Genomics & Informatics Vol. 8(2) 86-89, June 2010

Non-Synteny Regions in the Human Genome

Kichan Lee and Sangsoo Kim*

Department of Bioinformatics & Life Science, Soongsil University, Seoul 156-743, Korea

Abstract

Closely related species share large genomic segments called syntenic regions, where the genomic elements such as genes are arranged co-linearly among the species. While synteny is an important criteria in establishing orthologous regions between species, non-syntenic regions may display species-specific features. As the first step in cataloging human- or primate- specific genomic elements, we surveyed human genomic regions that are not syntenic with any other non-primate mammalian genomes sequenced so far. Based on the data compiled in Ensembl databases, we were able to identify 10 such regions located in eight different human chromosomes. Interestingly, most of these highly human- or primate- specific loci are concentrated in subtelomeric or pericentromeric regions. It has been reported that subtelomeric regions in human chromosomes are highly plastic and filled with recently shuffled genomic elements. Pericentromeric regions also show a great deal of segmental duplications. Such genomic rearrangements may have caused these large human- or primate- specific genome segments.

Keywords: synteny, pericentric, subtelomeric, genome rearrangement, mammalian genomes

Introduction

Closely related species that have diverged relatively recently, share large genome segments that show co-linearly arranged genomic features such as genes. These regions are said in synteny. For example, mammals share a great deal of syntenic regions, implying divergence from a common ancestral genome (Bourque, 2004). Often a genomic segment in a chromosome of a species is split into two different chromosomes in another species (Peng, 2006). It was also observed that the synteny break points were conserved among the species, allowing reconstruction of the ancestral chromosomes (Murphy, 2005). Since synteny blocks have inherited from a common ancestor through genome rearrangement and the genes in a syntenic block are in the same order, synteny plays a critical role in identifying orthologs between closely related species. On the other hand, non-syntenic regions would show species-specific features that are due to genome rearrangement. Since several mammalian genome sequences are available now (Rhead, 2010), we may identify regions of a species that are not syntenic with any other species. These regions would be highly specific to the species of interest. Ensembl Genome Browser hosts a number of sequenced whole genomes and provides information on syntenic regions (Flicek, 2010). Since the comparison of human and chimpanzee genome sequences showed no noticeable non-syntenic regions caused by genome rearrangements (CSAC, 2005), we focused on the comparison of the human genome versus all the other non-primate mammalian genomes. Using the data downloaded from Ensembl database, we cataloged the genomic regions that were not syntenic with any other non-primate mammalian genomes. Here we report how many such regions exist in human genome, and how they are distributed among the chromosomes and regionally within a given chromosome.

Methods

Cataloging syntenic regions

Our definition of synteny between a pair of genomes was based on the working principle used by Ensembl Genome Browser (http://www.ensembl.org/info/docs/compara/analyses.html). Ensembl filters the result of blastZ (Schwartz, 2003) runs between a pair of genomes by chaining the aligned blocks into bigger nets using blastZ-net, followed by grouping neighboring regions closer than 200kb. The final syntenic regions are defined by merging those grouped regions if they are not interrupted by non-syntenic blocks. The resulting synteny data files (dnafrag.txt, dnafrag_region.txt, genome_db.txt, method_line.txt, method_link_species_set.txt, and synteny_region.txt) are downloadable from Ensembl ftp site (ftp://ftp.ensembl.org/pub/). We made a synteny database from these six files using MySQL database management system. This database includes 10 mammalian species *(Rattus norvegicus, Macaca mulatta, Pan troglodytes, Canis familiaris, Monodelphis domestica, Mus musculus, Pongo pygmaeus, Equus caballus, Bos Taurus, and*

*Corresponding author: E-mail sskimb@ssu.ac.kr
Tel +82-2-820-0457, Fax +82-2-824-4383
Accepted 14 June 2010

Homo sapiens). Since we concerned human-specific features not shared with other non-primate mammals, we excluded non-human primates and used the remaining six genomes. The pair-wise synteny information between human and each of the other species was saved in BED format (http://genome.ucsc.edu/FAQ/FAQformat.html#format1).

