Revisiting the landscape of evolutionary breakpoints across human genome using multi-way comparison

Golrokh Vitae[1], Amine M. Remita[1], Abdoulaye Baniré Diallo[1],

1 Université du Québec À Montréal

*** diallo.abdoullaye@uqam.ca**

Abstract

Genome rearrangement is one of the major forces driving the processes of the evolution and disease development. The chromosomal position affected by these rearrangements are called breakpoints. The breakpoints occurring during the evolution of species are known to be non randomly distributed. Detecting their landscape and mapping them to genomic features constitute an important features in both comparative and functional genomics. Several studies have attempted to provide such mapping based on pairwise comparison of genes as conservation anchors. With the availability of more accurate multi-way alignments, we design an approach to identify synteny blocks and evolutionary breakpoints based on UCSC 45-way conservation sequence alignments with 12 selected species. The multi-way designed approach with the mild flexibility of presence of selected species, helped to have a better determination of human lineage-specific evolutionary breakpoints. We identified 261,391 human lineage-specific evolutionary breakpoints across the genome and 2,564 dense regions enriched with biological processes involved in adaptive traits such as *response to DNA damage stimulus*, *cellular response to stress* and *metabolic process*. Moreover, we found 230 regions refractory to evolutionary breakpoints that carry genes associated with crucial developmental process such as *organ morphogenesis*, *skeletal system development*, *chordate embryonic development*, *nerve development* and *regulation of biological process*. This initial map of the human genome will help to gain better insight into several studies including developmental studies and cancer rearrangement processes.

Introduction

Genome rearrangement is one of the major forces driving the process of evolution, population diversity and development of diseases including cancers. It happens when DNA breaks in specific positions (breakpoints) and reassembles in a way different from the initial genome conformation and changes the genome landscape [6]. It is now well-accepted that genome rearrangements do not happen randomly along the genome and not all genomic regions are susceptible to such dramatic modifications [7, 25, 35, 36]. Millions of years of evolution and natural selection driven by these structural modifications curved the genome in a way that regions carrying high functional pressure maintained their integrity and could be identified as orthologs with same order and on the same chromosome in a range of relatively close species. Hence, resistance of genomic regions to any major modification implies that these regions harbor functional importance to survival as well as reproduction of species and any modification could have a deleterious effect in the process of natural selection [30].

Studies on the genome synteny have shown that conserved regions are not only significantly enriched in putative regulatory regions [23, 31] but also are associated with transcriptional regulations and developmental processes [31, 41, 46]. On the other hand, are the regions that are more receptive to rearrangements, participating in speciation, adaptation and development of species-specific traits and behavior that are not detrimental to the survival nor reproduction of the species. These break-prone regions or "*breakpoint hotspots*" are also known to carry distinctive functional and structural markers [10]. The identification of such regions are difficult and controversial due to several assumptions. "Comparative genomics relies on the structuring of genomes into syntenic blocks" [19]. However, comparative studies still lack a clear definition of synteny. Since the term "synteny" was introduced, even the criteria that define the synteny conservation vary from one study to another. Also, in most of the methods, to identify synteny, synteny fragments were defined as a set of genes with a conserved order [9, 37, 38, 40]. These issues are well discussed by Ghiurcuta, 2014 [19]. The lack of clear definition, choice of granularity of study (minimum size), and different type of comparison could led to have the high probability of different results generated by same data [19]. One conclusion provided to tackle the drawback of these methods are to use multi sets of homologous markers instead of only genes as anchors as well as extending the pairwise comparison to multi-ways [19]. Moreover, the identification of synteny breaks is affected by the selection of species. This influences the size and position of synteny blocks and their corresponding breakpoints as breakpoint is not a tangible physical entity in a genome; it is a result of an analytical construct based on the comparison of selected genomes [42]. Other than exogenous factors, some biological mechanisms such as non-allelic homologous recombination (NAHR) and non-homologous endjoining (NHEJ) [20] and the presence of endogenous factors such as CpG islands [10], GC content[13], some repeat elements [21, 24] and G-quadruplex [10, 39] could affect the susceptibility of genomic regions to rearrangements. Due to some of these factors, DNA could undergo a conformation change (e.g. from B to Z) that could destabilize the molecule and make it prone to structural damage [2, 44, 47]. Also, phenomena such as segmental duplications (SDs) [1, 3, 32], known to be associated evolutionary breakpoints causing a double strand breaks on DNA structure. Hence, it is important to revisit Evolutionary Breakpoints (EBrs) with more accurate and well-drafted mammalian sequenced genomes and updated genomics features. In this paper we propose a new map of accurate and more precise evolutionary breakpoints region based on human lineage synteny breaks, multi-way comparison and ancestral breakpoint reconstruction. Furthermore, we explore the hotspots of EBrs (EBHRs) based on their coassociation to multiple structural and functional human genome features.

