions in non-coding regions that are brought into contact by the three-dimensional conformation of chromosomes in the nucleus, which also provides a direct explanation for their mechanism of occurrence. The distribution of open chromatin regions and the distance-dependent nature of chromatin-chromatin interactions result in this biased breakpoint pattern. Our model not only explains but also reproduces in silico the genome-wide pattern of evolutionary rearrangement breakpoints observed in eukaryotes. Notably, we observe the same striking linear relation between intergene size distribution and breakpoint rates in both mammals and yeasts, suggesting that the proposed model may be acting over a very broad evolutionary scale, and possibly in all eukaryote genomes.

The idea that chromatin dynamics influence mutational processes has recently been put forward in other contexts by several reports. Physical damage to DNA occur preferentially in open chromatin (Cowell et al., 2007; Kim et al., 2007) and in active sites of transcription (Chiarle et al., 2011; Klein et al., 2011). Additionally, convergent evidence from induced double-strand breaks (DSBs) (Kruhlak et al., 2006; Soutoglou et al., 2007; Jakob et al., 2009; Zhang et al., 2012) and natural translocations in vivo (Roix et al., 2003) suggest that rearrangements will preferentially take place between genomic sites closely located within the nucleus after DSB. Our results unify these observations at the whole-genome scale and suggest that mechanistic processes not only largely govern the initial occurrence of rearrangements but also, ultimately, their genome-wide distribution. Indeed, according to our model, the evolutionary fates of rearrangements mainly depend on the location of the breakpoints, rather than on the content of the region being rearranged. While genes and a small number of *cis*-regulatory interactions are under strong negative selection against disruption of synteny, rearrangement breakpoints occurring in non-coding regions are generally neutral and their distribution will mainly reflect their initial probability of occurrence. Unlike previously hypothesized, the biased distribution of rearrangement breakpoints would not be primarily the consequence of selection that maintains the local organization of genes and their conserved regulatory elements. An important consequence that follows is that gene order is mostly unconstrained too. While we do detect an influence of the presence of conserved regulatory regions on rearrangement rates in our model, the negative selection that it imposes on genome organization is marginal and probably restricted to very specific areas of the genome.

Interestingly, the occurrence of breakpoints in regions of open chromatin provides an attractive answer to the question of breakpoint reuse, and to the related question of fragile regions. Our simulations reproduce almost exactly the excess of small synteny blocks (Figure 3B) that led to the initial ''breakpoint reuse'' scenario (Pevzner and Tesler, 2003). Fragile regions therefore exist but rather than reflecting fragility of the DNA sequence itself, their vulnerability is a consequence of their chromatin state. Additionally, since higher-order chromatin organization is mostly conserved across mammals (Chambers et al., 2013), the space of possible rearrangements will be similar across species, resulting in higher co-occurrence and recurrence of breakpoints than expected under uniformity.

Finally, the model also clarifies another debated question: that transposable elements are a major cause of disruption of the integrity of the genome at the origin of evolutionary breakpoints (Ruiz-Herrera and Robinson, 2008; Cordaux and Batzer, 2009). Although intuitively attractive, the notion that repeated elements directly promote rearrangements via non-homologous recombination events has been hard to ascertain, especially given that similar elements can easily be found by chance at or near the position of breakpoints. There is evidence that repeated elements are involved in a number of recurrent rearrangements in human (Lupski and Stankiewicz, 2005), but these only represent a small fraction of all rearrangements, which are mostly non-recurrent (Kidd et al., 2010). In contrast, experimental results have shown that repeated elements do not influence rearrangements frequency, even though they can affect the pathway of choice for breakpoint repair (Elliott et al., 2005; Weinstock et al., 2006). This is in agreement with our results showing that recombination

between repeated elements alone cannot explain the breakage pattern of mammalian genomes.

Importantly, although our model has been developed on evolutionary rearrangement data, it is probably also relevant to other types of rearrangements, especially somatic and cancer rearrangements. In this context, our model provides a unified synthesis for many seemingly contradictory observations and suggests three main predictions. First, and paradoxically, rearrangements are expected to occur in gene-dense, actively transcribed regions of the genome. A recent survey of rearrangements in human cancers shows that this is indeed the case (Stephens et al., 2009). Additionally, this same study showed that the large majority of breakpoints in cancer occur with a distance of 2 Mb, much closer than expected by chance but consistent with the strong influence from the intra-nucleus chromatin interactions that our model accounts for. Second, evolutionary and cancer breakpoints are expected to significantly cluster and to share genomic characteristics, as previously reported (Murphy et al., 2005; Darai-Ramqvist et al., 2008), since they would be generated by the same mutational mechanisms and should exhibit similar genomic trends. Third, according to our model, cancer-associated rearrangements are expected to have tissue-specific characteristics reflecting the chromatin architecture of their tissue of origin and to reoccur in a tissue-specific manner.

Potentially, the most interesting application of our model lies in its ability to predict rearrangements probabilities. As proof of concept, we report here that our model appropriately reproduces not only the characteristics of rearrangement regions, but also the local rearrangement rates observed in a number of genomic structures. More generally, our results suggest that maps of open chromatin domains and 3D genomic contacts are sufficient to compute genome-wide, high-resolution rearrangement probabilities in any lineage or cell type. Such data are becoming increasingly available with the improvement and widespread use of functional genomics methods in the past few years. Predictions of local rearrangement probabilities would provide a baseline to detect regions that consistently deviate from their expected rearrangement pattern. This would, in turn, enable the identification of rare gene topologies that are more resistant to rearrangement in multiple lineages than would be expected (and that are probably functional). Notably, when we consider five independent mammalian lineages, our results suggest that large intergenes (>100 kb) have a breakage probability of approximately 10%. Therefore, we estimate that data from 100 species carefully selected to represent the mammalian phylogeny should provide sufficient statistical power to permit the well-resolved mapping of evolutionary constraint on genome re-organization in mammalian genomes. In the context of somatic rearrangements, cell-type-specific predictions of rearrangement probabilities could allow rearrangement-prone genomic regions to be identified. Such regions may indicate the existence of additional types of genomic fragility or the action of positive selection on some rearrangements. Using the same predictive approach, it will be possible to identify regions that are resistant to rearrangement because such rearrangements are lethal to the cell. The model we propose here may thus serve as a theoretical framework to better understand not only germline rearrangements leading to evolutionary fixation or to disease but also cell-type-specific somatic rearrangements occurring during tumorigenesis.

EXPERIMENTAL PROCEDURES

Supplemental Experimental Procedures are provided as Supplemental Information for all analyses.

Ancestral Genome Reconstructions and Estimation of Ancestral Genomic Features

Information on gene trees and gene order were downloaded from Ensembl v.57 (Flicek et al., 2013) for all available genomes (51 species). For yeasts, the gene order information was obtained from Genolevures for 11 species (Sherman et al., 2009); gene trees were built using TreeBest (Vilella et al., 2009). The ancestral genome reconstruction method computes pairwise comparisons of gene order for all pairs of species that are informative for the ancestor of interest; i.e., the ancestor is on the pathway between both species in the phylogenetic tree. Pairs of genes that are directly next to each other and in the same orientation in two such genomes are considered as gene adjacencies inherited from their last common ancestor. Conflicts were resolved using a weighted graph algorithm selecting the most likely ancestral gene order from the number of informative pairwise genome comparisons in support of each gene adjacency (Supplemental Information).

The length, GC content, and total conserved non-coding sequence as defined by GERP (Cooper et al., 2005) of ancestral intergenes were estimated based on their values in all sequenced modern descendants of the ancestor of interest (28 species for mammals, five for yeasts). In each case, the median modern value was used as an estimate of the ancestral value.

Identification of Evolutionary Rearrangement Breakpoints

The ancestral gene order was compared to each of the five modern genomes under study (human mouse, dog, cow, and horse) to identify rearrangement breakpoints, i.e., pairs of genes that have different neighbors in the modern genome than their ancestral counterparts. Cases due solely to gene gains, losses, and duplications were not considered as rearrangements as they may arise through other mechanisms (polymerase slippage, loss-of-function mutations, retrotransposition, etc.). Additionally, dubious rearrangement events consistent with errors in ancestral or modern genome assemblies were removed from the data set (Supplemental Information). Breakpoints were compared with a previously published set obtained in four out of the five species used in our analysis (Larkin et al., 2009). Larkin et al.'s data set describes the human coordinates of regions of discontinuity with another mammalian genome. We tested whether these human regions descend from one of the breakpoint regions we identified in the ancestral Boreoeutheria genome, in which case we consider that we successfully identified the same rearrangement event (Supplemental Information).

Statistical Modeling Using Generalized Linear Models: Poisson Regression

The multivariate regression analysis was carried out in R (http://www.R-project.org/) using the generalized linear models implemented in the *glm()* function. Intergenes were stratified into classes of similar length (bins of width 0.5 in log scale), then further into classes of GC content (bins of 0.2) or into top 50% and lower 50% according to the proportion of conserved non-coding elements. The mean value of each parameter was used as the predictor value for each class of intergenes in the regression. A stepwise regression procedure was carried out to progressively add new variables in the model, in an order determined by their initial performance in explaining the data (intergene length, then GC content or proportion of CNEs; see the Supplemental Information for details on the regression model and procedure). The goodness of fit at each step was estimated using a χ^2 test on the residual deviance and degrees of freedom of the model (likelihood ratio test). Non-significance ($p > 0.05$) denotes that variations between the model, and the data are consistent with statistical noise. A new parameter was retained in the model when a χ^2 test on the difference of residual deviances with and without the parameter (with one degree of freedom) was significant.

Of note, this was always in agreement with Akaike's Information Criterion (no over-fitting).

