Journal of Theoretical Biology 341 (2014) 111–122

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Revisiting the diffusion approximation to estimate evolutionary rates of gene family diversification

Erida Gjini [a,*], Daniel T. Haydon [c,d,e], J. David Barry [e], Christina A. Cobbold [b,d]

[a] Instituto Gulbenkian de Ciência Oeiras, Portugal
[b] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom
[c] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[d] The Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, Glasgow, United Kingdom
[e] Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, United Kingdom

HIGHLIGHTS

- We model genetic diversification of a multi-gene family by explicitly simulating the evolutionary processes of point mutation and gene conversion.
- We link the stochastic dynamics of diversification to the Wright–Fisher model in population genetics and the diffusion approximation.
- We compare simulations and the diffusion approach across many relevant parameter regimes, showing a very good match for large family size, long gene sequences and small relative conversion tract length.
- We apply the diffusion approximation to estimate rates of diversification within the antigen gene family of African trypanosomes.

ARTICLE INFO

Article history:
Received 18 December 2012
Received in revised form
21 June 2013
Accepted 2 October 2013
Available online 11 October 2013

Keywords:
Identity distribution
Wright-Fisher model
Mutation rate
Gene conversion
Trypanosome antigen archive

ABSTRACT

Genetic diversity in multigene families is shaped by multiple processes, including gene conversion and point mutation. Because multi-gene families are involved in crucial traits of organisms, quantifying the rates of their genetic diversification is important. With increasing availability of genomic data, there is a growing need for quantitative approaches that integrate the molecular evolution of gene families with their higher-scale function. In this study, we integrate a stochastic simulation framework with population genetics theory, namely the diffusion approximation, to investigate the dynamics of genetic diversification in a gene family. Duplicated genes can diverge and encode new functions as a result of point mutation, and become more similar through gene conversion. To model the evolution of pairwise identity in a multigene family, we first consider all conversion and mutation events in a discrete manner, keeping track of their details and times of occurrence; second we consider only the infinitesimal effect of these processes on pairwise identity accounting for random sampling of genes and positions. The purely stochastic approach is closer to biological reality and is based on many explicit parameters, such as conversion tract length and family size, but is more challenging analytically. The population genetics approach is an approximation accounting implicitly for point mutation and gene conversion, only in terms of per-site average probabilities. Comparison of these two approaches across a range of parameter combinations reveals that they are not entirely equivalent, but that for certain relevant regimes they do match. As an application of this modelling framework, we consider the distribution of nucleotide identity among VSG genes of African trypanosomes, representing the most prominent example of a multi-gene family mediating parasite antigenic variation and within-host immune evasion.

1. Introduction

Quantifying the contributions of different evolutionary processes to the generation of genetic diversity is important to understand the evolution, adaptation and persistence of organisms. Key functions are often encoded by multi-gene families such as the major histocompatibility complexes (MHC) in man and mouse, the *Amy* multigene family of *Drosophila melanogaster*, and variable antigen genes of parasites such as *Plasmodium falciparum* and African trypanosomes. Typically multigene families contain genes that have arisen primarily via gene duplication, a driving force in molecular evolution (Ohno, 1970; Lynch and Conery, 2000; Bailey et al., 2002).

* Corresponding author at: Instituto Gulbenkian de Ciência, Oeiras, Portugal. Tel.: +351 214407968.
E-mail address: egjini@igc.gulbenkian.pt (E. Gjini).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.10.001

Gene duplication is then followed by gene conversion and point mutation, besides other processes such as unequal crossing over, recombination, random genetic drift and selection. Gene conversion is a special type of non-reciprocal transfer of genetic material in which one segment of DNA contributes genetic information to another, making the recipient location identical to the donor, but not altering the donor sequence. This process is very important for the concerted evolution of gene families and the functions they encode across organisms (Ohta, 2010). The combined effects of gene conversion and point mutation determine the diversification of duplicated genes, with gene conversion playing a major role in accelerating the spread of beneficial mutations through all gene family members.

Theoretical treatment of concerted evolution of multigene families presents many challenges, because the pattern of polymorphism in multigene families is much more complicated than that in single-copy genes. However some major advances using population genetic approaches were accomplished as early as the 1970s and 80s (Ohta, 1976, 1983; Nagylaki, 1984; Walsh, 1983), and modern approaches using the coalescent to understand multigene family complexity and evolution are increasing (see e.g. Griffiths and Watterson, 1990; Innan, 2002). Typical population genetic analyses focus on fixation probabilities of alleles and equilibrium identity coefficients under different scenarios (Mano and Innan, 2008; Innan, 2009). The picture of evolution gets complicated when the gene copy number is not constant in time (see Tachida and Kuboyama, 1998 for a gene duplication model), when gene conversion occurs in a biased or sequence dependent manner (Walsh, 1983, 1987), when mutation is biased and when selective forces are at play.

Although many characteristics of identity coefficients have been modelled, the temporal dynamics driving gene families towards such equilibria have usually received more minor attention, and simulation of idealized single-nucleotide events, rather than explicit whole genetic events has generally been adopted, with few exceptions (Innan, 2002). Because of analytical tractability, small multigene families with two copies of genes have been modelled more frequently, and distribution of allele frequencies between genomes rather than within genomes have been investigated. Depending on whether gene homology is studied by nucleotide identity or amino-acid identity, a K-allele model (Kimura and et al., 1968) or an infinite-allele model (Kimura and Crow, 1964) have been used respectively.

With the increasing availability of genetic data comes the challenge of quantifying the rates and characteristics of mutation, gene conversion, and other evolutionary forces that shape gene families from the molecular signatures they leave on DNA sequences. The rate and tract length of gene conversion between duplicated genes are among the most difficult parameters to infer. The empirical approach usually requires mutation accumulation experiments in transgenic model systems, while polymorphism (SNP) data is usually analyzed from a more theoretical standpoint, when DNA sequence data are available. More recently maximum likelihood methods have been proposed that overcome the limitation of estimates being model-dependent (Mansai et al., 2011). Empirical approaches for estimating tract length of gene conversions rely on identification of donor and recipient genes and involve the analysis of selected markers (see Song et al., 2011 for a recent review). In contrast, evolutionary data are not very informative for the tract length, mainly because of their dependence on the overall accumulation of footprints of historical gene conversions that potentially overlap with one another.