Results

0.1 Identification of human lineage-specific synteny breaks

Synteny block is a cluster of (relatively close) markers conserved among related genomes that maintained their orientation and order. The regions bordered by synteny blocks are the position of the genome that have been subjected to rearrangement in one or more compared species since a common ancestor. Multi-way comparative analysis of human genome with 10 other mammals and chicken (see supplementary table S2), revealed 261,420 synteny blocks covering 52% of the human genome with a size distribution from 1 Kbp to 200 Kbp. The corresponding 261,391 genomic regions were identified as human lineage-specific synteny breaks or evolutionary breakpoints (EBrs). 30 break regions were eliminated due to their size higher than 2 Mbp. These regions are whether on centromeric or telomeric regions or on genomic regions with missing

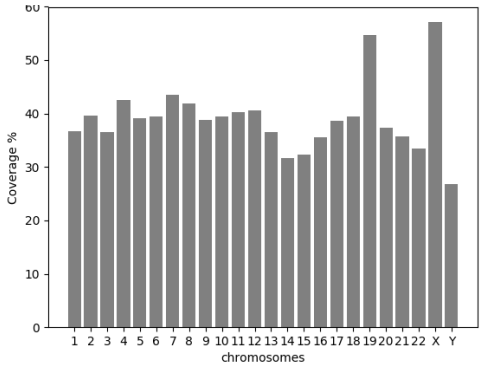

Figure 1. Coverage of chromosomes by EBrs

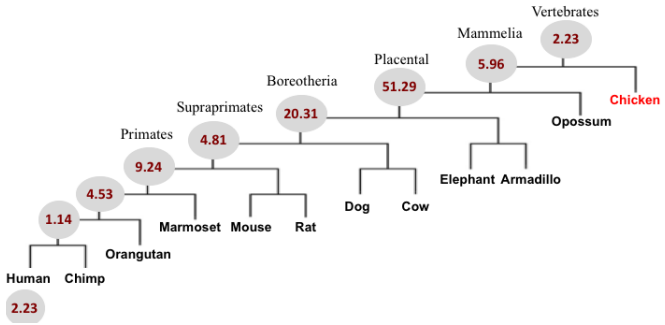

Figure 2. The percentages of human lineage specific breakpoints origin along the vertebrates reference tree

information. See table 1. EBrs are associated with 39.58% of the human genome. However, the coverage of breakpoints raise to over 40% in chr11, chr12, chr8, chr4 and chr7 and to over 55% in chr19 and chrX. See figure 1. The size of these regions were from 1 bp to 1.857 Mbp. The seven largest EBrs (size $>$ 1 Mbp) were located chr14, chrY, chr11, chrX, chr2 and chr7. Based on a Least Common Ancestor approach, we identify the lineage-specific breakpoints, over 50% of these breakpoints were reused and originated from separation of placental mammals. The next most representative ancestral node was the Boreotherian ancestor with 20% of EBrs that originated from. Others were originated from various ancestry nodes along the reference phylogeny tree with a representation between $<$ 1% to about 10%. The distribution of EBr ancestry origins are presented in figure 2.

Method	Input	Output	Size of output
Conservation anchor extraction	45-way multiple alignments	Conservation anchors	21,809,215
Fusion	Conservation anchors	Synteny clusters	346,515
filter according to length	Synteny clusters	Synteny blocks	261,420
Synteny break extraction	Synteny blocks	Synteny breaks	261,421
Size filter	Synteny breaks	EBrs	261,391
Permutation	EBrs	EBHRs	2,564
EBr desert extraction	EBrs + human genome	EBRRs	230

Table 1. Summary of EBr-related results

0.2 Localisation of breakpoint hotspots

Using an overlapping sliding window approach of 100 Kbp and a permutation test with 100,000 iterations, we identified 2,564 regions as EBr hotspot regions (EBHRs). Although evolutionary breakpoints are distributed across 90% of genomic windows, only 8.28% of genomic windows (2,564) were enriched in EBrs (p-value < 0.05). The results show that EBHRs contain 12 to 21 EBrs with an average of 14.85 ± 1.56 (except chromosome Y). EBHRs located on chromosome Y harbor 3 to 14 breakpoints in each window. The EBHRs distribution with respect to chromosomes shows that chromosomes are heterogeneously affected by these regions. Chromosomes are covered by EBHRs from 3% to 23%. Based on the coverage of EBHRs, chromosome were divided into three (3) categories: chromosomes with EBHRs representation under 8%, chromosomes with a coverage between 8% to 15% and chromosomes with an over 15% of EBHRs presence. See supplementary table S3. EBHRs is highly observed in chromosome 22 ($> 23\%$) when Chromosome 1 has the highest number of EBHRs (204 regions). However, it has a moderate coverage of EBHRs due to its size.