Genome-wide Simulations of Breakage

Pairs of breakpoints were simulated by drawing an intergenic base randomly in the human genome as the first breakpoint and then a second breakpoint at a distance *d* according to the probability distribution derived from (Lieberman-Aiden et al., 2009), which describes the probability of contact of two loci according to their distance. If the space between both breakpoints encompasses at least one gene, the breakpoints were recorded as detectable by our gene-based method (and otherwise as undetectable). This process was repeated until we obtained as many detectable breakpoints as observed between Boreoeutheria and the five lineages under study (see the Supplemental Information). To simulate rearrangements driven by specific genomic regions, a condition was applied to record breakpoints only when they were both drawn from open chromatin regions (identified by the ENCODE project, Supplemental Information), from transposable elements (TEs) of the same class (SINEs, LINEs, LTR, DNA) or from TEs strictly of the same type (AluY, MIRb, L1M4, and so forth), as annotated by RepeatMasker.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.046.

AUTHOR CONTRIBUTIONS

C.B. and H.R.C. designed the analyses. C.B. performed the analyses with help from J.A. M.M. contributed algorithms. The manuscript was written by C.B. and H.R.C.

ACKNOWLEDGMENTS

We thank Alexandra Louis and Pierre Vincens for assistance with computing resources. We thank Claude Thermes for sharing updated data on predicted human replication origins location and for critical reading of an earlier version of the manuscript. We thank Brian Cusack and Adam Wilkins for advices during the writing of the manuscript. This work was supported by a bursary from the French Ministry of Higher Education and Research to C.B. and by the programme Investissements d'Avenir launched by the French Government and implemented by the ANR (ANR-10-LABX-54 MEMO LIFE).

Received: July 15, 2014
Revised: December 17, 2014
Accepted: February 18, 2015
Published: March 19, 2015

REFERENCES

Alekseyev, M.A., and Pevzner, P.A. (2007). Are there rearrangement hotspots in the human genome? PLoS Comput. Biol. *3*, e209.

Baptista, J., Mercer, C., Prigmore, E., Gribble, S.M., Carter, N.P., Maloney, V., Thomas, N.S., Jacobs, P.A., and Crolla, J.A. (2008). Breakpoint mapping and array CGH in translocations: comparison of a phenotypically normal and an abnormal cohort. Am. J. Hum. Genet. *82*, 927–936.

Becker, T.S., and Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: a matter of resolution. Mol. Genet. Genomics *278*, 487–491.

Benko, S., Fantes, J.A., Amiel, J., Kleinjan, D.-J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C.T., et al. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat. Genet. *41*, 359–364.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. *14*, 2412–2423.

Bourque, G., Pevzner, P.A., and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res. *14*, 507–516.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. *4*, e138.

Bridges, C.B. (1923). The translocation of a section of chromosome-II upon chromosome-III in Drosophila. Anat. Rec. *24*, 426–427.

Carbone, L., Harris, R.A., Vessere, G.M., Mootnick, A.R., Humphray, S., Rogers, J., Kim, S.K., Wall, J.D., Martin, D., Jurka, J., et al. (2009). Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genet. *5*, e1000538.

Chambers, E.V., Bickmore, W.A., and Semple, C.A. (2013). Divergence of mammalian higher order chromatin structure is associated with developmental loci. PLoS Comput. Biol. *9*, e1003017.

Chauve, C., and Tannier, E. (2008). A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comput. Biol. *4*, e1000234.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Cowell, I.G., Sunter, N.J., Singh, P.B., Austin, C.A., Durkacz, B.W., and Tilby, M.J. (2007). gammaH2AX foci form preferentially in euchromatin after ionising-radiation. PLoS ONE *2*, e1057.

Darai-Ramqvist, E., Sandlund, A., Müller, S., Klein, G., Imreh, S., and Kost-Alimova, M. (2008). Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. Genome Res. *18*, 370–379.

Di Rienzi, S.C., Collingwood, D., Raghuraman, M.K., and Brewer, B.J. (2009). Fragile genomic sites are associated with origins of replication. Genome Biol. Evol. *1*, 350–363.

Dimitrieva, S., and Bucher, P. (2013). UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res. *41*, D101–D109.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., Beroukhim, R., and Getz, G. (2013). Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Elliott, B., Richardson, C., and Jasin, M. (2005). Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. Mol. Cell *17*, 885–894.

Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. *17*, 1898–1908.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48–D55.

Goode, D.K., Snell, P., Smith, S.F., Cooke, J.E., and Elgar, G. (2005). Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. Genomics *86*, 172–181.

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., Crooijmans, R., Groenen, M., Lucas, S., Ovcharenko, I., and Stubbs, L. (2007). Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. Genome Res. *17*, 1603–1613.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. PLoS Genet. *5*, e1000485.

Hinsch, H., and Hannenhalli, S. (2006). Recurring genomic breaks in independent lineages support genomic fragility. BMC Evol. Biol. *6*, 90.

Hufton, A.L., Mathia, S., Braun, H., Georgi, U., Lehrach, H., Vingron, M., Poustka, A.J., and Panopoulou, G. (2009). Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. Genome Res. *19*, 2036–2051.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. *5*, 299–310.

Irimia, M., Tena, J.J., Alexis, M.S., Fernandez-Miñan, A., Maeso, I., Bogdanović, O., de la Calle-Mustienes, E., Roy, S.W., Gómez-Skarmeta, J.L., and Fraser, H.B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. *22*, 2356–2367.

Jakob, B., Splinter, J., Durante, M., and Taucher-Scholz, G. (2009). Live cell microscopy analysis of radiation-induced DNA double-strand break motion. Proc. Natl. Acad. Sci. USA *106*, 3172–3177.

Jones, B.R., Rajaraman, A., Tannier, E., and Chauve, C. (2012). ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics *28*, 2388–2390.

Keung, Y.K., Beaty, M., Powell, B.L., Molnar, I., Buss, D., and Pettenati, M. (2004). Philadelphia chromosome positive myelodysplastic syndrome and acute myeloid leukemia-retrospective study and review of literature. Leuk. Res. *28*, 579–586.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837–847.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. *17*, 545–555.

Kim, J.A., Kruhlak, M., Dotiwala, F., Nussenzweig, A., and Haber, J.E. (2007). Heterochromatin is refractory to gamma-H2AX modification in yeast and mammals. J. Cell Biol. *178*, 209–218.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell *147*, 95–106.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Kruhlak, M.J., Celeste, A., Dellaire, G., Fernandez-Capetillo, O., Müller, W.G., McNally, J.G., Bazett-Jones, D.P., and Nussenzweig, A. (2006). Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. J. Cell Biol. *172*, 823–834.

Larkin, D.M., Pape, G., Donthu, R., Auvil, L., Welge, M., and Lewin, H.A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res. *19*, 770–777.

Lee, J.A., Carvalho, C.M.B., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235–1247.

Lemaitre, C., Zaghloul, L., Sagot, M.F., Gautier, C., Arneodo, A., Tannier, E., and Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. BMC Genomics *10*, 335.

Li, J., Harris, R.A., Cheung, S.W., Coarfa, C., Jeong, M., Goodell, M.A., White, L.D., Patel, A., Kang, S.-H., Shaw, C., et al. (2012). Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. PLoS Genet. *8*, e1002692.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lupski, J.R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. *1*, e49.

Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. Genome Res. *16*, 1557–1565.

Meaburn, K.J., Misteli, T., and Soutoglou, E. (2007). Spatial genome organization in the formation of chromosomal translocations. Semin. Cancer Biol. *17*, 80–90.

Mongin, E., Dewar, K., and Blanchette, M. (2009). Long-range regulation is a major driving force in maintaining genome integrity. BMC Evol. Biol. *9*, 203.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science *309*, 613–617.

Nadeau, J.H., and Sankoff, D. (1998). The lengths of undiscovered conserved segments in comparative maps. Mamm. Genome *9*, 491–495.

Nadeau, J.H., and Taylor, B.A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA *81*, 814–818.

Ouangraoua, A., Tannier, E., and Chauve, C. (2011). Reconstructing the architecture of the ancestral amniote genome. Bioinformatics *27*, 2664–2671.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res. *15*, 137–145.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., and Birney, E. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. *18*, 1829–1843.

Peng, Q., Pevzner, P.A., and Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. PLoS Comput. Biol. *2*, e14.

Pevzner, P., and Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc. Natl. Acad. Sci. USA *100*, 7672–7677.

Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. *20*, 623–635.

Roix, J.J., McQueen, P.G., Munson, P.J., Parada, L.A., and Misteli, T. (2003). Spatial proximity of translocation-prone gene loci in human lymphomas. Nat. Genet. *34*, 287–291.

Roukos, V., Burman, B., and Misteli, T. (2013). The cellular etiology of chromosome translocations. Curr. Opin. Cell Biol. *25*, 357–364.

Ruiz-Herrera, A., and Robinson, T.J. (2008). Evolutionary plasticity and cancer breakpoints in human chromosome 3. BioEssays *30*, 1126–1137.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

Sankoff, D., and Trinh, P. (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. J. Comput. Biol. *12*, 812–821.

Shaw, C.J., and Lupski, J.R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum. Mol. Genet. *13*, R57–R64.

Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.-L., and Durrens, P.; Génolevures Consortium (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res. *37*, D550–D554.

Soutoglou, E., Dorn, J.F., Sengupta, K., Jasin, M., Nussenzweig, A., Ried, T., Danuser, G., and Misteli, T. (2007). Positional stability of single double-strand breaks in mammalian cells. Nat. Cell Biol. *9*, 675–682.

Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005–1010.

Sturtevant, A.H. (1925). The effects of unequal crossing over at the bar locus in Drosophila. Genetics *10*, 117–147.

Tawn, E.J., and Earl, R. (1992). The frequencies of constitutional chromosome abnormalities in an apparently normal adult population. Mutat. Res. *283*, 69–73.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Vavouri, T., McEwen, G.K., Woolfe, A., Gilks, W.R., and Elgar, G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet. *22*, 5–10.

Véron, A.S., Lemaitre, C., Gautier, C., Lacroix, V., and Sagot, M.-F. (2011). Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. BMC Genomics *12*, 303.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Völker, M., Backström, N., Skinner, B.M., Langley, E.J., Bunzey, S.K., Ellegren, H., and Griffin, D.K. (2010). Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res. *20*, 503–511.

Weinstock, D.M., Elliott, B., and Jasin, M. (2006). A model of oncogenic rearrangements: differences between chromosomal translocation mechanisms and simple double-strand break repair. Blood *107*, 777–780.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

Zhao, H., and Bourque, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. Genome Res. *19*, 934–942.

Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W.C., Strausberg, R.L., and Fraser, C.M. (2004). Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. Genome Res. *14* (*10A*), 1851–1860.

Cell Reports

Supplemental Information

The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions

Camille Berthelot, Matthieu Muffato, Judith Abecassis, and Hugues Roest Crollius

1. Reconstruction of the ancestral Boreoeutheria genome

Computing ancestral gene adjacencies

The rationale of the method is that pairs of genes that are next to each other and in the same orientation in two genomes are considered as potentially inherited from their last common ancestor. If the ancestral genome of interest is on the branch between this last common ancestor and any of the two modern genomes, then this inherited gene configuration must have existed in this ancestral genome as well. The method computes all relevant pairwise gene order comparisons in a phylogeny-aware manner, and assigns a weight to each possible gene-gene adjacency, corresponding to the number of pairwise comparisons that support this adjacency as ancestral at the node of interest. These weights are then used to resolve a gene graph and extract the most probable gene order in the ancestor of interest. Of note, pairwise comparisons between species whose last common ancestor is more recent than the ancestor of interest are not included in the computation, as they are not informative for this ancestral gene order.

Information on gene trees and gene order were downloaded from Ensembl v.57 (Flicek et al., 2011; Vilella et al., 2009) for all available genomes (51 species) and used to reconstruct the gene order and orientation at the Boreoeutheria node. The reconstruction method first computes all pairwise comparisons of genomes that are informative for the ancestor of interest (the ancestor is on the path between both species in the tree).

Two modern genomes share a common ancestor, from which they both independently inherit some degree of conserved gene organisation. Locally, this conserved gene organisation can take the form of conserved gene adjacencies, where two genes a1 and b1 in species 1 are adjacent, and their respective orthologs a2 and b2 in species 2 are also adjacent. This definition of conserved gene adjacency may be more or less constrained. For example, it may request that the two adjacent genes also conserved their transcriptional orientation, or it may tolerate that some intervening genes separate the two genes as long as they did not exist in the common ancestor (i.e. lineage specific genes). We designed an algorithm to specifically identify locally conserved gene adjacencies between two genomes that are informative for the ancestral Boreoeutheria. The algorithm first filters the two genomes to only retain genes present in their last common ancestor. It then intersects the two genomes to retain the pairs of adjacent genes in the same transcriptional orientation, which will be considered as potentially ancestral (Figure S1A.1). Finally, it assigns each potential ancestral pair not only to the last common ancestor of the two species, but also to each ancestor along the branches that connect the two genomes, as long as no lineage specific gene has been inserted between the two genes considered. A formal description of the algorithm can be found below.

Algorithm 1 Compare two genomes and extracts a list of conserved adjacent gene pairs

Require: $\mathcal{G}_A$ et $\mathcal{G}_B$: two genomes to compare

$\text{Anc}_0 \leftarrow$ common ancestor of A and B
{comment : filtering of genomes $\mathcal{G}_A$ and $\mathcal{G}_B$}
for all Ancestral species Anc between A and Anc_0 **do**
  $\mathcal{G}_A^{\text{Anc}} \leftarrow \mathcal{G}_A$ filter to only retain genes in $\mathcal{L}_{\text{genes}}^{\text{Anc}}$
  $P_A^{\text{Anc}} \leftarrow \bigcup_{g_1,g_2 \text{ consecutive in } \mathcal{G}_A^{\text{Anc}}} \{(g_1, g_2), (\overline{g_2}, \overline{g_1})\}$
end for
for all Ancestral species Anc between B and Anc_0 **do**
  $\mathcal{G}_B^{\text{Anc}} \leftarrow \mathcal{G}_B$ filter to only retain genes in $\mathcal{L}_{\text{genes}}^{\text{Anc}}$
  $P_B^{\text{Anc}} \leftarrow \bigcup_{g_1,g_2 \text{ consecutive in } \mathcal{G}_B^{\text{Anc}}} \{(g_1, g_2), (\overline{g_2}, \overline{g_1})\}$
end for
{Intersection of gene pairs}
$\mathcal{C}_{\text{Anc}_0} \leftarrow P_A^{\text{Anc}_0} \cap P_B^{\text{Anc}_0}$
{comment : propagation of conserved pairs to intermediate ancestors}
for all Ancestral species Anc between A and Anc_0 **do**
  $\mathcal{C}_{\text{Anc}} \leftarrow \mathcal{C}_{\text{Anc}_0} \cap P_A^{\text{Anc}}$
end for
for all Ancestral species Anc between B and Anc_0 **do**
  $\mathcal{C}_{\text{Anc}} \leftarrow \mathcal{C}_{\text{Anc}_0} \cap P_B^{\text{Anc}}$
end for
return $\mathcal{C}$

Gene adjacencies inferred by Algorithm 1 should all be ancestral if no events have taken place in any lineage. In this case, all genes should be involved in at most two adjacencies (one upstream and one downstream of its own position in the ancestral genome). Rearrangements, gene gains and losses however create situations where a given ancestral gene is involved in additional adjacencies, thus creating a need to decide which adjacency is the most likely to be ancestral. To do this, the method labels each adjacency with a weight reflecting the number of times that it has been reported as conserved since this ancestor through relevant pairwise comparisons of extant genomes according to the species phylogeny (Figure S1A.2). Adjacencies are ranked by decreasing weight, and selected in turn from most to least conserved (Figure S1A.3). By selecting these adjacencies, gene pairs are sequentially classified as ancestral as long as they are not already involved in a conserved adjacency with a higher weight. The following definitions formalise this approach.

Let M be a set of genes. Each gene m is oriented, and thus possesses two ends: one input noted m^- and one output noted m^+. $\overrightarrow{m}$ represents the gene in the orientation (m^+,m^-) while $\overleftarrow{m}$ represents the gene in the alternative orientation (m^-,m^+). The operator $m \mapsto \overline{m}$ inverses the orientation of the gene. The set of gene ends is $\text{M}^{\pm} = \bigcup_{\text{m}\in\text{M}}\{m^-,m^+\}$. The set of genes with both alternative orientations is $\overleftrightarrow{\text{M}} = \bigcup_{\text{m}\in\text{M}}\{\overrightarrow{m},\overleftarrow{m}\}$. An adjacency is a pair of oriented genes $(m_1, m_2) \in \overleftrightarrow{\text{M}}^2$. It represents the non-oriented junction between the output of m_1 and the input of m_2. Such an adjacency is equivalent to $(\overline{m_2},\overline{m_1})$. The set of adjacencies is $\text{A} \subset \overleftrightarrow{\text{M}} \times \overleftrightarrow{\text{M}}$. Each adjacency is labelled with a weight corresponding to the number of pairwise genome comparisons where it is conserved using a valuation function v: $\text{A} \mapsto \mathbb{R}$. Adjacencies are indexed in $\text{A} = \{\text{a}_\text{i}\}_{1\leq i\leq n_a}$ with decreasing weights (the sequence $\text{v}(\text{a}_\text{i}))_{1\leq i\leq n_a}$ is decreasing). In case of equal weights, an arbitrary choice of a_i order is made. Adjacencies are then drawn

from A according to the valuation function v. An adjacency $a_i = (m_1, m_2)$ is classified as being ancestral if the output of m_1 and the input of m_2 are both free (i.e. not already used in an ancestral adjacency of higher weight).

Quality assessment

Using this method, we reconstructed 18,436 gene adjacencies (and therefore, intergenes) in the Boreoeutheria ancestral genome. In order to assess the robustness of the reconstruction, we counted the number of modern genomes where a pair of adjacent genes can be found in the same configuration as the inferred ancestral one: a gene configuration that exists in multiple genomes in different lineages is very likely to be the inherited ancestral configuration. On average, ancestral intergenes are observed in 13.7 descendant genomes out of 28 (SD = 6.2; Figure S2A), 73.5% of intergenes are supported by more than 10 descendant species, and more than 90% of the intergenes can be observed in at least 5 descendant genomes, including at least one in each of the two clades that originated from the Boreoeutheria ancestor (laurasiatherians and euarchontoglires). Of note, 16 out of 28 sequenced Boreoeutherian genomes used in this study are highly fragmented assemblies, often preventing the observation of potentially conserved gene adjacencies (Figure S2B).

Additionally, we compared our reconstruction of the Boreoeutheria ancestral genome to three published reconstructions using different methods (Table S1): a reconstruction by inferCARs (Ma et al., 2006), and two reconstructions by ANGES based on different sets of genomic markers (Jones et al., 2012). Both inferCARs and ANGES differ algorithmically from our method and can only efficiently use a much smaller number of genomic markers (typically 1000-5000). These reconstructions are therefore based on genomic regions over 100 kb in length where synteny is generally conserved across genomes, ignoring shorter non-alignable regions or inversions within these regions. In this respect, our method greatly improves upon previous reconstructions in terms of resolution with over 20,000 markers of median length 30 kb (Table S1). In terms of continuity, our reconstruction contains 18,436 gene adjacencies resulting in 133 contiguous ancestral regions (CARs). This is slightly less contiguous than previous reconstructions (17 to 29 CARs, which are thought to correspond to full ancestral chromosomes), but remarkably high in regard of the increase in resolution. Ends of CARs typically correspond to genes either highly duplicated or lost through many lineages, where reconstruction of the ancestral gene order is particularly difficult. Our ancestral Boreoeutheria genome was in high agreement with previous, lower resolution reconstructions, as despite this lower continuity, 86.1%-91.5% of marker adjacencies in previous reconstructions were successfully recovered in ours.

Estimation of ancestral intergene lengths

To estimate the length of the ancestral intergenes, we examined whether adjacent genes conserved in their ancestral configuration in multiple modern genomes are typically separated by intergenes of similar length, which would suggest that the intergene has little changed in size since the ancestor. We computed a table of the lengths of orthologous intergenes across all sequenced modern descendants of the Boreoeutheria ancestor (28 species in Ensembl v.57), and tested for overall correlation across values. The correlation of orthologous intergene lengths in modern genomes is remarkably high: for most ancestral intergenes, modern lengths group tightly around the median value (R^2 = 0.86; Figure S3A). When modern intergene lengths are randomly shuffled within their respective genomes, the overall correlation of

orthologous intergene lengths with their median value is drastically lower (R^2 = 0.04; Figure S3B).