In this study, we investigate the dynamics of within-genome diversification of a multi-gene family as a result of only two recurring processes: point mutation and gene conversion among its members. We adopt two approaches, one based on simulation of discrete events and the other based on a diffusion approximation to extract information about the magnitude of genetic diversity attainable in a family of genes, and how it depends on the rates and characteristics of these evolutionary forces and on the family size. We analyze the role of various parameters, such as gene length, family size and conversion tract length in the distribution of pairwise identity in a gene family. We link the classical Wright–Fisher model (Fisher, 1930; Wright, 1931) to the dynamics of multigene family diversification, providing an avenue for further quantitative exploration of genomic evolution.

Finally, we apply our modelling framework to the nucleotide diversity of antigen (*VSG*) genes in African trypanosomes, to examine the interplay of gene conversion and point mutation within a group of related genes, representative of a multi-gene family that has originated through duplication. The overall distribution of genetic identity in this antigen gene family has two main implications for the fitness of the parasite: first, it interferes with higher-scale processes such as mosaic gene formation, often driven by identity-related recombination (Barbet and Kamper, 1993; Marcello and Barry, 2007a) involving pseudogenes; second, it can determine antigenic cross-reactivity between parasite variants that appear sequentially in infections and are targeted by the host immune system. Applying the diffusion approximation to the empirical genetic identity distribution of this multi-gene family, we lay a new bridge between mathematical theory and parasite genetic data, and are able to extract the rates of the evolutionary processes that can shape antigen gene diversification.

2. Modelling framework

To model the evolution of pairwise genetic identity in a multigene family we first consider all conversion and mutation events as they happen, keeping track of the donors and times of their occurrence; then we consider only the infinitesimal effect of these processes on pairwise identity accounting for random sampling of genes and positions. The first approach is purely stochastic, based on many explicit parameters, such as conversion event rate and mutation rate per unit of time, as well as conversion tract length, gene length and gene number, and it serves to visualize exact trajectories of the system of genes. The second approach is an approximation of the biological stochastic process, implicitly taking into account the characteristics of point mutation and gene conversion, but depending basically on just three parameters: mutation and conversion probabilities per base pair per generation and gene length, which makes it more amenable to analytic treatment.

2.1. Stochastic simulation of genetic events

Consider a population of N genes, each of length L, subject to gene conversion between pairs of genes and random point mutation. The state of a gene is represented by an array of L integers, corresponding to the 4 nucleotide types (A–C–T–G), in line with the K-allele ($K=4$) model. At time 0 a random initial sequence of length L is generated and applied to all genes, making them identical. We model the stochastic occurrence of single genetic events in such a multigene family as a Poisson process, which we simulate using the Gillespie Algorithm (Gillespie, 1977). The rate at which a gene is converted per unit of time is denoted by γ, while the rate at which mutations occur is given by μ. The global event rates of the two processes per unit of time are γN and μN. Stochastic events (mutations and conversions) are indexed $1,2,\ldots,T\in\mathbb{Z}$, which occur at the times $t_1,t_2,\ldots,t_T\in\mathbb{R}$. The inter-event times are exponentially distributed with mean $1/(\gamma N+\mu N)$. Only one event can happen at a time. Point mutation is chosen

with relative probability $\mu/(\gamma+\mu)$, whereas conversion events are chosen with relative probability $\gamma/(\gamma+\mu)$. The gene where point mutation occurs is randomly chosen out of N genes, and its position is also uniformly drawn out of L possible sites. Any nucleotide type can mutate to a specific one of the $K-1$, (3 in our case) remaining types with equal probability $1/(K-1)$.

The length of a conversion tract is generally assumed fixed, and is denoted by l_c $(1 \le l_c \le L)$, although this assumption can be easily relaxed. Each conversion involves a donor and recipient gene, chosen with probabilities $1/N$ and $1/(N-1)$ respectively. Conversion tracts are initiated at a random uniform site along the gene and continued from left to right until l_c sites are copied by the recipient. To distinguish between family members and track the ancestry of different gene segments, genes can be arbitrarily indexed as $1, 2, ..N$, so after conversion events, both the state sequence from the donor and the origin of the imported segment are retained in the recipient. An illustration of the system evolution is given in Fig. 1.

Pairwise identity between the genes is computed after each event by comparing their state sequences. Denoting by $h_{ij}(t)$ $(0 \le h_{ij} \le 1)$, the relative frequency of identical sites between two genes i and j at time t, the mean pairwise identity in the family is given by

$$\bar{h}(t) = \frac{\sum_{i=1}^{N}\sum_{j\neq i} h_{ij}(t)}{N(N-1)}. \quad (1)$$