We also classified each EBHRs into different groups based on their predicted evolutionary origin according to the origin of the EBRs present in that regionn. The covered orgins are the five major human ancestry nodes along the reference species tree, Placental, Boreotheria, Superprimates and Primates. The results show that about 55% of EBHRs are mostly composed of EBrs originated around separation of human placental ancestor, more than 30% are mostly composed of EBrs originated from different ancestry nodes down to Boreotheria, and 10% of EBHRs were mostly composed of EBrs originated from ancestry node at and after Boreotherian ancestry nodes.

0.3 Biological processes enrichment of genes associated with EBHRs

The GO enrichment analysis of biological process showed that these EBHRs are enriched in genes associated with distinct biological processes compared to genome. Biological process to response to stress: *response to DNA damage stimulus*, *cellular response to stress*, and *DNA repair*; metabolic process of organic and non organic compounds and transformation of chemical substances: *metabolic process*, *nucleic acid metabolic process*, *macromolecule metabolic process*, *nitrogen compound metabolic process* and *primary metabolic process*; The chemical reactions and pathways resulting in the breakdown of a histone protein: *catabolic process* as well as *cellular process* such as in cell communication occurs among more than one cell, but occurs at the cellular level [5]. The list of enriched biological process and the corresponding p-values are provided in supplementary table S1.

0.4 EBHRs associations with genomic features

We mapped EBHR with 25 selected genomic markers known to have positive or negative effect on fragility of the genome. The full description of these markers is included in supplementary table S6. Overall overview of these regions showed that EBHRs have higher coverage of CpG islands (1.3 fold), direct repeats (1.3 fold), SINE (1.4), G-quadruplex (1.4 fold), self chains (1.6 fold), simple repeats (1.9 fold) and SDs (2 folds) compare to the genome. In terms of satellite repeats, the presence of this feature falls in half of EBHRs compare to the genome. The most illustrative coverage comparison of these features are illustrated in figure 3. The complete report on the features coverage are provided in supplementary table S6. It should be noted that although the it seems that 24.9% of EBHRs are covered by genes, however, not all the EBHRs overlap with genes. Half of EBHRs has no overlap with genes. Hence, when

genes coverage is calculated for the other half of EBHRs, the coverage of genes augments by almost 2 times (44.48%).

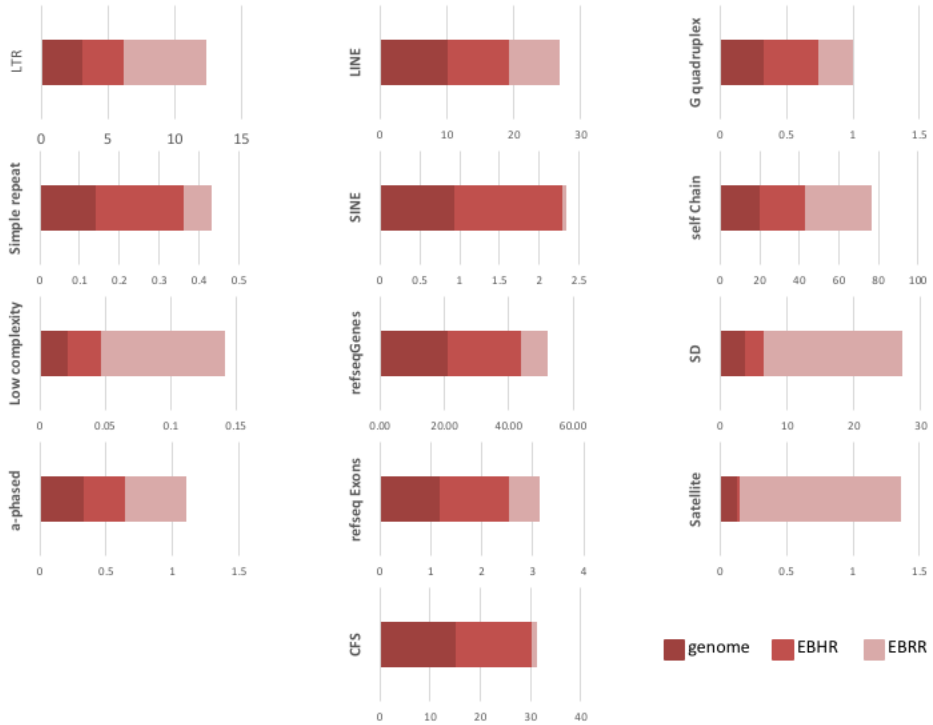

Figure 3. This figure shows the comparison between the coverage of selected genomic features on genome, EBHRs and EBRRs. x-axis represents the coverage percentage of each feature.