The R^2 coefficient of 0.86 is most parsimoniously explained by a median value that is directly informative about the ancestral state. This is consistent with an evolutionary process that randomly inserts or deletes DNA elements from an ancestral intergene, largely independently in each lineage, thus leading to a length distribution of modern lengths centred on the ancestral value. We conclude that in most cases, the median value of modern orthologous intergene lengths is a reliable estimate of the ancestral length. We retained these ancestral estimates only when they are supported by at least two genomes sequenced with coverages of 6x or more from different clades (primates, rodents and laurasiatherians), to ensure that the estimate is truly ancestral. Intergenes with high variability in modern lengths were not considered in the analysis, as the ancestral estimate was then deemed unreliable; we chose as a cutoff to retain only intergenes for which the interquartile range (range spanned by the 50% of values closest to the median) is no larger than 1.5 times the median value.

Using this approach, we obtained an ancestral length estimate for 16,115 intergenes out of 18,436 (87.4%). Remarkably, the distribution of ancestral intergene lengths is log-normal and very similar to that of a high-quality modern genome (Figure S2C). This demonstrates that both the reconstruction of ancestral intergenes and the estimation of their lengths do not exclude specific categories of intergenes based on their length (especially the longest ones, which may have been rearranged beyond our ability to reconstruct them).

Estimation of ancestral intergenic GC contents

Ancestral intergenic GC contents were estimated in a similar manner to intergene lengths, based on correlation of GC contents in orthologous intergenes across modern descendant species. Like intergene lengths, GC content is highly correlated across orthologous intergenes (R^2 = 0.81, compared to R2 = 0.04 after random shuffling; Figure S3C and S3D). We use the median modern value as an estimate of the ancestral GC content for ancestral intergenes that are supported by at least two genomes sequenced with a coverage of 6x or more in different clades, with a scatter cutoff at 10% (see 2.3). We could thus compute the ancestral GC content for 15,856 ancestral intergenes (86.0%). The ancestral values display a distribution similar to those of high-quality modern genomes (Figure S2D).

Estimation of ancestral conserved non-coding sequence content

We used conserved non-coding elements (CNEs) identified by GERP (Cooper et al., 2005) from an alignment of 33 eutherian genomes, downloaded from Ensembl. Elements that are well conserved between boreoeutherian genomes are most likely ancestral. We inferred the ancestral CNE content in an ancestral intergene as the median of the total CNE length in modern intergenes, similarly to intergene length and GC content (see above). All ancestral intergenes with an estimated length have a CNE content estimate, of which 34% do not contain any conserved non-coding element and an additional 12.5% show high disparities in total CNE length amongst species, suggesting that the multiple alignment may be poor in some of these intergenic regions. However, we retained all estimates for the analysis, based on the observations that (a) the correlation between modern orthologous CNE lengths and their median value is very high over the entire set (R^2 = 0.82; Figure S3E), (b) removing intergenes with uncertain total CNE estimates biases the dataset towards long intergenes, presumably by removing many short intergenes that truly do not contain CNEs, and (c) in any

case, analyses carried out on the remaining 53.5% of intergenes where the estimate is most reliable gave very similar results to those on the entire set (data not shown).

The total CNE lengths in the Boreoeutheria ancestor are skewed towards smaller values compared to genomes sequenced with high coverage, suggesting that the ancestral CNE lengths reconstructed in this way are underestimating the true ancestral constrained sequence content (Figure S2E). This is probably due to the use of the median modern value as an approximation for the ancestral state. Indeed, losses of CNEs or missing sequence data in some of the lineages (especially in low coverage genomes) will tend to bias the median value towards lower values than the true ancestral CNE content. A precise reconstruction of the ancestral CNE content is a complex problem beyond the scope of this work; considering that these values are used to categorize intergenes in only two relative groups of respectively high and low proportion of CNE, the fact that the ancestral CNE estimates are lower bound estimates is not a major concern.

2. Identification of evolutionary rearrangement breakpoints

Computing rearrangement breakpoints

In order to identify intergenes that have been affected by a rearrangement breakpoint, we compared the ancestral Boreoeutheria gene organisation to five extant Boreoeutherian genomes (human, mouse, dog, cow and horse) chosen for the quality of their assembly and annotation. The five lineages have radiated at short time intervals after the Boreoeutheria ancestor (Nery et al., 2012), so that the vast majority of rearrangement events are expected to be independent. Using the reconstructed ancestral gene adjacencies, we concatenated the ancestral genes into blocks of consecutive oriented genes. The ancestral gene order was then compared to the gene order in each of the five modern genomes under study to identify regions in the ancestral genome where the gene order has been modified in subsequent evolution. In such regions, genes have different neighbours in the modern genome compared to their ancestral counterparts, so that the ancestral intergenes have been lost (Figure S1A.4b). The breakpoint regions may be single ancestral intergenes, or runs of consecutive ancestral intergenes that no longer exist in the modern genome. In order to exclude regions where the perturbation of gene order is due solely to gain or loss of genes rather than a true chromosomal rearrangement, we reduced the ancestral and modern genomes to the order of the genes present in 1-to-1 copies in both genomes (excluding new genes, lost genes and duplications). Regions where gene adjacencies remain different between the ancestral and modern genomes correspond to true breakpoints.

In some complex cases, breakpoints could not be mapped to a precise intergene as the region was affected both by a breakpoint and gene gain/loss events (Figure S1B). As it is not possible to date the gain/loss event relatively to the breakpoint, we cannot decide which ancestral gene organisation was relevant when the rearrangement occurred. Such breakpoints, although probably real, were not included in the final dataset as they cannot be processed in the analysis. The distribution of the 751 breakpoints in the five lineages under study is shown in Figure 1B.

False positives filtering

Breakpoints identified through the method described above suffer from a potentially important caveat: any error in reconstructing the ancestral gene order, and any assembly or annotation error in a modern genome, will be identified as a *bona fide* breakpoint if it causes a gene to have different neighbors in the ancestral and modern genomes. Over 200 initially identified breakpoints were manually curated to detect patterns that resulted in false positives in the dataset. As a result, we identified four recurrent cases that could be automatically filtered out to reduce noise to a minimum in the breakpoints dataset.

Poorly supported ancestral gene adjacencies. When an ancestral gene adjacency is reconstructed solely based on genomes below the ancestral node and is not supported by any outgroup genome, there is reasonable grounds to suspect that the ancestral adjacency is erroneous. Conversely, an ancestral adjacency based on mostly outgroups and very few descendant genomes is equally suspect. We consider that breakpoints occurring in ancestral intergenes not supported by at least one outgroup and one high-quality descendant genome (amongst those not used for breakpoint analysis) are false positives due to errors in the

ancestral gene order reconstruction. Such cases generally arise when orthology and paralogy links are poorly resolved between different paralogs in the gene tree, resulting in inconsistencies in the apparent gene order from one genome to the next.

Monoexonic genes. Some breakpoints are the result of the apparent transposition of a single gene. When this gene possesses only one exon (monoexonic gene), however, there is suspicion that the transposition may be due to a retrotransposition event rather than to a true rearrangement event. Such monoexonic genes may also be annotation errors, when a spurious short CDS has been annotated based on low similarity of sequence with proteins present in other genomes. We removed breakpoints from the dataset if they involve a monoexonic gene in the modern genome.

Wrongly annotated 5'-end exons. False positives in the dataset were frequently caused by spurious (or rarely transcribed) exons annotated at the 5'-end of genes, present in only one genome, at a large distance from the next exon shared with other species. When the intron between a spurious 5' exon and the first true exon includes one or several genes, the apparent order of the genes, based on the relative order of the 5'-end of the genes, is modified. When a breakpoint is caused by the apparent transposition of a gene up to four genes upstream or downstream of its ancestral position, in conserved orientation, we consider that the breakpoint is a false positive caused by spurious exons.

Assembly errors in the cow genome. Breakpoint analysis showed that the cow genome contained over a hundred breakpoint events that correspond to single-gene events (a single gene is bordered by breakpoints on both sides), while other genomes contain between 10 (human) and 24 (dog) of such events. Manual curation revealed that in many cases, these genes are on a short contig bordered by large gaps that appear to be misassembled, either in the wrong orientation or simply misplaced. As most of these events appeared to be false positives, we removed them from the dataset. It should be noted, however, that including this category of breakpoints in the analysis did not significantly alter the results, showing that the analysis is robust to the inclusion of false positives in the dataset.

Comparison to previously reported breakpoints

The breakpoints set was compared to the independently obtained breakpoint regions published in (Larkin et al., 2009). Larkin's breakpoint regions (LBR) correspond to the human (hg18) genomic coordinates of 433 regions of discontinuity between the human genome and 10 other mammalian genomes. The rearrangement event is attributed to one of the lineages by comparing with the remainder of the species set. We used the liftover function included in Galaxy (Giardine et al., 2005) to map all LBR corresponding to breakpoints in the human, cow, mouse and dog genomes to the hg19 version of the human genome. We collected the human genes found within each LBR (plus two genes up- and downstream, to account for differences in gene limits between versions of the human genome), and tested whether the ancestral copies of these genes include the borders of a breakpoint from our set in the corresponding lineage. When such an overlap between breakpoints was found, we considered that they correspond to the same breakpoint region.