The quantitative level of resolution of pairwise identity in this model directly depends on the length of each gene, as the smallest shift in pairwise identity equals $1/L$. We note that both mutation and conversion event rates are assumed to be independent of time and of current identity between interacting sequences and that family size, N, remains constant. The dynamics of mean pairwise identity, summarizing one evolutionary realization of a gene family, can be obtained analytically as a function of the number of events that have occurred, or generally as a function of continuous time (see Appendix A). Notice that the per-site rate of gene conversion is γl_c, while the per-site rate of point mutation is μ. Using $c = 2\gamma l_c/(N-1)$ and $m = 2\mu$ as the corresponding rates per pair of aligned sites on two genes, we have

$$\bar{h}(t) = \frac{c+m/3+me^{-(c+4m/3/L)t}}{c+4m/3}, \quad (2)$$

tending, as time tends to infinity, to the equilibrium value:

$$\bar{h}^* = \frac{c+m/3}{c+4m/3}. \quad (3)$$

The latter equilibrium identity formula matches the identity coefficient $\hat{c}_1$ derived by Ohta (1982) in her K-allele model without interchromosomal recombination. Notice that although the equilibrium identity coefficient does not depend on the number of sites in a gene, L, the temporal dynamics of approaching equilibrium are gene length dependent: keeping c and m constant, when there are more sites available for evolution, the dynamics is slower. Fig. 2 illustrates the dynamics of identity evolution in a gene family as a function of time. Not only the mean identity but also the entire distribution of identity among gene pairs changes through point mutation and gene conversion, tending to a stationary distribution as time increases. Notice that the fluctuations in pairwise identity are higher when the conversion tracts are longer, even though the number of events per unit of time is the same. Instead of a constant conversion tract length, a random tract length can be assumed at each event, for example a geometrically distributed tract length, which is widely supported in the empirical and theoretical literature on gene conversion processes (Hilliker et al., 1994; Betran et al., 1997; Song et al., 2011). However, simulations of multi-gene dynamics with constant conversion tract length do not deviate much from simulations with non-constant lengths, in particular in parameter regions where l_c/L is small, thus also the distributions of identity at equilibrium are comparable (see for example Fig. S5).

2.2. Diffusion approximation for a single gene pair

A convenient analytical framework related to the evolution of identity frequencies is the Wright–Fisher model (Fisher, 1930; Wright, 1931) with asymmetric mutation. In the baseline gene family evolution model described above, the L positions in any pairwise alignment would correspond to 'haploid individuals', and the 0/1 identity indices are the analogous representations of 'alleles'. Three dynamic components affect the frequency of identical sites between two genes at each generation: mutation driving towards low identity, gene conversion driving towards high identity, and fluctuations arising from random sampling of nucleotides involved in these processes, analogous to random genetic drift.

Thus, focusing on just one gene pair, for a simpler analytical description of identity evolution, we can consider the quantity $P(x,t)$, denoting the probability that at time t, the pairwise identity between any two genes is x $(0 \le x \le 1)$. Obtaining a formula for $P(x,t)$, directly from the stochastic event-based approach, accounting for all the process details is much more challenging. Denoting the number of identical sites between two genes as Y, their pairwise identity is then the frequency Y/L, a quantity that changes

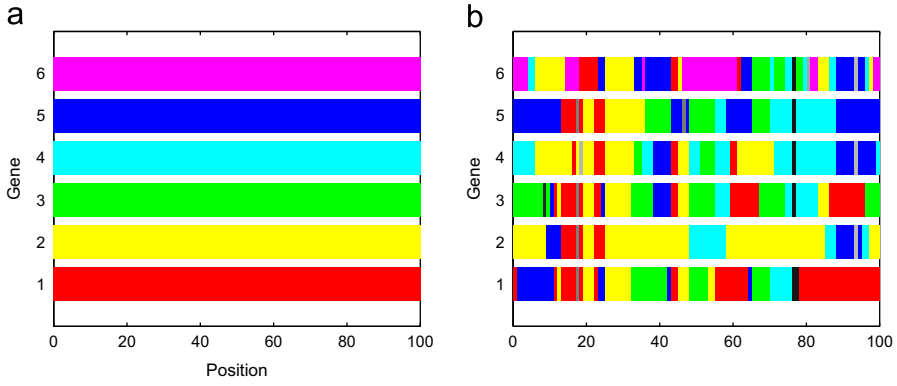

Fig. 1. The gene family changes as a function of stochastic events, comprising partial gene conversions and point mutations. Each gene is denoted by a particular colour at the beginning of the simulation. After conversion events, besides changes in the state sequence of the recipient gene according to the imported segment from the donor, also the colour information of that gene portion in the recipient is altered. The genes appear as mosaics in the long term as a result of imported segments of multiple origin. Mutations are represented in a greyscale, with the more recent mutation being assigned a darker shade. Parameters used: $\gamma = 0.5, \mu = 0.1$, with $l_c = 10, N = 6$, $L = 100, T = 100$. (a) Gene family at $T=0$ and (b) gene family after 100 events. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