0.5 The most resistant genomic regions to EBrs

We found that less than one percent of genomic windows, 230 regions have a complete absence of EBrs. These regions that are mostly continuous, located in 50 genomic locations on all chromosomes. We called these regions evolutionary breakpoint deserts or EBr refractory regions (EBRRs). The 5 largest locations are belonging to chrY and chr19. See figure 4. More than half of EBRRs regions (120 windows) do not overlap with any gene annotations. However, GO enrichment analysis of overlapping genes showed that, these regions are located among genes that are enriched in biological processes completely different from other genomic regions. These regions were highly enriched in critical biological processes involved in development such as *anatomical structure arrangement*, *anterior/posterior pattern formation*, *chordate embryonic development*, *cranial nerve development and morphogenesis*, *embryonic development*, *embryonic morphogenesis*, *embryonic organ development*, *embryonic organ morphogenesis*, *skeletal system morphogenesis*, *facial nerve development*, *nerve development*, *organ morphogenesis* and regulatory processes such as *regulation of biological process*, *regulation of cellular process*, *regulation of gene expression*, *regulation of metabolic process*, *regulation of transcription*, etc. See supplementary table S4 for the complete list of enriched biological process. These genes included thirteen members of Hox-cluster genes that encode conserved transcription factors in most bilateral species [34]. Another interesting group of genes were the 33 members of zinc finger genes located on the 19p12.

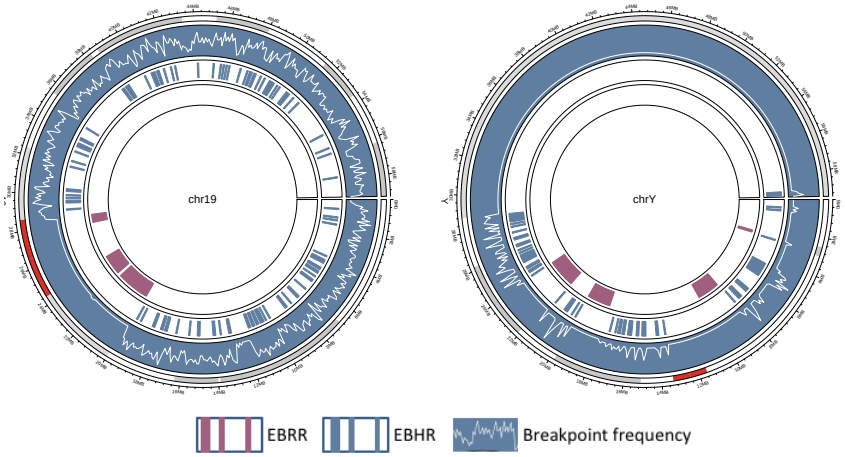

Figure 4. Chromosome 19 showed only one single EBr in a complete chromosome band of 4.4 Mbp. This region codes for 33 members of zinc finger family which is highly conserved in primates. Chromosome Y is one of the most remarkable chromosomes on human genome. The smallest human genome with the highest presence of EBrs, EBHRs and EBRRs.

The other genes were also all conserved genes among mammals or vertebrates. BCL11A, that mutation in this gene is known to be associated with intellectual developmental disorder with persistence of fetal hemoglobin (IDPFH) [33], SRY (sex-determining region Y), known to be associated with 46,XX testicular disorder of sex development and Swyer syndrome [33], FOXP1, known to be associated with Mental retardation with language impairment and autistic features (MRLIAF) [33], TCF4 known to be associated with Pitt-Hopkins syndrome is a condition characterized by intellectual disability and developmental delay, breathing problems, recurrent seizures (epilepsy), and distinctive facial features [33], are some examples of genes and their importance in terms of survival and reproduction. Study of the genomic features in these regions showed that the landmark of these regions are satellite repeats. About 8% of these regions overlap with satellite repeats which is 60 times more than their presence in the genome (0.13%) and 500 times more than their presence in EBHRs (0.02%). Moreover, these regions have higher coverage of low complexity repeats, LTR and a-phased repeats. On the other hand, they show a lower presence of simple repeats, SINE and common fragile sites. See figure 3.

Discussion

0.6 Evaluation of synteny identification

We conducted a multi-way comparative analysis of 12 genomes and identified synteny blocks and their corresponding 261,391 evolutionary breakpoints (EBrs) on the human genome. Using a permutation approach, 559 regions of 100 Kbp were identified as evolutionary breakpoint hotspot regions (EBHRs). One of the basics of comparative genomics is the identification of synteny. However, there is no precise definition of synteny. Species selection, their evolutionary distances, pairwise or multi-way comparison, size of the anchors could all have huge impacts on the resulting synteny blocks (Reviewed by Ghiurcuta, 2014 [19]). In this study 12 species were selected with the human as the reference species from different principal branches of species tree. We selected species according to their evolutionary distance from the human as well as the