Sixty percent of the LBR are included in our breakpoints set (see Table S2). Of the remaining 40%, only the human LBR could be rigorously investigated in detail. 21 human LBR (26%) had no correspondences in the hg19 assembly and correspond to assembly errors in the hg18 version of the human genome rather than *bona fide* breakpoints. The additional 15

human LBR with no match in our dataset were manually inspected and correspond to regions of complex history (Figure S1B), which we had eliminated from our analysis. The LBR in other genomes were provided as orthologous human coordinates, but no information was provided on the alignments and mapping algorithms used. It was therefore impossible to obtain the original LBR coordinates in each mammalian genome, and thoroughly investigate the discrepancies between our and Larkin *et al*.'s datasets as above. However, we used a synteny browser to manually investigate a large number of cases. About half of the LBR with no match in our dataset are not orthologous to a breakpoint in the reported species in the current genome assembly. These LBR most likely correspond to assembly errors that have been corrected since Larkin *et al*.'s work. The other half were regions of complex history which were removed from our analysis. This repartition was consistent with the distribution of discrepancies between datasets observed for human breakpoints.

Of note, the LBR are relatively large regions of the human genome (mean: 700 kb, 9.7 genes), and several LBR overlap more than one breakpoint in our intergene-scale dataset.

3. Poisson regression analysis

Regression model

In order to test whether the three parameters estimated in the ancestor (intergene length, GC content, CNE content) are significantly correlated with breakage, we used Poisson regression, a generalized linear regression method that models the distribution of rare events (here, breakpoints) in a set of intervals (here, intergenes) according to characteristics of these intervals. Importantly, Poisson regression relies on the fact that, after a logarithmic transformation, most rare events occurrence rates can be appropriately modelled as a weighted sum of explicative parameters with Poisson-distributed errors. Throughout the study, and for the sake of simplicity, we refer to the expected value of the breakage rate R for a given class of intergenes x (rigorously, $E(R|x)$) as the "breakage rate", noted r. The multivariate regression analysis was carried out in R (http://www.R-project.org/) using the generalized linear models implemented in the *glm()* function.

The null hypothesis is that breakpoints are distributed randomly, in direct proportion to the size of intergenes, with a proportionality coefficient equal to the average number of breakpoints per intergenic base pair (total number of breakpoints divided by the total intergenic length). To test this, let r be the breakage rate (mean number of breakpoints per intergene), and L be the mean length of a class of intergenes. If breakage is random, we expect:

$$r = a \,.\, L \quad \Leftrightarrow \quad log(r) = log(L) + \alpha$$

Therefore, in a log-log representation, we expect the breakage rate per intergene to be a linear function of intergene length, with $x = y$. If other factors influence breakage, under the assumptions of Poisson regression, $log(r)$ will typically be a linear function of both $log(L)$ and these parameters. For example, if %GC is the percentage of guanine and cytosine bases in intergenes, and if %CNE is the percentage of conserved non coding elements in intergenes, the Poisson regression used to test their influence on breakage rate would look like:

$$log(r) = \alpha \,.\, log(L) + \beta \,.\, \%GC + \gamma \,.\, \%CNE + \delta \quad \Leftrightarrow \quad r = d \,.\, L^{\alpha} \,.\, e^{\beta \%GC} \,.\, e^{\gamma CNE}$$

Stepwise regression analysis

Intergenes were divided into classes of similar length (bins of width 0.5 in log scale), then further into classes of GC content (bins of width 0.2) or into top 50% and lower 50% according to the proportion of conserved non-coding elements. The mean value of each parameter was used as the predictor value for the class of intergenes in the regression. A stepwise regression procedure was carried out to progressively add new variables in the model, in an order based on their initial performance in explaining the data when used as a single regression parameter (intergene length, then GC content and proportion of CNEs).

The goodness of fit at each step of the model was estimated in two different ways. Firstly, a Chi^2 test was performed on the residual deviance and degrees of freedom of the model, to compare the regression model to a saturated model with no deviance and no degrees of freedom (perfect fit). When the test is not significant ($P > 0.05$), then variations between the model and the data can be attributed to statistical noise. Secondly, the proportion of deviance

accounted for by the model was calculated using McFadden's pseudo-R^2.

At each step of the regression procedure, the new parameter was retained in the model when a Chi^2 test on the difference in residual deviances with and without the parameter (with one degree of freedom) was significant. Of note, this was always in agreement with Akaike's Information Criterion, so no issues of over-fitting arose.

4. Breakpoints distribution in yeast genomes

Ancestral genome reconstruction

We tested whether the relationship between breakage rates and ancestral intergene lengths is specific to mammals or can be also uncovered in very distant eukaryotic genomes such as yeasts. Gene annotations were obtained from the Génolevures database for 11 yeast species (Sherman et al., 2009). Gene trees were built using TreeBeST (Vilella et al., 2009). The gene order in the last common ancestor of *Kluyveromyces* and *Lachancea* yeasts, usually referred to as "Ancestor I", was reconstructed as described in Supplementary Material 2. The reconstructed ancestral genome contains 4,608 gene-to-gene adjacencies. The lengths of the intergenic spacers were estimated as described in Supplementary Material 2, based on the 5 sequenced descendants of Ancestor I. The correlation between modern and ancestral intergene lengths is shown in Figure S2F.

Correlation between intergene lengths and breakage rates

Rearrangement breakpoints that occurred in the *Lachancea kluyveri*, *Lachancea waltii*, and *Kluyveromyces lactis* lineages were obtained as described for mammals in Supplementary Material 3. We identified 505 breakpoints (Figure 1D), a finding consistent with a previous report in these lineages (Gordon et al., 2009).

Poisson regression was carried out as previously described to model the correlation between the mean number of breakpoints per intergene and intergene length. Strikingly, like in mammals, breakage rate in yeasts correlates almost perfectly with intergene length in log space (Figure 2E; Chi^2 test, $P = 0.18$; McFadden's pseudo-$R^2 = 0.76$). The regression equation is:

$$log(r) = 0.44 \,.\, log(L) - 4.94 \quad \Leftrightarrow \quad r = 7.2 \,.\, 10^{-3} \,.\, L^{0.44}$$

The fact that evolutionary breakpoints are distributed in intergenes according to their lengths, in two eukaryotic groups as distantly related as yeasts and mammals, argues in favor of a mechanism linked to the general structure of eukaryotic genomes rather than linked to more recently acquired features under selection.

It should be noted that the exact value of the power law linking intergene length and breakage rate is different between yeasts (0.44) and mammals (0.28). In the light of the conclusions of this work, this may reflect the differences in coding to non-coding DNA ratios between yeasts and mammals: yeast genomes are much more compact and gene-dense than mammalian genomes, which may result in a higher proportion of non-coding chromatin in an open state on average, to complete regulatory functions. Our model predicts that the power law parameter will increase as the proportion of non-coding chromatin in an open state increases. In the ideal case where the entire genome is open, breakage rates would be strictly proportional to intergene lengths, and the power law parameter would equal 1.

Of note, these results are not directly consistent with a previous model of breakpoint occurrence in yeast genomes, which concluded that intergene length is only weakly correlated with breakage probability (Poyatos and Hurst, 2007). The latter study used logistic regression to model breakage probability, a method that deals with binary variables (in this case, presence or absence of orthologous intergenes in pairwise comparisons of yeast genomes).

Unlike Poisson regression, logistic regression does not offer a straightforward expectation for the relationship between interval length and event probability under a random distribution, and does not imply that the correlation between both should be sought in logarithmic space. This likely explains why the near perfect power law relationship between breakage rates and intergene lengths was not observed using a logistic regression.

5. Correlation of genomic features with intergene length in the human genome

Non-coding features of the human genome were screened for candidates to explain the power law linking ancestral intergene lengths and the mean number of breakpoints per intergene (breakage rate). This power law relationship implies that the breakpoint density, expressed in breakpoints per Mb, is not constant along the genome, as expected at random, but is high in gene-dense regions and decreases when intergenic distances increase. Appropriate candidate features are expected to display a similar distribution to that of breakpoints, i.e. their density per Mb should decrease sharply when intergene lengths increase (Figure 2B).

Repeated elements

Repeated elements exist in up to thousands of almost identical copies in mammalian genomes and have been proposed as a possible cause of rearrangements through non-allelic homologous recombination. The genomic coordinates and annotations of repeated elements identified by RepeatMasker (http://www.repeatmasker.org) in the human genome (hg19) were downloaded from the UCSC Genome Browser (Fujita et al., 2011).

Segmental duplications, or low-copy repeats (LCR), are not identified by RepeatMasker and were handled separately from other repeated elements. The coordinates of segmental duplications longer than 1 kb and over 90% identity in the human genome were downloaded from the UCSC Genome Browser.

Recombination rate

Measures of the local recombination rate in the human genome (hg18) were obtained from the HapMap Project phase II (http://hapmap.ncbi.nlm.nih.gov/downloads/). The recombination map was updated by transferring the coordinates of the markers to the hg19 version of the human genome using the liftover tool on the Galaxy server (Giardine et al., 2005). The recombination rate in a given intergene of the human genome was calculated as the mean of the local recombination rates at all markers belonging to this intergene. 7084 intergenes do not contain any marker and were excluded from the analysis.

Replication origins

The coordinates of computationally predicted replication origins in the human genome, described in (Huvet et al., 2007), were communicated by the authors. The set of 874 replication origins used in this work is a refined and updated version of the set originally published in 2007. The density of replication origins in intergenes was found to decrease as intergene lengths increase, reminiscent of the pattern observed for breakpoints (Figure 2B). However, this appears to be a secondary consequence of a similar correlation with intergene length, rather than an indication that replication origins cause rearrangements. Replication origins and timing are conserved features in mammalian genomes (Ryba et al., 2010), allowing us to map the replication origins in the Boreoeutheria ancestral genome. Although their small number precluded their use in the regression analysis, we were able to test whether the locations of ancestral replication origins significantly co-occur with breakpoints, i.e.

whether they are observed in the same intergenes more often than expected at random. We found that 5.1% of breakpoints co-occur with a replication origin. This overlap is statistically significant when all intergenes are considered at once (Fisher's exact test, $P = 0.01$). However, when intergenes of homogeneous length are considered, the co-occurrence is not significant (intergenes < 20 kb: $P = 0.29$; intergenes > 20 kb: $P = 0.13$). This is evidence of a secondary correlation, rather than of a dependency between replication origins and rearrangement breakpoints. Interestingly, both breakpoints and replication origins independently follow distributions related to intergene length, and could be independent consequences of the same genomic feature. Indeed, replication origins are specified by an open chromatin signature (Audit et al., 2009), which is consistent with our conclusions that rearrangements occur between open chromatin regions in contact in the nucleus (although not specifically between regions containing replication origins).