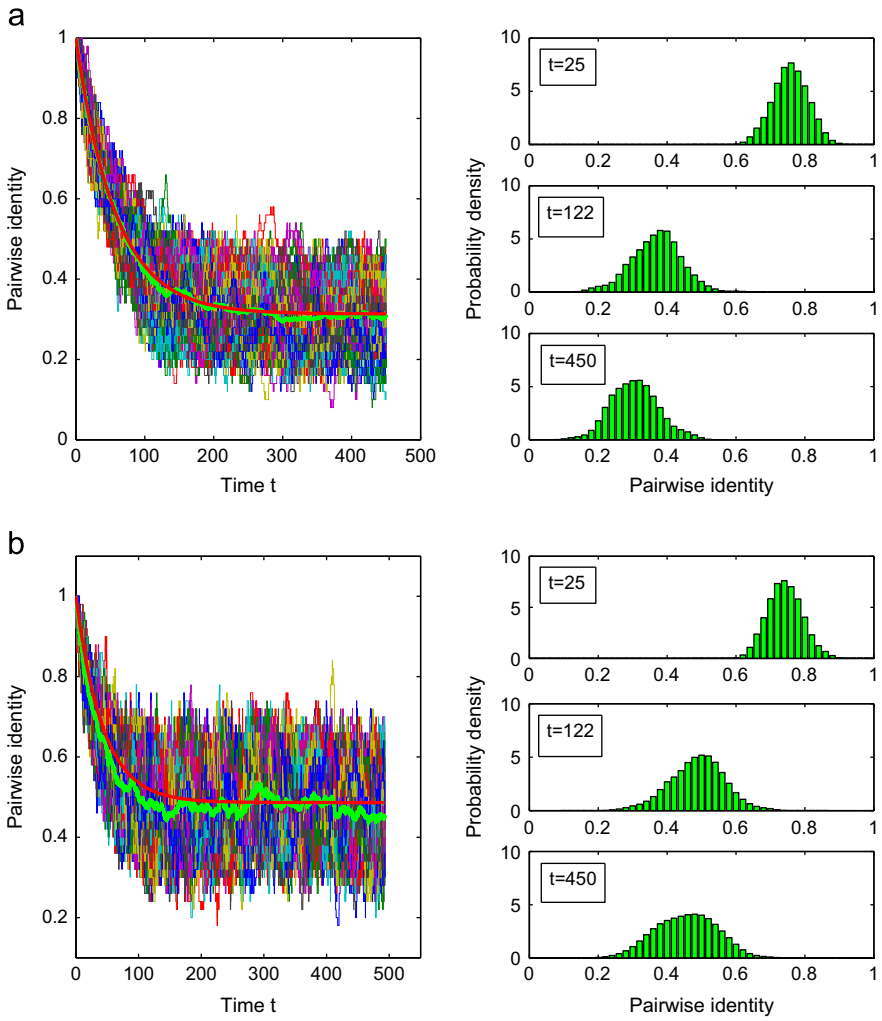

Fig. 2. Dynamics and distribution of pairwise identity in a multigene family at different time points. Each gene pair (coloured line) follows independently the same stochastic processes of conversion and mutation. The simulation mean over gene pairs (green line) is very well approximated by our analytical expression for $\overline{h}(t)$ (given in red). Parameters used in the simulation are $N=20, \mu=0.3, \gamma=0.7$ and $L=50$ and $l_c=1$ (a), $l_c=5$ (b). The equilibrium mean pairwise identity is $\overline{h}^*=0.3133$ (a), and $\overline{h}^*=0.4865$ (b). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Table 1
Summary of parameters and their relative scalings across the two modelling frameworks. Although the discrete event explicit simulation approach is biologically more realistic, the advantage of the diffusion approximation is its analytical tractability.

Discrete event simulation	**Diffusion approximation**
Length of one gene L	Length of one alignment L
Nucleotide state $\{1,2,3,4\}$	Identity state for base pair $\{0,1\}$
Number of genes N	$N=2$ (implicit)
Inter-event times $t_i \sim \exp(1/N(\gamma+\mu))$	Discrete generations (τ)
General continuous time t	Continuous time $\delta t = 1/L \to 0$
Gene pairs: $N(N-1)/2$	One gene pair
Conversion tract length l_c	1 base pair (implicit *per-aligned sites*)
Mutation rate per gene μ (t^{-1})	Probability $1\to0$ per bp per generation
Mutation rate per pair of sites $m=2\mu$	$\nu_{10}=m\tau$
Gene conversion rate per gene γ (t^{-1})	Probability $0\to1$ per bp per generation
Per site pair conversion rate $c=2\gamma l_c/(N-1)$	$\nu_{10}=\tau(c+m/3)$
No limits required	$\lim_{L\to\infty} L\nu_{10}=\theta, \lim_{L\to\infty} L\nu_{01}=\sigma$
Minimum change of identity $1/L$	$x=i/L\in\mathbb{R}$

at each generation of the gene family. When a pair of genes experiences random point mutation, some nucleotides that were identical in the previous generation become different in the next generation ($1\to0$) with per-site probability ν_{10}. When a gene pair undergoes gene conversion, some nucleotides that were originally different become identical ($0\to1$), with per-site probability ν_{01} (see Table 1 for the link of these parameters with simulation).

A key mathematical approach to deal with genetic drift is the diffusion approximation (Fisher, 1922; Wright, 1945; Kimura, 1955; Crow and Kimura, 1970). Under this approximation, the proportion of individuals of a particular allelic type is treated as a continuous random variable whose distribution obeys a diffusion equation. In the following, we describe the dynamics of $\{X(t): t\geq 0\}$, where t denotes time and $X(t)$ the relative frequency of identical positions

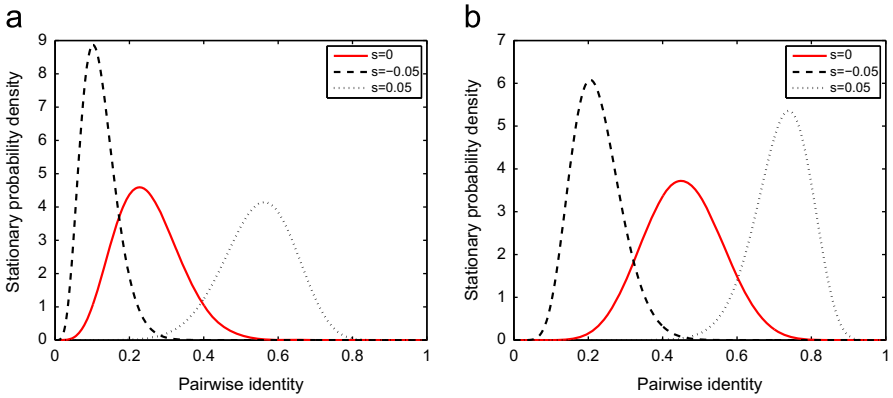

Fig. 3. The stationary distribution of pairwise identity in a multi-gene family as described by the diffusion approximation. (a) $c=0, m=0.03$ and (b) $m=0.02, c=0.01$. Gene length equals $L=300$. In the presence of selection against identity ($s=-0.05$), the mean identity at equilibrium and the variation around it can be much lower than the expected mean in the no-selection case. When there is positive selection for identity instead ($s=0.05$), the distribution shifts to the right, maintaining a similar variance as in the no-selection case if $c=0$ and reducing the variance if $c>0$.

in a typical gene pair. A diffusion process is characterized by two basic quantities: the mean and the variance of the infinitesimal displacement, corresponding to drift and diffusion. Let $Y(n)$ be the number of identical positions in the pairwise alignment of length L at generation n. From random sampling, the transition probability is given by

$$P(Y(n+1)=j|Y(n)=i)=\binom{L}{j}\phi_1^j(1-\phi_1)^{L-j}, \tag{4}$$

where $\phi_1=i/L(1-\nu_{10})+(L-i)/L\nu_{01}$ is the proportion of nucleotides that are of type 1 (identical) after mutation and conversion have occurred. To go from discrete generations to continuous time and calculate explicitly drift and diffusion, one studies the scaled process where $\delta t=1/L$: $X_L(t)=Y(\lfloor Lt\rfloor)/L, t>0$, and $\lfloor Lt\rfloor$ denotes the largest integer less than or equal to Lt. In this formulation, the limit $L\to\infty$ is equivalent to $\delta t\to 0$. The infinitesimal drift parameter $a(x)$ and diffusion parameter $b(x)$ can be easily derived for this model, exploiting $\lim_{L\to\infty}L\nu_{10}=\theta$ and $\lim_{L\to\infty}L\nu_{01}=\sigma$, and substituting $x=i/L$:

$$a(x)=-\theta x+(1-x)\sigma, \quad b(x)=x(1-x), \tag{5}$$

leading finally to the diffusion approximation:

$$\frac{\partial P}{\partial t}=\frac{1}{2}\frac{\partial^2}{\partial x^2}[x(1-x)P]-\frac{\partial}{\partial x}[(-\theta x+\sigma(1-x))P]. \tag{6}$$

Eq. (6) describes how the distribution of pairwise identity changes over time between two interacting genes under neutral evolutionary forces. The initial condition is given by a Dirac delta function at $x=1$, as each gene pair starts at 100% identity. The solution of this equation with no-flux conditions at the boundaries $x=0$ and $x=1$ has been obtained earlier (Crow and Kimura, 1956; Goldberg, 1950). When $t\to\infty$, $P(x,t)$ tends to an equilibrium, $P^*(x)$, yielding a balance between mutation and gene conversion independent of initial conditions:

$$P^*(x)=\frac{\Gamma(2\sigma+2\theta)}{\Gamma(2\sigma)\Gamma(2\theta)}x^{2\sigma-1}(1-x)^{2\theta-1}, \tag{7}$$

where $\theta,\sigma>0$ and Γ denotes the gamma function. The probability density in Eq. (7) corresponds to a beta distribution with mean $\sigma/(\theta+\sigma)$, and variance equal to $\sigma\theta/[(\sigma+\theta)^2(\sigma+\theta+1)]$. Different values of the parameters σ and θ lead to different shapes of the stationary distribution. The ratio θ/σ controls the mean of $P^*(x)$: the higher θ/σ, the closer to 0 the mean is, and vice versa, the closer to 1. The absolute magnitudes of these parameters, instead control the variance of the distribution: low values of θ and σ, lead to a wider distribution, in particular values below 1 lead to a polarized U-shaped distribution. The mathematical formula for the stationary identity distribution illustrates perfectly the counter-balancing effects of mutation and gene conversion. The high identity part of the spectrum ($x>0.5$) is more susceptible to the mutation process, constantly pushing the distribution to the left, whereas the low-identity part of the spectrum ($x<0.5$) is more affected by the gene conversion process, pulling the distribution to the right.

2.2.1. Diffusion approximation with selection

In the presence of selection, favouring one of the two 'alleles' 0/1 per pair of aligned sites between two genes, the diffusion approximation takes a different form. The proportion of nucleotides that are of type 1 (identical) after mutation, conversion and selection have occurred is $\phi_1=(i(1+s)/(i(1+s)+L-i))(1-\nu_{10})+((L-i)/(i(1+s)+L-i))\nu_{01}$, where s is the relative selective advantage of identical sites vs. variable ones. Following the same steps as above, under the assumption that $\lim_{L\to\infty}sL=\lambda$, we obtain

$$\frac{\partial P}{\partial t}=\frac{1}{2}\frac{\partial^2}{\partial x^2}[x(1-x)P]-\frac{\partial}{\partial x}[(\lambda x(1-x)-\theta x+\sigma(1-x))P], \tag{8}$$

with stationary distribution

$$P^*(x)=Ce^{2\lambda x}x^{2\sigma-1}(1-x)^{2\theta-1}, \tag{9}$$

where C is a normalizing constant $C=[\Gamma(2\sigma)\Gamma(2\theta)_1F_1(2\sigma,2\sigma+2\theta,2\lambda)]^{-1}$, where ${}_1F_1$ denotes the hypergeometric function. Two examples of the diffusion approximation with and without selection can be seen in Fig. 3.

Simulation of the Wright–Fisher asymmetric mutation model with selection is straightforward, as one always simulates changes on a pairwise alignment, directly at the 0/1 allele level (mismatch vs. match) with explicit rates. However, implementing selection in the full simulation framework is less straightforward, as the basic units of simulation in that case are individual genes, each with their history and particular nucleotide array of length L (4 alleles). It is possible to include selection with respect to pairwise identity in the multi-gene simulation framework in the form of frequency-dependent selection at each site: first, all genes can start off with some random pairwise differences, then before mutation and gene conversion events take place at each generation, for each position from 1 to L, and each gene 1 to N, a new state is chosen with multinomial probabilities given by f_1,f_2,f_3,f_4 denoting the frequencies of each of the 4 alleles at that position across N genes. If the more frequent allele is chosen with higher probability, then we have positive frequency dependent selection, which selects for

more identity at each site. If the rarer allele out of the four is chosen with higher probability instead, then we would have negative frequency-dependent selection which selects for more diversity at each site.

3. Results

3.1. Comparison of the simulation and diffusion approaches

In order to go from the discrete event framework to the diffusion framework, we need to specify only the generation time for the genes in question and take appropriate scalings of the original parameters (Table 1). Instead, to go from the diffusion approximation to a discrete event formulation, many more characteristics of the system need to be specified. Thus, for each diffusion approximation, there are many discrete event models, making these two approaches only approximately equivalent. It is easy to see that different combinations of event rates per gene γ, μ and specific parameters such as l_c, N and L can lead to the same per-site rates in a pair of genes. For example, doubling the conversion length can be compensated by reducing the rate of gene conversion by a half, thus yielding the same average per-nucleotide rate. But are the evolutionary dynamics of the system the same?

With stochastic events randomly assigned to members of the family during simulation, all gene pairs can be assumed to follow approximately the same stochastic process, giving rise to independent temporal trajectories (Fig. 2). As has been noted earlier in studies of concerted evolution (Ohta, 1982), the diversity among gene members is primarily determined by the balance between mutation, gene conversion and population size. In the simulation framework, increasing gene family size, N, serves to resolve the distribution of identity, i.e. the proportion of gene pairs out of $N(N-1)/2$ at a particular identity level, while increasing the number of sites per gene, L, serves to approximate in a continuous manner shifts in pairwise identity itself after each stochastic event. In our simulations, we notice that the higher the mutation rate per gene in the family, the higher the diversity that emerges and is maintained. In contrast, high conversion rate and long average conversion tracts contribute to maintain higher identity between genes.

Although we do not provide an expression for the variance of pairwise identity in the system, model simulations show that the variance in identity between genes is sensitive to the length of conversion tracts exchanged: if short relative to the total gene length, identity variation between different pairs is smaller, given the same per-nucleotide probability. In contrast, when the converted tracts are longer, even small conversion rates can bring about major fluctuations in pairwise identity. When the number of genes is small, such fluctuations across different gene pairs are correlated and this can introduce a bias and affect the mean equilibrium identity, bringing it below or above what is expected from the average of per-nucleotide rates alone. However, this bias is corrected in large gene families ($N > 5$), where a long conversion tract, despite accelerating diversification between the two interacting genes, affects a smaller proportion of the total pool of gene pairs, hence contributes less to changes in the overall identity distribution. When converted segments are short relative to gene sequence length instead, the mean and variance of pairwise identity appear independent of family size, supporting independent evolution of pairwise identity between family members.

Keeping other parameters fixed, the length of each gene seems to have an effect only on the variance of pairwise identity. Unsurprisingly, if L is short, the possible jumps in identity occur only in steps greater than or equal to $1/L$, increasing the variance. This effect is reduced when considering larger L. Although we can study and describe simulation statistics for many combinations of parameters, the analytical challenge of the exact model compels us to seek approximations in suitable parameter regimes, based on averaging across gene pairs, therefore we use the diffusion approximation. Recalling that the diffusion approximation is based on a specific relation between L and the per-site probabilities of change: $\lim_{L\to\infty} L\nu_{01} = \sigma$ and $\lim_{L\to\infty} L\nu_{10} = \theta$, the match between the simulations and the diffusion approximation is first dependent on the fulfilment of this scaling criterion.

The approximation derived in section 2.2. for the pairwise identity probability distribution can be tested by comparison of the theoretical predictions to the results of computer simulations of the gene conversion and mutation events. To obtain the stationary identity distribution from simulations of the explicit conversion/mutation events, we first have to estimate the approximate time and the number of events it takes the system to reach equilibrium, given c and m. From Eq. (2), we can compute the time it takes for a typical pair to reach identity level within α of the equilibrium value: $t_\alpha = -L\ln(\alpha/m(c+4m/3))/(c+4m/3)$. Thus, in a family of size N, given conversion length l_c and generation time τ, the minimal number of events to be simulated until equilibrium is approximately $T_{min} = t_\alpha N[c(N-1)/(2l_c)+m/2]$. We usually assume $\alpha = 0.05$ and simulate a number of events equal to $6T_{min}$ and compute the stationary distribution from the last T_{min} events, for different combinations of parameter values. We compute the distribution from the proportion of gene pairs at given identity levels between 0 and 1 in steps of $1/L$ over consecutive continuous time intervals corresponding to these last events in the stationary phase. The final simulation-derived distribution is calculated by

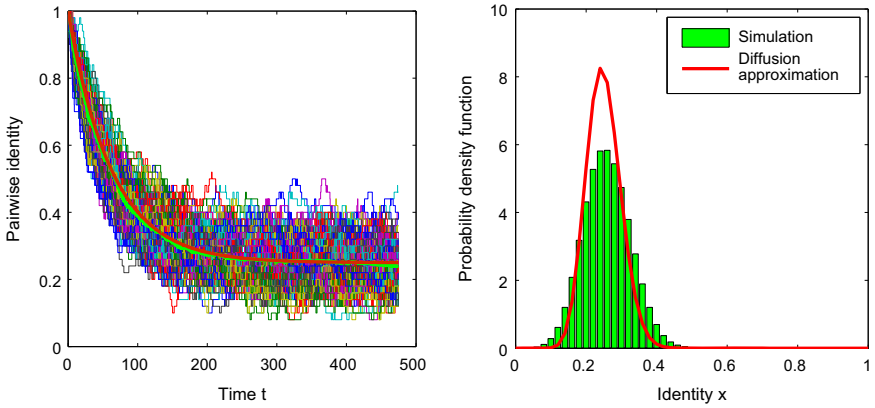

Fig. 4. Illustration of simulated pairwise trajectories toward the stationary identity distribution (single run) and the diffusion approximation. Mutation-only case in a multi-gene family with parameters: $\mu = 0.3, L = 50, N = 20$.

taking the average of these consecutive distributions, which should remove correlations between consecutive time points along one pairwise trajectory. Notice that there is stochasticity in the order, timing, location, and type of different events, even when simulating with the same parameters. This has implications for the resulting distribution (see Fig. S1 for an example). Thus, to get a more accurate picture of the match between the two formulations, we can simulate many runs from the same initial conditions, as opposed to a single run, for each parameter combination, and check whether the diffusion approximation is contained in the 90% confidence region predicted by the simulations (Fig. 6).

Unsurprisingly, for the mutation-only case, the two approaches agree very well. Fig. 4 shows the stationary distribution resulting from one run of the gene family evolutionary dynamics through mutation, with the diffusion approximation superimposed. As the number of genes increases, variance between different runs is reduced and the two distributions come closer (Fig. S2). Although for a given combination of c, m and L, there is only one diffusion approximation, the exact choice of conversion length l_c and number of genes N on the simulation-derived distribution leads to a variable match between the two in some cases. In general, the diffusion approximation agrees well with the simulation model for N and L large, e.g. $N \geq 6$, $L \geq 50$, as the correlations between individual trajectories of gene pairs vanish in these limits and identity changes in smaller discrete jumps per unit of time (see Fig. S3).

Another requirement to obtain a good match between the theoretical and simulation-derived distribution is that of small conversion length relative to L, for example when $l_c/L \leq 0.05$. Comparisons of theoretical and simulated distributions for a single stochastic run are given in Fig. 5 and numerical comparisons of parameter estimates for a set of true values are described in Table S1. Notice, the requirement of small l_c/L is stronger in the parameter regimes where mutation events are more frequent than gene conversion events, because longer conversion tracts in that case transport a higher number of substitutions to the recipient gene and affect more the new pairwise identity of that gene with other genes in the family. However, our simulations show that even for l_c in the range of 10% of total gene length, when the gene length is large, the diffusion approximation provides a reasonably accurate description of identity at equilibrium (Fig. S3). For illustration, we have also simulated the case of a geometric distribution for conversion length, but this has not resulted in major deviations from the expectations with a constant conversion length (Fig. S4). In fact, if the mean conversion length is small relative to total length (5–10%), and if the total gene length L is large, we have a very good match between the diffusion approximation and the stationary identity distributions from simulations (Fig. S5), reflected also in parameter estimates close to the true values (Table S2). We can expect that for a gene sequence of length about 1000 bases, the diffusion approximation should hold for a range of conversion lengths up to 50–100 nucleotides.

3.2. Application: VSG gene family of African trypanosomes

We apply the diffusion approximation to the diversity of genes in the trypanosome antigen gene family assuming the current identity distribution reflects equilibrium. African trypanosomes provide one of the most prominent antigenic variation examples among parasites. The antigen genes coding for the variable surface protein coat of trypanosomes are the Variant Surface Glycoprotein (*VSG*) genes and they form a multigene family. As antibody responses against an antigenic variant accumulate, parasites switch to expression of a new variant, evading current host immunity and producing a distinct wave of parasitaemia. Through repeated iterations of this process, where only one VSG gene is expressed at a time, the African trypanosome is able to sustain chronic infection in its hosts for long periods (see Lythgoe et al., 2007; Gjini et al., 2010 for within-host models of this process). The structure and organisation of the huge *VSG* archive ($\geq$ 1600 genes) and molecular mechanisms that govern its expression have been studied extensively (Barry, 1997; Borst et al., 1997; Morrison et al., 2005), yet our quantitative understanding of the evolutionary forces involved remains limited. Genomic sequence determination of *Trypanosoma brucei* has now greatly expanded the *VSG* sequence dataset, allowing for more in-depth analyses of

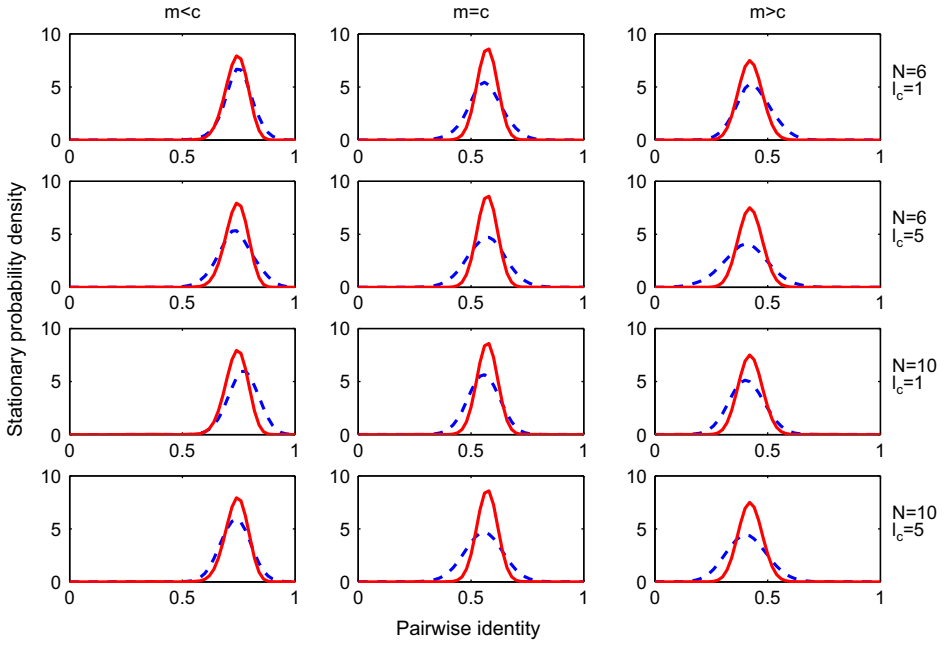

Fig. 5. Comparison of the diffusion approximation with simulation-derived identity distributions from a single run at equilibrium. The theoretical diffusion approximation (solid line) is superimposed on the distribution derived from discrete event simulations in the gene family (dashed line) for cases when the per-aligned site probabilities satisfy: $m = 0.2, c = 0.5$ (first vertical panel); $m = c = 0.5$ (middle vertical panel); $m = 0.5, c = 0.2$ (third vertical panel) and family size N and conversion length l_c vary in each row. γ and μ are adjusted in each case to keep m and c constant, $L = 50$.