quality of their genome assembly. Species from more distant lineages of mammalian evolutionary tree and a non-mammalian vertebrate as out-group allows a higher resolution. However, it yields to shorter size of synteny blocks compared to previous studies. From multi-way alignment blocks, blocks shared by at least 7 selected species were accepted as conservation anchors. This looseness lowers the dependency of this method to the precise species list and could lower the bias of species selection and missing data. Also, the fact that about 80% of our EBrs have a Least Common Ancestor originated from the primate node to farther up on the reference tree, shows that these breakpoints are reuse and a careful replace of some selected species should not have a dramatic effect on the identified synteny blocks and their positions. Nonetheless, comparing identified synteny blocks with Conserved Ancestral Regions (CARs) that reconstructed by Ma, 2006 [27], showed that 98.4% of identified synteny blocks overlaps with 59.98% CARs. Considering the estimation of Ma, 2006 [27] closest to the real ancestral genomic regions, the presence of synteny blocks in conserved ancestral regions is not far from the presence of these synteny regions in the human contemporary genome calculated in this study (52%). Due to the use of human oriented multi-way conservation alignments as anchors, over 7% of EBrs have a size equal to 1 bp as conservation anchors are contiguous on human genome. Comparison of EBrs with evolutionary breakpoints produced in the study of Lemaitre, 2009 [26] showed that EBrs were similar to what Lemaitre, 2009 [26] identified as breakpoint regions (BPR). It should be noted that, in this study pairwise comparison of orthologous genes based on five other mammals (with 4 that were common with species in this study) were used to identify synteny and their BPRs have similar size distribution of EBrs produced in our study. Other than the choice of the species, the use of genes as conservation anchors limited their study to only the coding regions. However, still 3% of EBrs produced in this study overlaps with 69.8% of their breakpoint regions.

0.7 Location of genes based on different selective pressure

The results presented in this study highlights the dynamic nature of the genome. EBrs are dense in regions coding for functions related to adaptive response such as *cellular process*, *metabolic process*, *DNA repair*, *response to DNA damage stimulus*, *cellular response to stress* and *catabolic process*. On the other hand, genes with high selection pressure such as members of homeobox gene family of transcription factors, sex determining region Y (SRY), BAF chromatin remodeling complex subunit (BCL11A), Forkhead box transcription factors (FOXP1) and zinc finger gene family are located in EBRRs. All the genes in these regions are conserved among mammals and vertebrates. SRY initiates male sex determination. Mutation in this gene is known to be associated with 46,XY complete gonadal dysgenesis or 46,XY pure gonadal dysgenesis or Swyer syndrome [33]. Translocation of part of the Y chromosome containing this gene to the X chromosome causes XX male syndrome [18]. The homeobox genes encode a highly conserved family of transcription factors that play an important role in morphogenesis in all multicellular organisms [18]. HOXB genes are known to encode conserved transcription factors in most bilateral species [34]. FOXB1 belongs to subfamily P of the forkhead box (FOX) transcription factor family and known to be associated with Mental retardation with language impairment and autistic features (MRLIAF). Zinc finger genes code for transcription factors in all eukaryotes. 33 members of zinc finger genes located on the 19p12 chromosome band. This 4.4 Mbp region is the largest EBr desert in our dataset having only one single EBr within this region. This gene cluster is known to be highly conserved in primate [4, 14]. This region is presented in figure 4. These results are in harmony with previous study indicating that refractory regions are strongly enriched for genes involved in development [15, 30] and any disruption in these regions should have detrimental effect [43]. Whereas, regions with high tendency of

rearrangements are more involved in adaptive process such as metabolic process, DNA repair and response to stress [30].

In conclusion the identified EBHRs provides a better insight on the nature of genome by emphasizing on the difference of the functions that EBr hotspots and EBr refractory regions harbor. It points out the dynamic process of evolution, adaptation and natural selection in interaction with the selective pressures of several genomic regions that maintain the integrity functions crucial for the survival of the species. The EBHRs and EBBRs reported in this paper will constitute a good asset for further studies including the affinity of some genomic regions to rearrangement associated to diseases known to be driven by genome rearrangements such as cancers.

Materials and Methods

0.8 Basic definitions

- ***Distance*** δ***:*** A distance between two consecutive blocks, B_i, Bj, is the minimum distance of one's head with the tail of the other for each species: $\delta = (B_j.s_t - B_i.s_h)$.
- ***Multi-way conservation alignment:*** is a series of multiple alignment of conserved segments in a set of genomes based on a reference genome. The multiple alignments used in this study were generated using multiz and other tools in the UCSC/Penn State Bioinformatics comparative genomics alignment pipeline. Conserved elements are identified by phastCons [22].
- ***Synetnic block:*** A syntenic block is defined as a large region of the genome that corresponds to a collection of contiguous blocks that maintained their positions and orders among a group of species since a common ancestor.
- ***Synteny break (evolutionary breakpoint or EBr)***: Given two consecutive conserved blocks, any discontinuity (as in distance, chromosome or orientation) between two sequences of one or more compared species that originated from any ancestry node of the reference genome (by Lowest Common Ancestor (LCA) approach) is considered a synteny breaks.
- ***Breakpoint region:*** Given two consecutive synteny blocks, the distance between the two blocks on the reference genome.
- ***Breakpoint-hotspot:*** Breakpoint-hotspot regions is a genomic region that represents a number of breakpoints significantly higher with respect to other genomic regions.