Topological domains boundaries

Topological domains are highly self-interacting domains of the genome that result from the 3D organisation of chromosomes in the nucleus. We used the coordinates of 2267 boundaries of topological domains in human ES cells and 2117 boundaries in mouse ES cells described in (Dixon et al., 2012). Limits of topological domains are roughly conserved in location between human and mouse, and may delineate domains that have functional relevance. As such, rearrangement breakpoints may be preferentially retained when they occur at topological domain boundaries, since they would reorganize the order of large functional domains without altering their internal organization.

We found that topological domains boundaries occur more frequently in gene-dense regions, confirming the authors' original results (Figure 2B). However, while 53.8%-75.9% of topological boundaries are shared between human and mouse at an average resolution (within 200 kb of one another), only 643 occur in orthologous intergenes in both genomes (with a 20 kb incertitude range, which is the resolution of the boundaries mapping in the original study). While these 643 boundaries are likely conserved and inherited from the Boreoeutheria ancestor, the other boundaries have experienced turn-over in one or both lineages, suggesting that domains are more plastic than initially thought. We mapped the conserved boundaries on the ancestral Boreoeutheria genome to test whether rearrangements breakpoints significantly co-occur with topological domain boundaries. Only 8.1% of breakpoints co-occur with a topological domain boundary, suggesting that boundaries are not a major determinant of the distribution of breakpoints. The overlap of breakpoints and boundaries is not statistically significant (Fisher's exact test, $P = 0.08$), and the marginal co-occurrence is most likely a secondary consequence, again, of preferential occurrences in open chromatin regions.

Open chromatin regions

The genomic coordinates of open chromatin regions in four human cell lines (HUVEC, hESC, HeLaS3 and GM12878), as defined by the ENCODE project (Dunham et al., 2012), were downloaded from the UCSC Genome Browser (Regulation group – Open Chrom Synth track – Syn Pk tables; http://genome.ucsc.edu/cgi-bin/hgTables?command=start). The correlation between proportions of open chromatin and intergene lengths in HUVEC cells is reported in the main text while the results for other cell lines are reported in Figure S4B.

6. Breakage simulations in the human genome

Random inversions

Although the distribution of inversion lengths in mammalian genomes is unknown, recent reports have shown that rearrangements preferentially occur between loci that are in physical contact due to chromosome folding in the nucleus (Véron et al., 2011; Zhang et al., 2012). We used the average probability of intrachromosomal contact between any two loci according to their genomic distance *d*, described in (Lieberman-Aiden et al., 2009), as an approximation for the probability of inversion.

Random inversions were simulated by drawing an intergenic base at random in the human genome, considered as the first breakpoint. The second breakpoint was chosen at a distance *d* either up- or downstream on the chromosome according to the probability of contact ($d \geq 1$kb). If both breakpoints fall into the same intergene (no perturbation of the gene order), the breakpoints are considered undetectable by our method. If the space between both breakpoints encompasses at least one gene, the breakpoints are detectable, and so on until the number of visible breakpoints is the same as observed between Boreoeutheria and the five lineages under study. The length distribution of visible and invisible inversions across 100 iterations of the simulation is displayed in Figure 2B. The simulation produces a large number of short inversions (<10 kb), most of which are not visible based on gene order alone. The simulation was performed using a custom-made Python script.

Inversions with breakpoints constrained to interspersed features

We tested for a possible additional effect of interspersed features in promoting inversions between specific loci in contact in the nucleus. The best (and only) candidate feature that may explain the distribution of breakpoints, according to the correlation analysis (see Supplementary Material 6), is the local proportion of open chromatin. Inversions were simulated as described above, except that inversions were recorded only if both breakpoints fall into open chromatin regions. This approximates a situation where either double-strand breaks or misrepairs are much more likely to occur in open chromatin regions, so that rearrangements occur almost exclusively between open chromatin regions. This simulation was performed using the open chromatin patterns of either one of four cell lines (HUVEC, hESC, GM12878 and HeLaS3; Figure 2D and S12) obtained by the ENCODE consortium (Dunham et al., 2012). The expected breakage rates according to intergene length are reported in Figure 2F and S10C.

We also performed similar simulations by constraining breakpoints to identical repeated sequences instead of open chromatin. This experiment serves two purposes: it firstly acts as an additional negative control to exclude that the distribution of breakpoints obtained considering open chromatin patterns would also be expected from any interspersed intergenic feature. Secondly, it formally excludes that the biased distribution of breakpoints result from recombination between repeated elements in contact in the nucleus. This simulation was performed by selecting breakpoints either when they occur in repeated elements of the same class (SINEs, LINEs, LTRs, or DNA transposons; Figure 2E), or strictly of the same type, for which similarity is maximal (AluY, MIRb, L1M4 and so forth; Figure S4D).

Characteristics of simulated breakpoints

We collected the genomic features of 100 kb windows centered on simulated breakpoints. These regions were compared to control windows centered on randomly selected intergenic bases ("random control"). This set of control regions reflects the expectations of the random distribution model while correcting for the fact that the possible locations of our simulated breakpoints exclude genes, ensuring that any difference observed between the breakpoint and control regions are not an artefact due to the exclusion of genes in our breakage model. Features were then tested for enrichment or depletion in simulated breakpoint regions using Wilcoxon's rank sum test.

Supplementary Tables

Method	# Markers	Average marker length (± sd)	Orientation aware	Loss/gain aware	Duplication aware	# Contiguous ancestral regions	# Adjacencies	# Recovered adjacencies
InferCARs (Ma et al. 2006)	1,338	2.0 Mb (± 3.4)	Yes	No	No	29	864	777 (89.9%)
ANGES 1 (Jones et al. 2012)	773	2.1 Mb (± 2.6)	No[a]	No[a]	No[a]	26	752	688 (91.5%)
ANGES 2 (Jones et al. 2012)	2,650	464.3 kb (± 553.0)	No[a]	No[a]	No[a]	17	1645	1416 (86.1%)
New method	22,184	31.4 kb (± 64.9)	Yes	Yes	Yes	133	18,436	-

[a] This option is available but was not used for the published (optimal) reconstruction of the Boreoeutheria ancestor

Table S1. Refers to Figure 1B and Experimental Procedure "*Ancestral genome reconstructions and estimation of ancestral genomic features"*. Comparison between the Boreoeutheria ancestral genome (this study, "New method") and previously published reconstructions using inferCARs (Ma et al., 2006) and ANGES (Jones et al., 2012). Two different reconstructions with ANGES were considered: one based on conserved genomic markers by (Ouangraoua et al., 2011)("ANGES 1"), and one based on conserved genomic markers by (Gavranovic et al., 2011)("ANGES 2"). The table contains the number of markers used for the reconstruction; their average length in the human genome (and standard deviation); whether the reconstruction method takes into account the orientation of the markers; whether the method handles markers that are present only in a subset of the species due to gains or losses; whether the method handles markers that are duplicated in a subset of the species; the number of contiguous ancestral regions (CARs) in the reconstructed Boreoeutheria genome; the number of marker adjacencies in the reconstructed Boreoeutheria genome; and the number of marker adjacencies that are successfully recovered in our reconstruction. Non-recovered adjacencies can result from errors in either reconstruction, or loss of continuity in ours (the extremities of the markers are on different CARs in our reconstruction).

	Human	Mouse	Dog	Cow	Laurasiatherians	**Total**
Number of LBR	81	166	75	99	12	**433**
After liftover to hg19	60	165	75	96	12	**408**
LBR overlapping breakpoints						
Counts	45	98	39	57	7	**246**
Percents	75%	59%	52%	59%	58%	**60%**
Number of breakpoints	100	176	99	289	20	**684**
Number of breakpoints in LBR						
Counts	58	114	41	62	13	**288**
Percents	58%	65%	41%	21%	65%	**38%**
Average per LBR	1.29	1.16	1.05	1.09	1.86	**1.17**

Table S2. Refers to Figure 1B and Experimental Procedures "*Identification of evolutionary rearrangement breakpoints"***.** Overlap between the set of breakpoints identified in this study ("breakpoints") and the evolutionary breakpoint regions identified in Larkin et al., 2009 ("LBR"). When updating the genomic coordinates of the LBR from hg18 to hg19 (using the liftover utility of the Galaxy server), 21 human LBR out of 81 (26%) no longer exist in the current version of the human genome assembly, i.e. likely correspond to assembly errors in the previous version. Manual curation of LBR in the mouse, dog and cow genomes indicate that roughly the same proportion of LBR were assembly errors in these genomes as well.

