Genome-Scale Detection of Tandemly-Duplicated Gene Structures

Keisuke Onishi[1] ss27183@mail.ecc.u-tokyo.ac.jp
Hao Zhang[2] hzhang@jbirc.aist.go.jp
Takuro Tamura[2] ttamura@jbirc.aist.go.jp
Takashi Gojobori[2,3] tgojobor@genes.nig.ac.jp
Shintaroh Ueda[1] sueda@biol.s.u-tokyo.ac.jp

[1] Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2] Japan Biological Information Research Center, 1-1 Time24 Bldg.10F, 2-45 Aomi, Koto-ku, Tokyo 135-0064, Japan
[3] Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

Keywords: dotplot, DSCAM, tandemly-duplicated structures

1 Introduction

Genomic sequences of several animals were almost completely sequenced, and it was shown that human genome has relatively small numbers of genes. This suggests that not mere gene number but proteomic and/or interactomic combination is more crucial for phenotypic complexity. Recently, a *Drosophila* gene, DSCAM, was reported to have surprising genomic structure that can generate more than 30,000 isoforms potentially by alternative splicing [1]. Though our genome has its putative ortholog on chromosome 21 and other homologs, they do not have such cassette exons as *Drosophila* [3]. These facts prompted us to investigate unknown genes that have cassette exons in the human genome. If human has no such genes, there may be some alternative genomic structures with regard to locally-duplicated structures. Therefore, we scanned human and *Drosophila* genomes using dotplot. Dotplot is a very useful tool to detect tandemly-duplicated structures even when there are no homology with known genes. When interspersed repetitive elements are effectively masked, we can identify such structures very accurately with a very low background regardless of protein-coding or non-coding. By this method, we extracted many known genomic structures such as gene clusters and domain clusters, and other unknown tandemly-duplicated structures from human and *Drosophila* genomes.

2 Method and Results

Genomic sequences of human (June 28, 2002 Genbank freeze: `http://genome-test.cse.ucsc.edu/downloads.html#human`) and *D. melanogaster* (Release 3: `http://www.fruitDrosophila.org/sequence/download.html`) were used. "Masked" human sequence was divided into 1Mb or 100kb each, and "unmasked" *Drosophila* sequence was first divided in the same way, then masked by RepeatMasker with default option [4]. These genomic fragments were queried against themselves for BLASTZ search of PipMaker locally (human) or on the web (*Drosophila*) [2]. Then, the outputs were visualized as dotplots, and plots were selected by the criteria of tandemly-duplicated structures spanning at least 10kb. Then, they were checked by BLASTN or BLAT search (`http://genome-test.cse.ucsc.edu/cgi-bin/hgBlat`). Several regions of particular interests were analyzed further by genomic comparison with other species.

In this analysis, we identified such gene structures as shown in Table 1 and several unknown structures. As a whole, human genome has much more domain-duplicated genes or protein-coding gene clusters than *Drosophila*, compared with their gene number ratio. For example, the former has nearly one thousand of zinc finger or olfactory receptor genes, whereas, the latter has no such structures at least as detectable clusters. In addition, human genome has several non-coding gene clusters like C/D box small nucleolar genes. Such non-coding structures but for *r*RNAs or *t*RNAs were not detected in the *Drosophila* genome. Besides, there are some complex structures involved in immune systems or cell-cell interaction (protocadherin clusters) in the human genome, but no genes with cassette exons like *Drosophila* DSCAM were not found in the human genome.

Table 1: Genes with tandemly-duplicated structures.

	Human	*Drosophila*
Domain repeat	Titin	Titin
	Collagen	Dumpy
	DAZ	
Gene Cluster		
(protein-coding)	Histone	Cytochrome P450
	Hox (A, B, C and D)	
	Olfactory receptor (including taste receptor)	
	Zinc finger	
(non-coding)	C/D box small nucleolar	5S *r*RNA
	TTTY	
Exon	Protocadherin (alpha and gamma)	DSCAM
(or gene fragment) cluster	Immunoglobulin (kappa light, gamma, heavy and lambda)	
	T-cell receptor (alpha and beta)	
Other Unknown structures	many	few

Only representatives were shown for each category.

3 Discussion

Here, we detected much more tandemly-duplicated structures including unknown structures in human than in *Drosophila*. This suggests that human have many protein-coding or non-coding gene clusters and that they may be contribute to alternative proteomic or interactomic diversity.

References

[1] Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L., *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity, *Cell*, 101:671–684, 2000.

[2] Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W., PipMaker-a web server for aligning two genomic DNA sequences, *Genome Res.*, 10:577–586, 2000.

[3] Yamakawa, K., Huot, Y.K., Haendelt, M.A., Hubert, R., Chen, X.N., Lyons, G.E., and Korenberg, J.R., DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system, *Hum. Mol. Genet.*, 7:227–237, 1998.

[4] Smit, A.F.A. and Green, P., 1999. (`http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl`)