Theoretical Population Biology 93 (2014) 38–51

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

On the retention of gene duplicates prone to dominant deleterious mutations

Giulia Malaguti, Param Priya Singh, Hervé Isambert *

Institut Curie, CNRS-UMR168, UPMC, 26 rue d'Ulm, 75005 Paris, France

ARTICLE INFO

Article history:
Received 15 October 2013
Available online 12 February 2014

Keywords:
Population genetics models
Whole-genome duplication
Small scale duplication
Dominant deleterious mutations
Purifying selection

ABSTRACT

Recent studies have shown that gene families from different functional categories have been preferentially expanded either by small scale duplication (SSD) or by whole-genome duplication (WGD). In particular, gene families prone to dominant deleterious mutations and implicated in cancers and other genetic diseases in human have been greatly expanded through two rounds of WGD dating back from early vertebrates. Here, we strengthen this intriguing observation, showing that human oncogenes involved in different primary tumors have retained many WGD duplicates compared to other human genes. In order to rationalize this evolutionary outcome, we propose a consistent population genetics model to analyze the retention of SSD and WGD duplicates taking into account their propensity to acquire dominant deleterious mutations. We solve a deterministic haploid model including initial duplicated loci, their retention through sub-functionalization or their neutral loss-of-function or deleterious gain-of-function at one locus. Extensions to diploid genotypes are presented and population size effects are analyzed using stochastic simulations. The only difference between the SSD and WGD scenarios is the initial number of individuals with duplicated loci. While SSD duplicates need to spread through the entire population from a single individual to reach fixation, WGD duplicates are *de facto* fixed in the small initial post-WGD population arising through the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. WGD duplicates prone to dominant deleterious mutations are then shown to be indirectly selected through purifying selection in post-WGD species, whereas SSD duplicates typically require positive selection. These results highlight the long-term evolution mechanisms behind the surprising accumulation of WGD duplicates prone to dominant deleterious mutations and are shown to be consistent with cancer genome data on the prevalence of human oncogenes with WGD duplicates.

1. Introduction

Gene duplication has long been recognized as a major source of genetic innovation in the course of evolution through the retention and divergence of specific gene duplicates arising by chance (Ohno, 1970; Holland et al., 1994; Sidow, 1996). Gene duplicates are also thought to confer some mutational robustness against loss-of-function mutations (Winzeler et al., 1999; Gu et al., 2003; Kamath et al., 2003; Gu, 2003). Conversely, however, the duplication of genes prone to dominant deleterious mutations, such as gain-of-function mutations, is expected to lead to an enhanced susceptibility to genetic diseases and, hence, be opposed by purifying selection (Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009). Yet, surprisingly, such "dangerous" gene families prone to dominant deleterious mutations have often been greatly expanded by duplication in the course of evolution, see *e.g.* Ise et al. (2000) and Esteban et al. (2001).

In particular, gene families frequently implicated in cancer and other genetic diseases in vertebrates have been greatly expanded through two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates (Singh et al., 2012). By contrast, gene families lacking such a susceptibility to dominant deleterious mutations have been more typically expanded through small scale duplication (SSD) (Singh et al., 2012). More generally, gene duplicates originated from SSD or WGD events have been shown to exhibit antagonist retention patterns, with gene families expanded through WGD having typically few additional SSD genes and, vice versa, for gene families expanded mostly through SSD which exhibit few additional retained duplicates from WGD (Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; Singh et al., 2012). This implies that the mode of duplication through SSD or WGD events directly impacts the selection process

* Corresponding author.
E-mail address: herve.isambert@curie.fr (H. Isambert).

http://dx.doi.org/10.1016/j.tpb.2014.01.004
0040-5809/

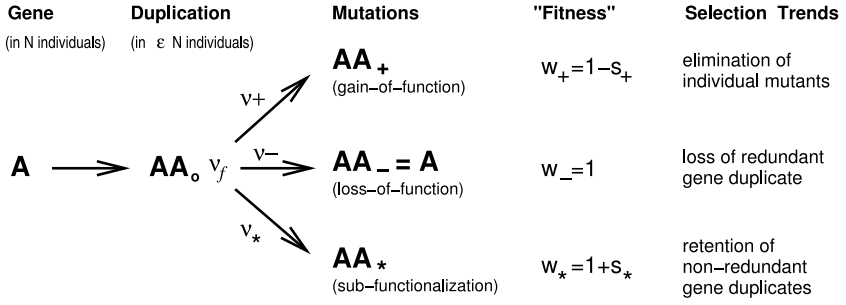

Fig. 1. *Haploid model for the retention of gene duplicates*. The model consists of two initial duplicated loci $AA_\circ$ in a haploid population with mutation rates (ν_i) towards deleterious gain-of-function mutants (AA_+ with $w_+ = 1 - s_+ < 1$), neutral loss-of-function mutants at a single locus (AA_- with $w_- = 1$) and neutral or beneficial fixed duplicates through sub-functionalization ($AA_\star$ with $w_\star = 1 + s_\star \geqslant 1$). The only difference between the two duplication scenarios is the initial fraction ϵ of individuals with duplicated loci, which is $\epsilon \simeq 1/N$ in the post-SSD population of size N, while it is $\epsilon \simeq 1$ in the post-WGD population arising through WGD-induced speciation. This WGD-induced speciation results from the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. See main text for model details.

of gene duplicates. Hence, their retention cannot be explained by the same *ad hoc* selection mechanism independent of the SSD or WGD modes of duplication.