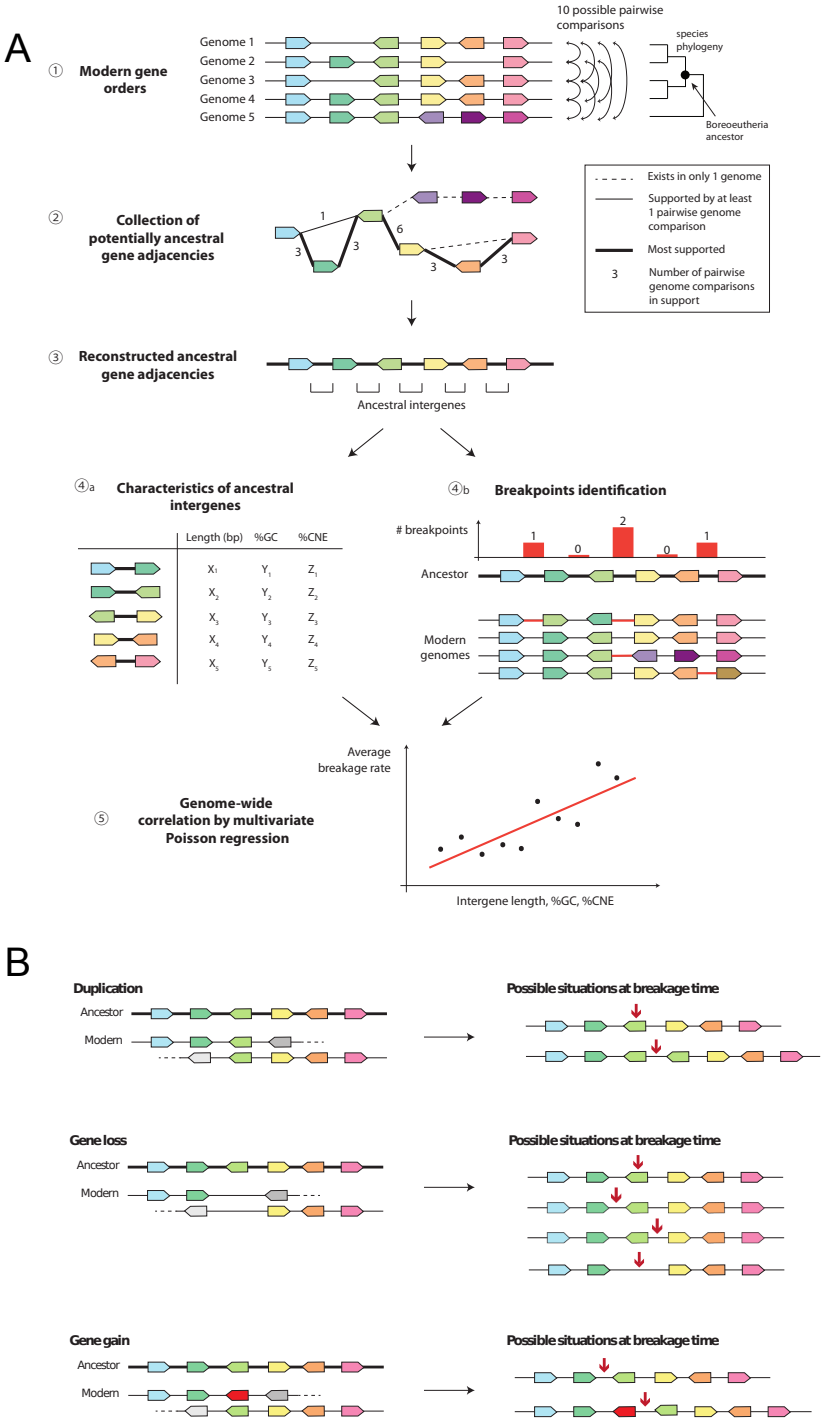


Figure S1. Refers to Figure 1A. (A) Outline of the regression analysis. (1) Genome comparisons are performed for all pairs of species that are informative for the ancestral gene order (i.e. their common ancestor is the same as, or predates, the ancestor targeted in the reconstruction), to detect gene adjacencies that are identical in both genomes (ancestral). (2) All gene adjacencies identified in modern genomes are collected in a graph where nodes are ancestral genes and links are conserved adjacencies, weighted by the number of pairwise comparisons that report their conservation. (3) Ancestral adjacencies are selected from this collection under the reasoning that the best supported adjacencies are ancestral. (4a) The ancestral gene adjacencies define ancestral intergenes, with characteristics (length, %GC, %CNE) that can be robustly estimated by parsimony from modern values. (4b) Meanwhile, the reconstructed ancestral gene order is compared to independent descendant genomes to count the number of times each intergene has been affected by a rearrangement breakpoint later on during evolution. (5) The ancestral characteristics of intergenes and their breakage rate are then correlated using multivariate Poisson regression to quantify their contribution, if any, to breakage probability **(B)** Complex breakpoint regions. When a breakpoint region has been affected by another genic event (breakpoint and duplication of a gene, breakpoint and gene loss, breakpoint and new gene gain), it is not possible to assess with certainty the gene configuration at the time of the rearrangement event. In case 1, the original ancestral gene order is now distributed over two loci on the same or on different chromosomes, and a duplicated gene (light green) is now found in each locus. At the time of the breakpoint, the ancestral light green gene may have been duplicated because of the breakpoint (through breakage and repair) or the duplication may have just preceded the breakpoint. In any case, the ancestral pre-rearrangement intergene is ambiguous. In case 2, a gene loss coincides with the loss of an adjacency. The ancestral rearrangement may have affected the gene itself (leading to loss), any of the two flanking intergenes, or have occurred after the gene loss event. In case 3, the red gene represents a gene that did not exist in the ancestral genome. It may have genuinely appeared in the lineage that follows, or more often correspond to an annotation artefact. In either case it is not possible to ascertain the ancestral intergene prior to the breakpoint

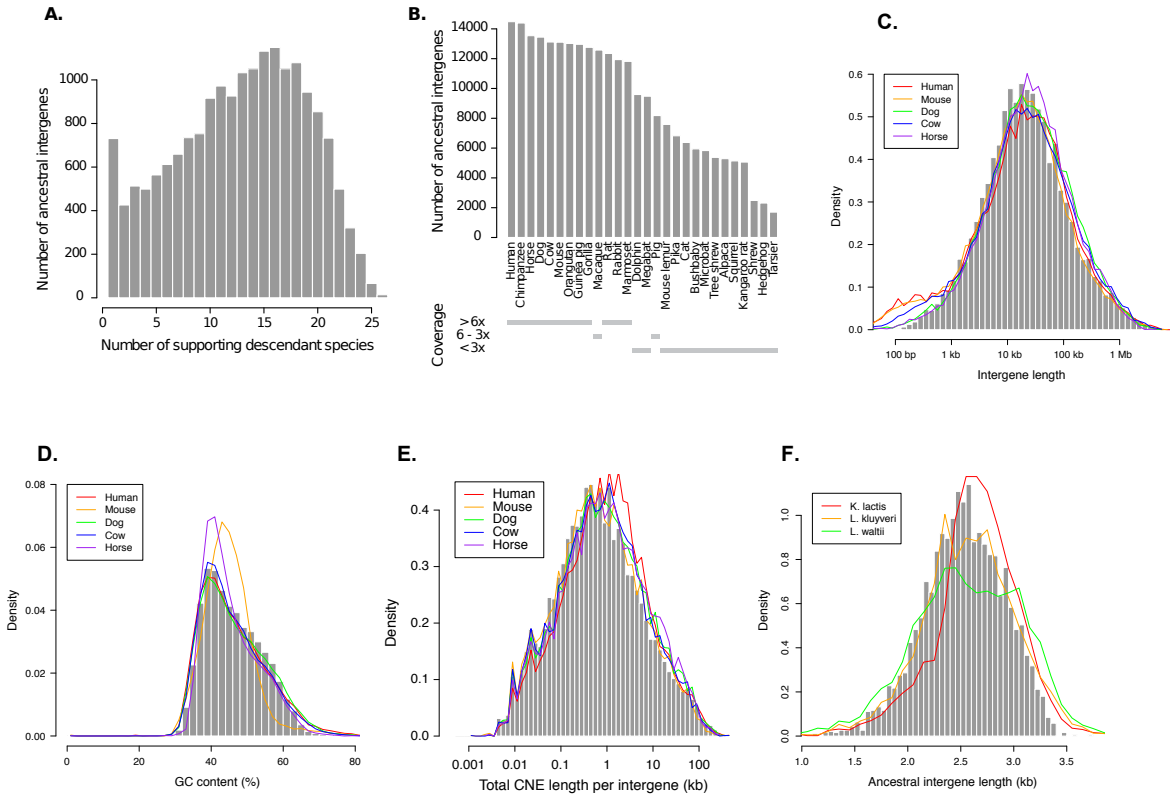


Figure S2. Refers to Experimental Procedures "*Ancestral genome reconstructions and estimation of ancestral genomic features",*

Quality assessment of reconstructed ancestral gene order in the Boreoeutheria genome. **(A)** Distribution of the support scores of ancestral intergenes, defined as the number of modern boreoeutherian genomes in which each ancestral intergene is conserved. **(B)** Contribution of each modern boreoeutherian genome to the ancestral reconstruction. The histogram represents the number of ancestral gene adjacencies observed in each modern genome. Genomes sequenced with a low coverage (< 3x) are available as highly fragmented assemblies, and are less informative than high-coverage genomes for the ancestral gene order reconstruction. **(C)** Distribution of ancestral intergene lengths estimates (in grey) and modern intergene lengths in five high-quality descendant genomes. **(D)** Distribution of ancestral intergenic %GC estimates (in grey) and modern intergenic %GC in five high-quality descendant genomes. **(E)** Distribution of ancestral intergenic CNE content estimates (in grey) and modern intergenic CNE contents in five high-quality descendant genomes. **(F)** Distribution of ancestral intergene lengths estimates (in grey) and modern intergene lengths in the three modern genomes used for breakpoint analysis.