Inferring non-syntenic regions

In order to identify human genome regions that are not synteny with any other non-primate mammalian species, the synteny regions saved in the previous step were masked from the human genome contig intervals. This procedure was operated with the Galaxy tool (http://main.g2.bx.psu.edu/) (Taylor, 2007) using the genome data downloaded from UCSC Genome Browser (http://genome.ucsc.edu/) (Rhead, 2010).

Results

A total of 2,106 synteny blocks between human genome and each of six non-primate mammalian (mouse, rat, dog, cow, pig, and opossum) genomes were retrieved from the database that mirrored some relevant tables of Ensembl database (ftp://ftp.ensembl.org/pub/) (Flicek, 2010). The genomic intervals of the syntenic regions were uploaded into Galaxy web server (http://main.g2.bx.psu.edu/) (Taylor, 2007). The overlapping intervals were merged into 117 intervals, which were then masked from the human genome contig intervals (249 in total) that were downloaded from UCSC Genome

Table 1. Non-synteny regions in the human genome

Chromosome	Start	End	Length (bp)	Cytoband	Locus
Chr1	121,086,695	121,485,434	398,739	P11.2	Pericentromeric
Chr2	91,595,103	92,326,171	731,068	p11.1	Pericentromeric
Chr9	46,561,039	47,060,133	499,094	p11.2	Pericentromeric
Chr12	132,806,992	133,841,895	1,034,903	q24.33	Subtelomeric
Chr13	114,425,993	114,639,948	213,955	q34	Subtelomeric
	114,739,948	115,109,878	369,930	q34	Subtelomeric
Chr17	79,759,049	81,195,210	1,436,161	q25.3	Subtelomeric
Chr21	10,365,976	10,647,896	281,920	p11.2	Pericentromeric
	10,697,896	11,188,129	490,233	p11.2-p11.1	Pericentromeric
Chr22	50,414,777	51,244,566	829,789	q13.33	Subtelomeric

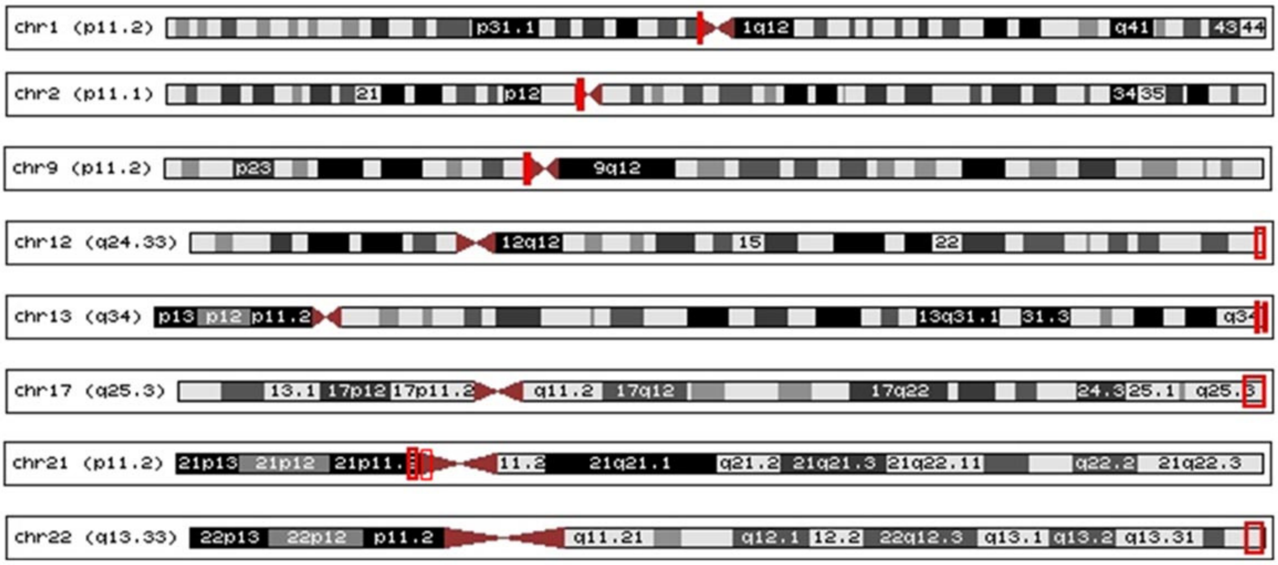

Fig. 1. Ideograms of the chromosomes where non-synteny regions are marked by red boxes. In chromosomes 13 and 17, two segments closely separated are shown, while for the other chromosomes only one segment per chromosome is shown.

Browser using Galaxy tools. This procedure resulted in a total of 28 genomic intervals that are not syntenic with any of the six genomes (data not shown). In collecting syntenic blocks, Ensembl merged genomic alignment blocks that were co-linear and less than 200kb apart. If a genome assembly gap exists flanking a syntenic block, the block would not be stitched with other syntenic blocks, we filtered out the blocks shorter than 200kb or located within 200kb from an assembly gap. We also excluded non-syntenic blocks in chromosome Y, a large portion of which are of segmental duplications. Finally we ended up with 10 such blocks as shown in Table 1. One locus was located on each of chromosomes 1, 2, 9, 12, 17, and 22, while two loci per each of chromosomes 13 and 21 were found. As shown in Fig. 1, the genomic loci were found in five subtelomeric and five pericentromeric regions. Subtelomeric regions were known to be highly plastic and filled with recently shuffled genomic elements (Ambrosini 2007; Riethman, 2005). Pericentromeric regions were also known to have undergone frequent segmental duplications either inter-chromosomal or intra-chromosomal (IHGSC 2001; 2004). Such genomic rearrangement in recent evolutionary history would have created human- or primate-specific genomic regions that are not syntenic with any of the non-primate mammalian genomes.

Discussion

We surveyed along human genome the regions non-syntenic with any of the known non-primate mammalian genomes. Gaps in genome sequence assembly may introduce artifacts in identifying such regions. In order to avoid such artifacts, we applied a strict filtering step where putative non-syntenic blocks near gaps were excluded. The source data used in this analysis were based on the synteny blocks defined by Ensembl Genome Browser (Flicek, 2010), which ignored short indels and tried to identify a small number of large chunks co-linear between a pair of genomes. Consequently we ended up with 10 extremely non-syntenic regions. If one included shorter synteny blocks in the calculation, more non-syntenic regions would have been identified. With the current definition of synteny blocks, we would identify only the most extremely non-syntenic regions.

Interestingly the non-syntenic regions reported here were confined to subtelomeric or pericentromeric regions, where recent segmental duplications prevailed (IHGSC, 2001; 2004). Although it is likely that invasions by transposable elements may also create species- or lineage-specific genomic elements, they tend to be dispersed around the genome and would not form a chunk large enough to be identified as synteny blocks. Instead the results of segmental duplications would create large chunks that are non-syntenic.

References

Ambrosini, A., Paul, S., Hu, S., and Riethman, H. (2007). Human subtelomeric duplicon structure and organization. *Genome Biol.* 8, R151.

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.

CSAC (Chimpanzee Sequencing and Analysis Consortium) (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69-87.

Flicek, P., Aken, B.L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y.A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S.P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Smith, J., and Searle, S.M. (2010). Ensembl's 10th year. *Nucl. Acids Res.*, Database issue, D557-D562.

IHGCS (International Human Genome Sequencing Consortium) (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945.

IHGSC (International Human Genome Sequencing Consortium) (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

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., Hitte, C., Meyers, S.N., Milan, D., Ostrander, E.A., Pape, G., Parker, H.G., Raudsepp, T., Rogatcheva, M.B., Schook, L.B., Skow, L.C., Welge, M., Womack, J.E., O'brien, S.J., Pevzner, P.A., and Lewin, H.A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Sci.* 309, 613-617.

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

Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., Pohl, A., Pheasant, M., Meyer, L.R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R.A., Giardine, B., Dreszer, T.R., Clawson, H., Barber, G.P., Haussler, D., and Kent, W.J. (2010). The UCSC Genome Browser database: update 2010. *Nucl. Acids Res.*, Database issue, D613-D619.

Riethman, H., Ambrosini, A., and Paul, S. (2005). Human subtelomere structure and variation. *Chromosome Res.* 13, 505-515.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103-107.

Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. (2007). Using galaxy to perform large-scale interactive dataanalyses. *Current protocols in bioinformatics,* Chapter 10, Unit 10.5.