These different retentions of SSD and WGD duplicates have been frequently associated to dosage balance constraints (Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Aury et al., 2006; Makino and McLysaght, 2010). However, extensive statistical analysis combining multiple properties of human genes (such as dosage balance constraints, association to cancers and genetic diseases and expression levels) have recently demonstrated (Singh et al., 2012) that the retention of WGD duplicates in vertebrates is more directly related to their susceptibility to dominant deleterious mutations than to dosage balance constraints or expression levels.

In this paper, we further strengthen this observation, showing that human oncogenes involved in different primary tumors have retained many WGD duplicates as compared to other human genes. This intriguing observation on the different retention patterns of WGD and SSD duplicates calls for a consistent population genetics model taking into account their propensity to acquire dominant deleterious mutations. To this end, we propose such a model focusing first on a simple, analytically tractable approach valid for large population sizes, before resorting to numerical simulations to analyze the consequences of stochastic fluctuations arising from finite population size. In particular, in order to analyze the retention of SSD *versus* WGD duplicates, we first use a simple deterministic model of two duplicated loci with neutral fixable genotypes in a haploid population of fixed size N and uncoupled mutation/selection dynamics. The only difference between SSD and WGD scenarios concerns the initial condition for each mode of duplication: the SSD case corresponds to a gene duplication in the genome of a *single* individual in the initial population, while the WGD case implies the genome duplication of *all* individuals in the small initial population arising through WGD. This is because WGD also induces a speciation event due to the ploidy incompatibility of the post-WGD individuals with the rest of the pre-WGD population. Although simplified, the asymptotic solutions of this deterministic population genetics model allow to capture the main evolutionary process responsible for the different retention of SSD *versus* WGD duplicates caused by dominant deleterious mutations. This haploid model is also extended into a simplified diploid model with three neutral haplotypes and one dominant deleterious haplotype. Then, to go beyond deterministic solutions for large populations, we use the formalism of one-step-process master equations and stochastic simulations to analyze the effect of finite population sizes on the retention of SSD *versus* WGD duplicates. All in all, this population genetics model supports the idea that the enhanced retention of "dangerous" WGD duplicates prone to dominant deleterious mutations is an indirect consequence of the initial speciation events triggered by WGD and the ensuing purifying selection in post-WGD species.

These results are then compared to the retention biases of SSD *versus* WGD duplicates for gene families with oncogenic properties and responsible for a broad range of primary tumors in human. Our application to genomic data will focus on the example of human oncogenes for which increasing amounts of data have recently become available from large scale cancer genome sequencing studies. Yet, unlike typical models on cancer genomics, *e.g.* Michor et al. (2004), Merlo et al. (2006), Beerenwinkel et al. (2007) and Bozic et al. (2010), our analysis of driver mutations from cancer genome data will *not* aim at modeling the *in situ* proliferation and selection of tumor cells within healthy tissues. Instead, it will concern the long-term evolution mechanisms that favored the surprising retention of WGD duplicates prone to dominant deleterious mutations in vertebrate genomes.

2. Model

We model the fixation of gene duplicates following either a SSD or a WGD event. In the following, we will first assume an haploid deterministic model to limit the number of two-locus combinations and stochastic effects to be considered. Extensions to diploid models and stochastic effects due to finite population size will then be analyzed in some details. Finally, the analytical and numerical solutions of these deterministic models will be compared to simulations of the corresponding stochastic population genetics models.

2.1. Haploid model for SSD and WGD duplicate retention

We start from the duplication event $A \rightarrow AA$ in a haploid genome, assuming that the newly duplicate gene is initially functionally redundant (Force et al., 1999; Lynch and Force, 2000a; Lynch and Conery, 2000; Lynch et al., 2001). Therefore, we assign to the initial (unstable) genotype with two redundant duplicated loci $AA_\circ$ a neutral fitness parameter $w_\circ = 1$. Then, we will consider three possible mutation-selection scenarios, corresponding to the emergence of three different phenotypes from the initial genotype $AA_\circ$, with mutation rates ν_-, $\nu_\star$, ν_+, and fitness parameters w_-, $w_\star$ and w_+, Fig. 1. Classical models suggest three alternative outcomes in the evolution of duplicate genes (Force et al., 1999; Lynch and Force, 2000a; Lynch and Conery, 2000; Lynch et al., 2001; Zhang, 2003). (i) One copy may become silenced by the accumulation of degenerative mutations and eventually become non-functionalized, while the other (fully functional) copy is retained. In our model, this corresponds to a neutral phenotype due to a loss-of-function of one of the duplicates (AA_-) with neutral fitness $w_- = 1$. (ii) Both copies may be reciprocally preserved