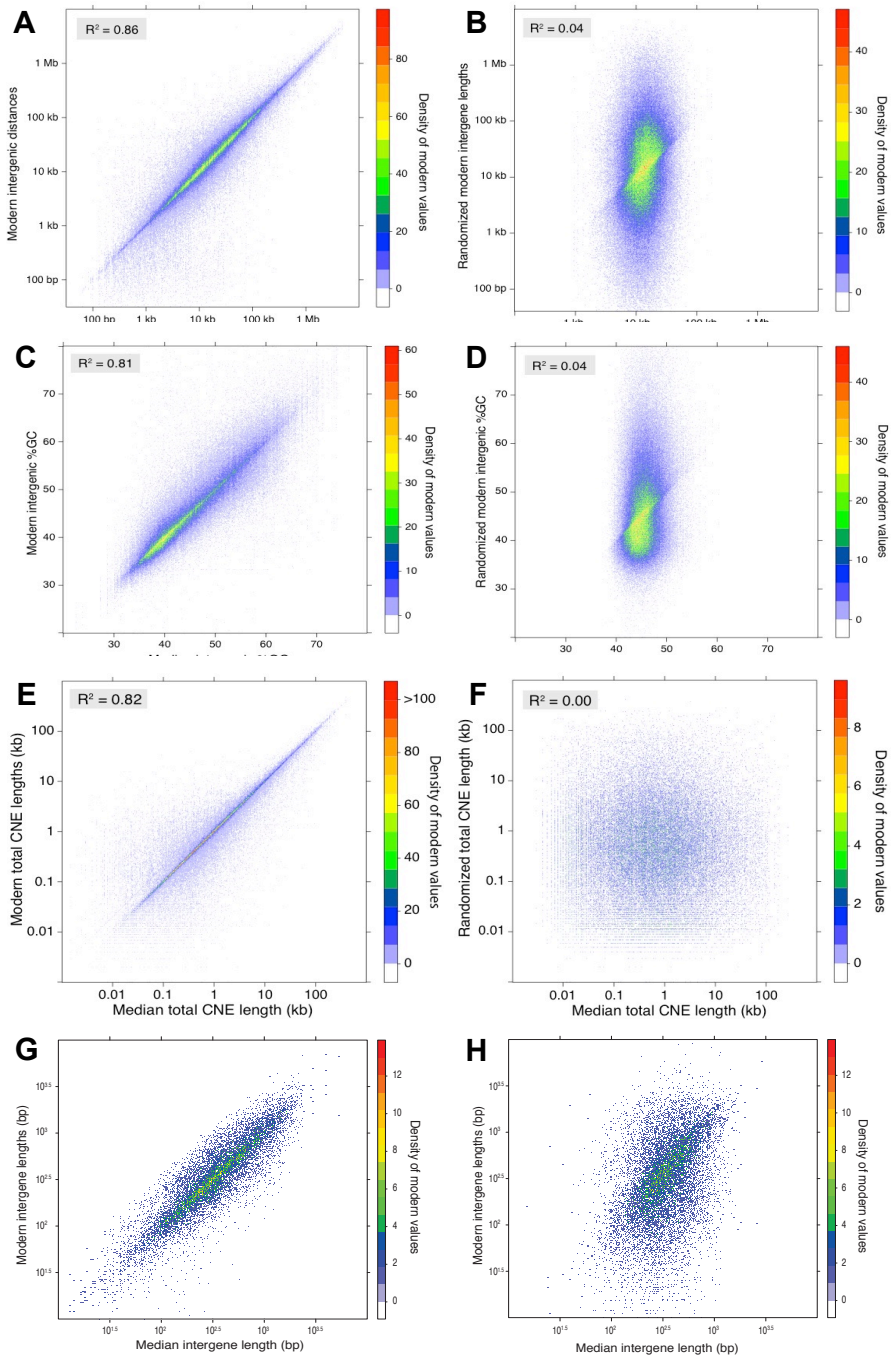


Figure S3. Refers to Experimental Procedures "*Ancestral genome reconstructions and estimation of ancestral genomic features",* Reliability of ancestral feature estimations. (A) Correlation of lengths across orthologous mammalian intergenes. The lengths of orthologous modern intergenes are plotted against their median value, used as an estimate for the ancestral value. Datapoint density is accounted for by the color scale on the right (in bins of width 0.01 in log scale on both axes). (B) Correlation expected from randomized values. The intergene lengths shown in A were randomly shuffled within each genome independently. (C) Correlation of %GC across orthologous mammalian intergenes. The %GC of orthologous modern intergenes are plotted against their median value, used as an estimate for the ancestral value. Datapoint density is accounted for by the color scale on the right (in bins of 0.1% on both axes). (D) Correlation expected from randomized values. The intergenic %GC shown in D were randomly shuffled within each genome independently. (E) Correlation of modern content in CNE across orthologous mammalian intergenes. The total CNE lengths of orthologous modern intergenes are plotted against their median value, used as an estimate for the ancestral value. Datapoint density is accounted for by the color scale on the right (in bins of width 0.01 in log scale on both axes). (F) Correlation expected from randomized values. The intergenic CNE contents shown in E were randomly shuffled within each genome independently. (G) The lengths of orthologous modern yeast intergenes are plotted against their median value, used as an estimate for the ancestral value. Datapoint density is accounted for by the color scale on the right (in bins of width 0.01 in log scale on both axes). (H) Correlation expected from randomized values. The intergene lengths shown in G were randomly shuffled within each yeast genome independently.

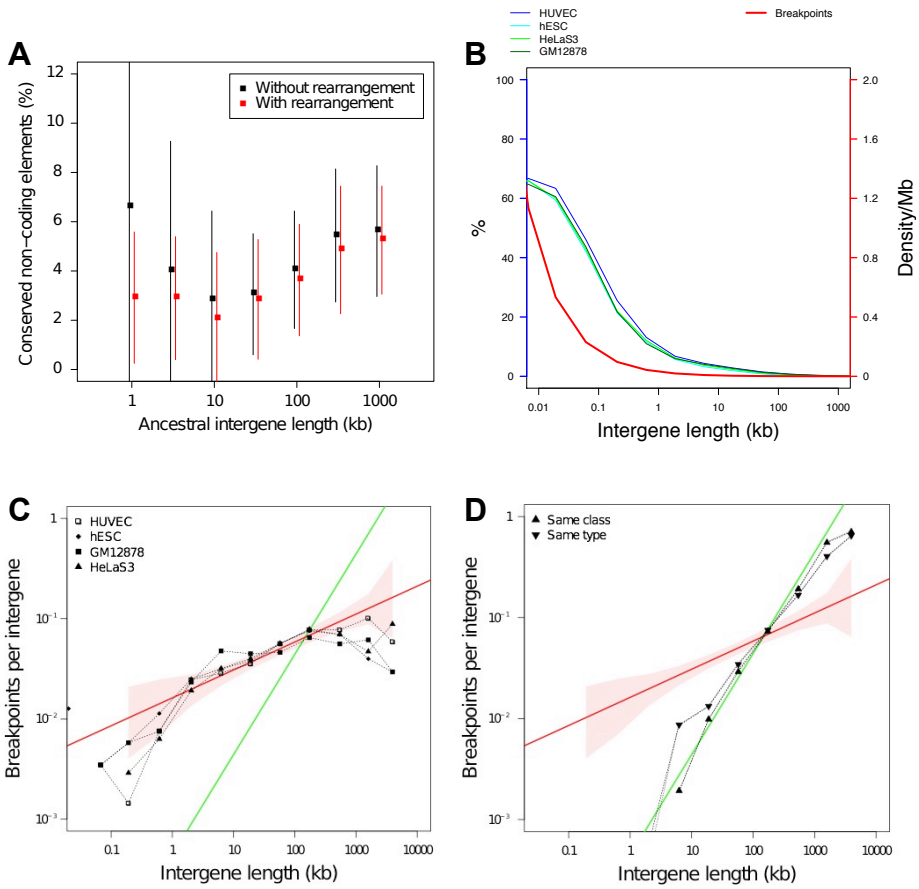


Figure S4. Refers to Figure 2. (A) Intergenes without rearrangement breakpoints contain a higher density of conserved non-coding elements than those with rearrangement breakpoints, in all classes of intergene length. (B) Mean proportion of open chromatin in intergenes as a function of their length in four different cell lines (HUVEC, hESC, GM12878 and HeLaS3). Although the individual regions in an open chromatin state differ across cell lines, the general profile is remarkably similar, with a high proportion of open chromatin where intergenes are small and decreasing proportions as intergenes become larger. This trend is similar to the density of breakpoints (shown in red, and referring to the red axis on the right). Values on the right axis should be multiplied by 10. (C) Breakage rate as a function of intergene length in simulated rearrangements between open chromatin regions in 3D contact in the nucleus is four different cell lines (HUVEC, hESC, GM12878, HeLaS3). While the particular genomic regions in an open state may differ across cell lines, the pattern obtained is remarkably similar in all cases and results in a power law correlation between intergene length and mean number of breakpoints. This power law reproduces exactly the observations made on real breakpoints (red line: regression equation; shaded red area: 95% confidence interval of the model; green line: random expectations). (D) Breakage rate as a function of intergene length in simulated rearrangements between repeated regions in 3D contact in the nucleus. Rearrangements were considered plausible between any pair of repeated elements of the same class (SINEs, LINEs, LTR, DNA transposons, etc.) or only between pairs of repeated elements strictly of the same type (AluY, MIRb, L1M4, etc.). The distribution of breakpoints is then entirely different from the observations in the real breakpoints data (red line: regression equation; shaded red area: 95% confidence interval of the model) and rather follow the expectations of a random distribution (green line).

Supplementary References

Audit, B., Zaghloul, L., Vaillant, C., Chevereau, G., D' Aubenton-Carafa, Y., Thermes, C., and Arneodo, A. (2009). Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells. Nucleic Acids Res. *37*, 6064–6075.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. Nucleic Acids Res. *39*, D800–806.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39*, D876–882.

Gavranovic, H., Chauve, C., Salse, J., and Tannier, E. (2011). Mapping ancestral genomes with massive gene loss: a matrix sandwich problem. Bioinformatics *27*, i257-265.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. PLoS Genet. *5*, e1000485.

Huvet, M., Nicolay, S., Touchon, M., Audit, B., D' Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. (2007). Human gene organization driven by the coordination of replication and transcription. Genome Res. *17*, 1278–1285.

Larkin, D.M., Pape, G., Donthu, R., Auvil, L., Welge, M., and Lewin, H.A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Research *19*, 770–777.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Muffato, M., Louis, A., Poisnel, C.-E., and Roest Crollius, H. (2010). Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics *26*, 1119–1121.

Nery, M.F., González, D.J., Hoffmann, F.G., and Opazo, J.C. (2012). Resolution of the laurasiatherian phylogeny: evidence from genomic data. Mol. Phylogenet. Evol. *64*, 685–689.

Ouangraoua, A., Tannier, E., and Chauve, C. (2011). Reconstructing the architecture of the ancestral amniote genome. Bioinformatics *27*, 2664-2671.

Poyatos, J.F., and Hurst, L.D. (2007). The determinants of gene order conservation in yeasts. Genome Biol. *8*, R233.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.-L., and Durrens, P. (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res. *37*, D550–554.

Véron, A.S., Lemaitre, C., Gautier, C., Lacroix, V., and Sagot, M.-F. (2011). Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. BMC Genomics *12*, 303.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell.