Genomics Data 3 (2015) 49–54

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Comparison of inter- and intraspecies variation in humans and fruit flies

Juliann Shih [a,b], Russ Hodge [c], Miguel A. Andrade-Navarro [c,d,e,*]

[a] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[b] Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA
[c] Max Delbrück Center for Molecular Medicine, Germany
[d] Faculty of Biology, Johannes-Gutenberg University of Mainz, Mainz, Germany
[e] Institute of Molecular Biology, Mainz, Germany

ARTICLE INFO

Article history:
Received 19 September 2014
Received in revised form 12 November 2014
Accepted 12 November 2014
Available online 23 November 2014

Keywords:
Evolution
Population
Variation
Human genome
Drosophila

ABSTRACT

Variation is essential to species survival and adaptation during evolution. This variation is conferred by the imperfection of biochemical processes, such as mutations and alterations in DNA sequences, and can also be seen within genomes through processes such as the generation of antibodies. Recent sequencing projects have produced multiple versions of the genomes of humans and fruit flies (*Drosophila melanogaster*). These give us a chance to study how individual gene sequences vary within and between species. Here we arranged human and fly genes in orthologous pairs and compared such within-species variability with their degree of conservation between flies and humans. We observed that a significant number of proteins associated with mRNA translation are highly conserved between species and yet are highly variable within each species. The fact that we observe this in two species whose lineages separated more than 700 million years ago suggests that this is the result of a very ancient process. We hypothesize that this effect might be attributed to a positive selection for variability of virus-interacting proteins that confers a general resistance to viral hijacking of the mRNA translation machinery within populations. Our analysis points to this and to other processes resulting in positive selection for gene variation.

Introduction

Traditionally, we have traced the evolution of genes by comparing homologous versions in different organisms. Such homologies reflect a basic conflict: between the conservation of sequence features related to gene functions and to the structures of translated protein products on the one hand, and on the other, processes that produce genetic variation and make sequences drift away over millions of years of evolution.

With the sequencing of multiple versions of genomes of single species, we have now the chance to observe a different aspect of the forces that shape molecular evolution by studying gene sequence variation within a species. There is a general expectation that the variation of a gene's sequence *within* a species and *between* species will agree, leading to a similar constraint of evolutionary drift at both levels. However, we wondered whether we could detect particular genes that displayed increased variability within a single species, for example to provide fast adaptation of a population to variable environments, or to escape pathogens that recognize a protein and co-evolve with the species. Such genes might be detected by rates of (evolutionary short-range) intraspecies variability that are higher than expected when compared to (evolutionary long-range) interspecies variability.

To explore whether such cases can be detected through an unbiased, genome-wide level analysis, we took advantage of the recent evaluation of genetic variation in human [1] and *D. melanogaster* [2] genomes. To contrast short-range intraspecies variation with interspecies variation, we obtained and concentrated our analysis on pairs of one-to-one orthologous genes between these two species. Then we defined the degree to which each pair of human and fly orthologs demonstrated identity to each other and their respective intraspecies variation. Comparisons of the results allowed us to find, overall, the expected correlation between long and short evolutionary range conservation: most genes are highly constrained in their evolution and therefore will not change much both within and across species. However, we could also detect outlier genes that are highly variable in human or fly populations while being highly conserved between these species. We carried out the analysis both at the levels of nucleotide sequences and at the (translated) amino acids to compare and distinguish evolutionary constraints that might possibly act differently on these levels.

Results

The evolutionary divergence of *Homo sapiens* and *D. melanogaster* (fruit-fly) from a common ancestor has been estimated at

* Corresponding author at: Miguel Andrade Computational Biology and Data Mining group Johannes-Gutenberg University of Mainz Faculty of Biology Institute of Molecular Biology Ackermannweg 4, 55128 Mainz, Germany.

http://dx.doi.org/10.1016/j.gdata.2014.11.010

approximately 782.7 million years ago [3]. Despite this, Ensembl Compara's phylogenetic-approach homolog prediction tool [4] indicates that a total of 14.9% of human genes and 46.0% of fly genes have orthologs (genes in different species that descended from the same ancestral sequence) to one or more fly and human genes, respectively.

A significant challenge in evolutionary biology is to determine the relationship between a gene's functions and its ability to avoid being phased out on an evolutionary scale, due to either negative selection or, more likely, disuse. Also implicit in such relationships is the level of conservation between human genes and their fly ortholog(s) and vice versa, which can be calculated in terms of the percentage of identity of nucleotides and/or amino acids. According to current thinking, conservation between a fly and human ortholog pair would imply that both genes have a function implicated in the survival of each species, while mutations potentially result in phenotypic disadvantages. Although we are aware of no studies that have proven this in fruit-flies, a high correlation has been established between the essential functions of mouse genes and their level of evolutionary conservation in humans [5]. Thus between different human individuals, as well as between different fly strains, high interspecies DNA/amino acid transcript conservation should indicate "essential" gene functions and should also confer high intraspecies conservation in the same gene.

The advent of next-generation sequencing technologies has made the full sequencing of the genomes of large numbers of individuals and strains significantly more efficient. Two large-scale projects taking advantage of these technologies have been Bart Deplancke's catalog of insertions, deletions, complex variants, and single nucleotide polymorphisms (SNPs) in 39 *D. melanogaster* Genetic Reference Panel (DGRP) inbred lines and their effects on gene expression [2], and the 1000 Genomes Project, which catalogs variants from 1092 human individuals from 14 different populations [1]. The 1000 Genomes Project also shows that evolutionary conservation is a key determinant of the strength of purifying selection, meaning that there is a correlation between the essential nature of the functions of a protein-coding gene and the conservation of base pairs (or corresponding amino acids) that it exhibits among different individuals of the same species.

Because evolutionary essentiality has been shown to cause both intra- and interspecies conservation (and therefore to restrict variation), here we strive to formally establish a correlation between percentage identity between human–fly orthologs (taken from reference genomes GRCh37.p12 and BDGP5) and intraspecies variation (taken from the DGRP and 1000 Genomes Project), while developing an analysis that would point to any genes that might escape this "rule".

To accomplish this, we determined the intraspecies variation and interspecies percentage identity of 3082 one-to-one orthologous pairs of genes between *H. sapiens* and *D. melanogaster* (Supplementary Table 1; see Methods for details). For each of the orthologous pairs, fly-to-human and human-to-fly percentage identities (for both nucleotides and amino acid sequences) were determined (Fig. 1; see Methods for details). For nucleotides, the distributions are rather symmetrical and show a peak at around 45% identity, while for amino acids the peak percentage identity is slightly lower, at around 30%, and the distributions show a skew to the left (towards lower values of identity).

We also computed the intraspecies variation score calculated for each gene, normalized for the length of the gene (Fig. 2; see Methods for details). The nucleotide distributions present a maximum, whereas the amino acid distributions peak near zero variation. This difference is due to the nucleotide variability allowed by synonymous substitutions in protein-coding genes. The median for the nucleotide distribution is higher for humans than for flies, whereas the medians of the amino acid distributions are rather similar. Comparing distributions of intraspecies variation is problematic because of fundamental differences in the geographic distribution of the human and fly populations that were chosen (see Discussion).

We found a correlation of intra-species variation between orthologous genes; that is, if the human and the fly genes had high intraspecies

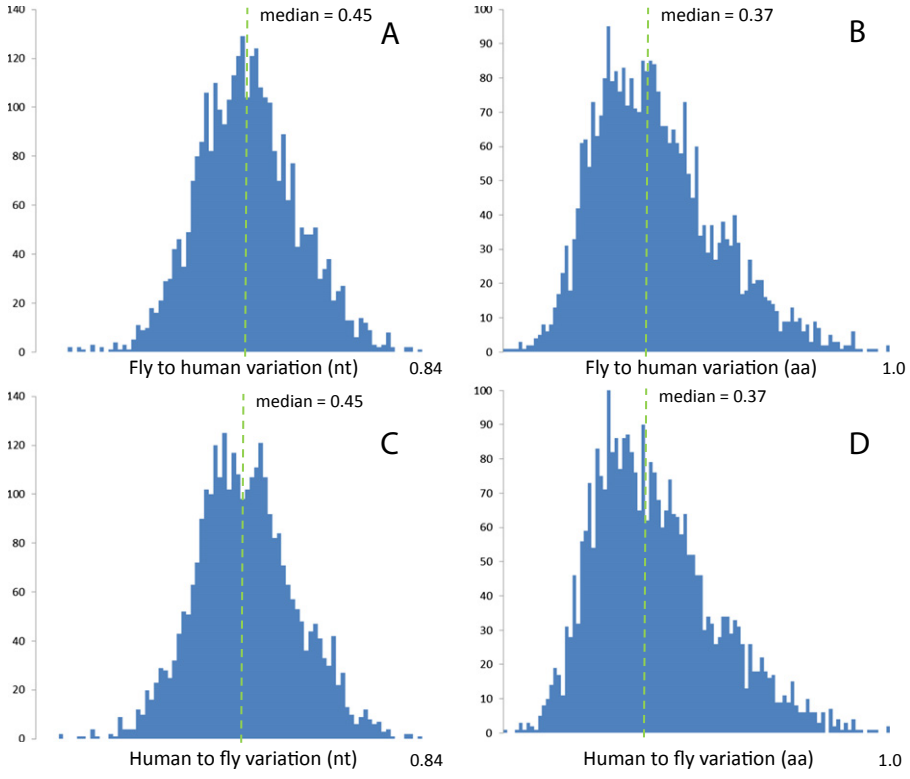

Fig. 1. Distributions of interspecies similarity (in percentage identity). Fly to human in nucleotides (A) and amino acids (B). Human to fly in nucleotides (C) and amino acids (D).