0.9 Synteny identification

To perform a comparative analysis on the human genome, the most appropriate candidates to capture more evolutionary patterns in the analysis are of two kinds : 1) neighboring species that share common features, and 2) more or less distant species that could have a broad divergence with the reference species (human). For the first group, three well-studied primates have been chosen, chimpanzee, orangutan and marmoset. For the second group, two well-sequenced and well-studied species from three major branches of mammalian phylogenetic tree were selected as follows: rat and mouse from Supraprimates, dog and cow from Laurasiatheria as well as elephant and armadillo as non-Boreotherian mammals. We also added two outliers, one non placental mammal, opossum and one non mammalian vertebrate, chicken. 45-way conservation sequences

alignments were obtained from UCSC genome browser in MAF format from: `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/maf`. Conservation alignment blocks having at least seven species among the 12 selected species were extracted as conservation anchors. Unselected species were removed from the blocks and missing selected species were added having unknown chromosome and gaps as their sequences. Synteny is constructed as follows:

Given:

- a set of selected species: $S = \{s_R, s_a, s_b, ..., s_n\}$, with S_R as reference species.
- a species reference tree T with set of ancestry nodes of s_R $A = \{A_1, A_2, A_3, ..., A_n\}$.
- a multi-way conservation sequence alignment blocks $B = \{B_1, B_2, B_3, ..., B_n\}$ each species sequence in a block B_i identified by its chromosome, $B_i.s_c$, its orientation, $B_i.s_o$ and its head $Bi.s_h$ and tail $B_i.s_t$

Discontinuity between two blocks, B_i and B_j with respect to each species is defined by any of three conditions:

- difference in chromosomes: $B_i.s_c \neq B_j.s_c$
- difference in orientations: $B_i.s_o \neq B_j.s_o$
- distance, $\delta(B_j.s_t - B_i.s_h)$, greater than a defined threshold G: $\delta > G$

for each two blocks, B_i and B_j:

1. S* $\longleftarrow$ get subset of S with discontinuity
2. if LCA(S*) $\subset A$ then:

   next
3. else

   $B_i \longleftarrow$ fus(B_i, B_j)
   $B_j \longleftarrow B_{j+1}$
   go to step 1.

For each two continuous anchors (with respect to the reference genome), species with any discontinuity (chromosome, orientation and distance) were identified. If the lowest common ancestor of species with discontinuity is an element of human ancestry nodes of the reference species tree, the two blocks would be considered as discontinuous and the algorithm will continue to compare the next couple in the line. Otherwise, the two blocks will be fused together to be compared with the next block. In each iteration, the list of species with discontinuity and their LCA will be documented. The results of this step is a list of larger conservation clusters that will go through a size filter (minimum size of 1000 bp). The clusters passed through the filter are the resulting synteny blocks. These steps are well illustrated in figure 5.

0.10 Identification of lineage-specific evolutionary breakpoint

Genomic regions bounded by two consecutive synteny blocks with a size smaller than 2 Mbp are considered as reference species lineage-specific breakpoints, or synteny breaks or evolutionary breakpoints (EBrs). The size constraint is necessary to avoid ambiguous regions that could not be associated correctly to breakpoints (e.g. sequences of heterochromatin). The origin of each EBr is the lowest common ancestor of species with discontinuity comparing the two bounded synteny blocks in the synteny identification step.

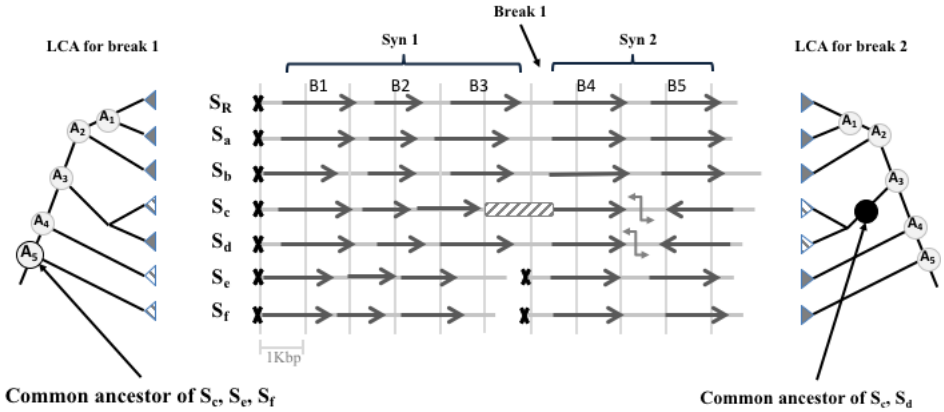

Figure 5. Extraction of synteny blocks and their corresponding breaks: In this illustration of a multi-way alignment blocks with respect to a reference species (S_R). S_a to S_f are the selected species present in the alignment blocks. The reference tree show the phylogenetic relationship of these species. A_1 to A_5 are the S_R ancestry nodes. Each arrow represents a conserved region on each genome. Directions in each arrow represent the orientation of that region in each species. The distance between each two vertical lines represents 1 Kbp. Each X shows the start of a chromosome.

0.11 Prediction of breakpoint-hotspots

To identify the genomic regions that are significantly enriched in breakpoint or breakpoint hotspot regions, we followed the permutation strategies suggested by **(author?)** [11]. We used a non-overlapping sliding window approach of size 100 Kb. In each window, the number of breakpoints was counted. Breakpoints were considered to fall into a window if they have at least one position overlapping that window. For 100,000 iterations, each breakpoint in the data set were simulated based on the following constraints:

- Breakpoint should be simulated within the original chromosome.
- Simulated breakpoint should not fall into centromeric or telomeric regions.
- Breakpoint should be simulated with respect to its size.
- Overlaps would be allowed to lower the calculation costs.

A p-value is calculated for the number of simulated breakpoints per window with compare to the original count. Windows with a p-vaule ¡ 0.05 is considered as significantly enriched windows or breakpoint hotspot regions.

0.12 Collection of genomic features

CpG islands, nested repeats, segmental duplications, CG content and self chain markers were downloaded from the UCSC Genome Browser [29]. Annotation on non-B DNA conformation were downloaded from Non-B DB [8]. Conserved elements of amniotes were obtained from catalog of conserved elements from genomic alignments (CEGA) [12]. Common fragile sites were provided by the supplementary material of Fungtammasan, 2012 [16] paper. RNA G-quadrupex were downloaded from RNA G-quadruplex database (G4RNA) [17]. Benign structural variation were donwloaded from Database of Genomic Variants [28]. RefSeq genes and exons where obtained from NCBI FTP portal. G quadruplex were downloaded from G4 database [45] in 2013.

However, the database is not accessible anymore. The complete list of these features are represented in supplementary table S5.

0.13 Identification of EBRRs

Regions with no identified EBrs were identified and their positions were compared to genomic regions that are not well sequenced or annotated. We found 1,462 genomic windows outside telomeric and centromeric regions that fall into these regions such as short arms of chromosome 13, 14, 15, 21, 22, as well as chromosome bounds 1q12, 9q12 and 16q11. These regions were eliminated from our analysis. Hence we left with 230 windows of 100 Kbp long with no EBr.

Supporting Information

Supporting data for this study is available at:
`https://github.com/bioinfoUQAM/RECOMB-CG-2019_supp`

Acknowledgments

We would like to thank Julie Horvath, Aida Ouangraoua, Abou Abdallah Malick Diouara and Bruno Daigle for helpful discussions. Special thanks to Emmanuel Mongin, by whom this project has been initially inspired. This work is supported by Natural Science and Engineering council of Canada (NSERC) and the Fonds de Recherche du Québec-Nature et Technologie (FRQNT) funds to ABD. AMR is a NSERC fellow. AMR and GV are FRQNT fellows. ...

References

1. L. Armengol, M. A. Pujana, J. Cheung, S. W. Scherer, and X. Estivill. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human molecular genetics*, 12(17):2201–2208, 2003.

2. A. Bacolla, J. A. Tainer, K. M. Vasquez, and D. N. Cooper. Translocation and deletion breakpoints in cancer genomes are associated with potential non-b dna-forming sequences. *Nucleic acids research*, 44(12):5673–5688, 2016.

3. J. A. Bailey, G. Liu, and E. E. Eichler. An *Alu* transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics*, 73(4):823–834, 2003.

4. E. J. Bellefroid, J.-C. Marine, A. G. Matera, C. Bourguignon, T. Desai, K. C. Healy, P. Bray-Ward, J. A. Martial, J. N. Ihle, and D. C. Ward. Emergence of the znf91 krüppel-associated box-containing zinc finger gene family in the last common ancestor of anthropoidea. *Proceedings of the National Academy of Sciences*, 92(23):10757–10761, 1995.

5. D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan, and R. Apweiler. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

6. M. Blanchette. Evolutionary puzzles: An introduction to genome rearrangement. In *Computational Science-ICCS 2001*, pages 1003–1011. Springer, 2001.

7. G. Bourque, P. A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome research*, 14(4):507–516, 2004.

8. R. Z. Cer, K. H. Bruce, U. S. Mudunuri, M. Yi, N. Volfovsky, B. T. Luke, A. Bacolla, J. R. Collins, and R. M. Stephens. Non-b db: a database of predicted non-b dna-forming motifs in mammalian genomes. *Nucleic acids research*, 39(suppl_1):D383–D391, 2010.

9. A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

10. S. De and F. Michor. Dna secondary structures and epigenetic determinants of cancer genome evolution. *Nature structural & molecular biology*, 18(8):950–955, 2011.

11. S. De, B. S. Pedersen, and K. Kechris. The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. *Briefings in bioinformatics*, 15(6):919–928, 2013.

12. A. Dousse, T. Junier, and E. M. Zdobnov. Cega—a catalog of conserved elements from genomic alignments. *Nucleic acids research*, 44(D1):D96–D100, 2015.

13. Y. Drier, M. S. Lawrence, S. L. Carter, C. Stewart, S. B. Gabriel, E. S. Lander, M. Meyerson, R. Beroukhim, and G. Getz. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of dna breakage and rearrangement-induced hypermutability. *Genome research*, 23(2):228–235, 2013.

14. E. E. Eichler, S. M. Hoffman, A. A. Adamson, L. A. Gordon, P. McCready, J. E. Lamerdin, and H. W. Mohrenweiser. Complex β-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome research*, 8(8):791–808, 1998.

15. P. G. Engström, S. J. H. Sui, Ø. Drivenes, T. S. Becker, and B. Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome research*, 17(12):1898–1908, 2007.

16. A. Fungtammasan, E. Walsh, F. Chiaromonte, K. A. Eckert, and K. D. Makova. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome research*, 22(6):993–1005, 2012.

17. J.-M. Garant, M. J. Luce, M. S. Scott, and J.-P. Perreault. G4rna: an rna g-quadruplex database. *Database*, 2015, 2015.

18. L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant. The ncbi biosystems database. *Nucleic acids research*, 38(suppl_1):D492–D496, 2009.

19. C. G. Ghiurcuta and B. M. Moret. Evaluating synteny for improved comparative studies. *Bioinformatics*, 30(12):i9–i18, 2014.

20. W. Gu, F. Zhang, and J. R. Lupski. Mechanisms for human genomic rearrangements. *Pathogenetics*, 1(1):4, 2008.

21. H. Kehrer-Sawatzki and D. N. Cooper. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Research*, 16(1):41–56, 2008.

22. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

23. H. Kikuta, M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engström, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*, 17(5):545–555, 2007.

24. J. Lee, K. Han, T. J. Meyer, H.-S. Kim, and M. A. Batzer. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS One*, 3(12):e4047, 2008.

25. C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier, and B. Audit. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics*, 10(1):335, 2009.

26. C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier, and B. Audit. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics*, 10(1):335, 2009.

27. J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome research*, 16(12):1557–1565, 2006.

28. J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992, 2013.

29. L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, et al. The ucsc genome browser database: extensions and updates 2013. *Nucleic acids research*, 41(D1):D64–D69, 2012.

30. E. Mongin, K. Dewar, and M. Blanchette. Long-range regulation is a major driving force in maintaining genome integrity. *BMC evolutionary biology*, 9(1):203, 2009.

31. E. Mongin, K. Dewar, and M. Blanchette. Mapping association between long-range cis-regulatory regions and their target genes using synteny. *Journal of Computational Biology*, 18(9):1115–1130, 2011.

32. W. J. Murphy, D. M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler, L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617, 2005.

33. H. page: National Library of Medicine (US). Genetics Home Reference [Internet]. Bethesda (MD). The library. Internet, June 2019.

34. J. C. Pearson, D. Lemons, and W. McGinnis. Modulating hox gene functions during animal body patterning. *Nature Reviews Genetics*, 6(12):893, 2005.

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

36. P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, 13(1):37–45, 2003.

37. P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, 13(1):37–45, 2003.

38. S. K. Pham and P. A. Pevzner. Drimm-synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, 26(20):2509–2516, 2010.

39. D. B. Pontier, E. Kruisselbrink, V. Guryev, and M. Tijsterman. Isolation of deletion alleles by g4 dna-induced mutagenesis. *nature methods*, 6(9):655, 2009.

40. S. Proost, J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, and K. Vandepoele. i-adhore 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research*, 40(2):e11–e11, 2011.

41. A. Sandelin, P. Bailey, S. Bruce, P. G. Engström, J. M. Klos, W. W. Wasserman, J. Ericson, and B. Lenhard. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC genomics*, 5(1):99, 2004.

42. D. Sankoff. The where and wherefore of evolutionary breakpoints. *J Biol*, 8:66, 2009.

43. F. Spitz, C. Herkenne, M. A. Morris, and D. Duboule. Inversion-induced disruption of the hoxd cluster leads to the partition of regulatory landscapes. *Nature genetics*, 37(8):889, 2005.

44. G. Wang, L. A. Christensen, and K. M. Vasquez. Z-dna-forming sequences generate large-scale deletions in mammalian cells. *Proceedings of the National Academy of Sciences*, 103(8):2677–2682, 2006.

45. H. M. Wong, O. Stegle, S. Rodgers, and J. L. Huppert. A toolbox for predicting g-quadruplex formation and stability. *Journal of nucleic acids*, 2010, 2010.

46. A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology*, 3(1):e7, 2004.

47. J. Zhao, A. Bacolla, G. Wang, and K. M. Vasquez. Non-b dna structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences*, 67(1):43–62, 2010.