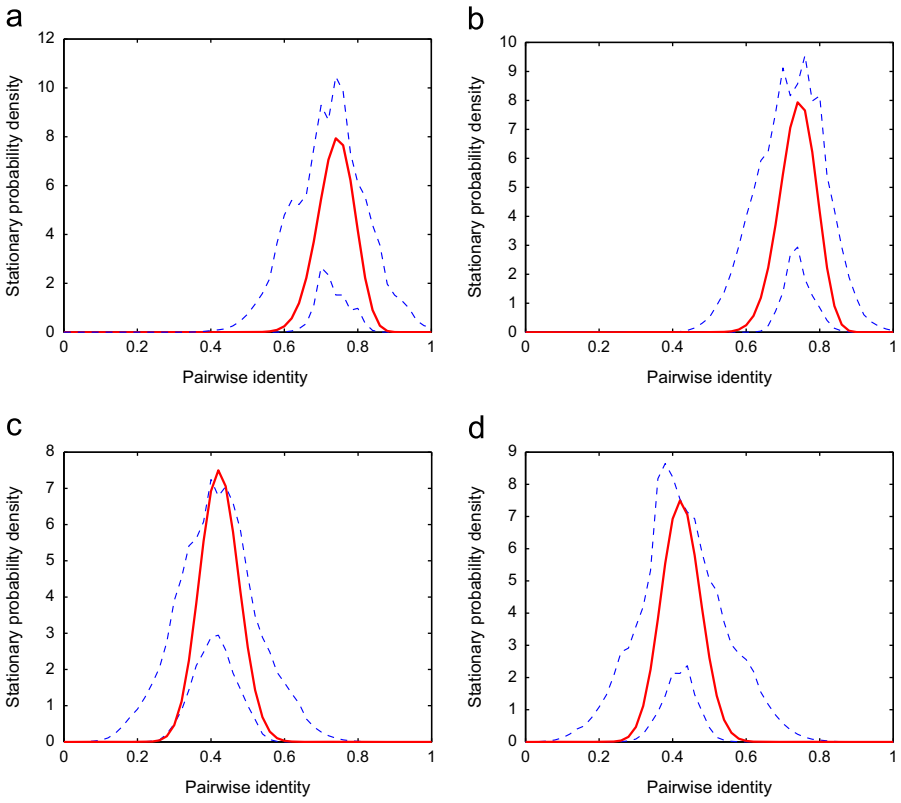

Fig. 6. Confidence regions for the stationary identity distribution obtained from simulations and the diffusion approximation. For each parameter combination 50 independent runs are simulated and stationary distributions are pooled together. The dashed lines represent 5% lower confidence bound and the 95% upper confidence bound. We use $L=50, l_c=4$. The diffusion approximation is given by the solid line, typically contained within 90% of the simulations. (a) $N=4, c=0.5, m=0.2$; (b) $N=6, c=0.5, m=0.2$; and (c) $N=4, c=0.2, m=0.5$; (d) $N=6, c=0.2, m=0.5$.

sequence variability, which might shed light on the balance between structural constraints, the processes of epitope diversification and the potential for antigenic variation in this parasite (Carrington et al., 1991). *VSG* genes contain an N-terminal domain of 350–400 residues, encoding the portion of the *VSG* protein that contains exposed surface loops with variable antigen epitopes (Miller et al., 1984; Hsia et al., 1996), and a more conserved C-terminal domain of 40–80 residues, encoding the part of the protein anchored to the plasma membrane (Carrington et al., 1991). N-terminal domains form three subfamilies (*nA*, *nB*, *nC*) on the basis of their phylogenetic structure, as shown by Marcello and Barry (2007b). Here, as an illustration of our quantitative approach, we consider diversification of *nA* and *nB* genes, within a single genome.

Using pairwise identity data from the *nA* and *nB VSG* subfamilies (gene sequences available on the *VSG*db database: www.vsgdb.net), obtained by aligning respectively their $N=412$ and $N=362$ gene sequences (Marcello and Barry, 2007a), we apply the diffusion approximation to estimate mutation and gene conversion rates in a real biological context. Allowing for variation across the *VSG* family, we first apply the model separately to nA and nB N-terminal domains and then to their combined distribution. We assume that evolutionary parameters are constant over time. Each N-domain varies in length between 900 and 1050 nucleotides, thus we use the average $L=975$. We first estimate σ and θ as the global rates of diversification by fitting the empirical identity distribution of *VSG* N-domains, to the theoretical formula for $P^*(x)$ given by Eq. (7). The fit is performed using a nonlinear least square routine in MATLAB® (The MathWorks, 2011). The best-fitting estimates are the ones that minimize the sum of squared deviations of the empirical distribution from the theoretical formula. To test for selection and estimate selection strength, we subsequently also apply the diffusion approximation with selection (Eq. (8)) to the empirical distribution and compare the results. The details of the fits and the estimated diffusion approximation parameters are presented in Fig. 7 and Table 2.

For nA gene sequences, given their extremely high nucleotide diversity we obtain $m\tau=0.0424$, (95% CI: 0.0438, 0.0452) from the diffusion parameter θ and derive $c\tau=\sigma/L-m\tau/3$, resulting in a negative estimate for this parameter equal to -0.0017, (95% CI: -0.0008, 0.0001). For nB gene sequences we obtain $m\tau=0.0267$ (95% CI: 0.0276, 0.0285) and $c\tau=0.0003$ (95% CI: 0.0009, 0.0015), supporting conversion event rates per pair of aligned sites in these genes about two orders of magnitude less frequent than mutation events. When the confidence intervals for the conversion event probability per base pair per generation contain zero, we can infer that this particular multigene family data supports only the action of random mutation. Indeed, the mean identity of about 25% in both nA and nB N-domains, matches the expectation of maximum identity attainable under the neutral mutation-only model in a 4-allele system. If we treat nA and nB N-domains together, we obtain $m\tau=0.0324$, (95% CI: 0.0316 , 0.0332) and $c\tau=2.4479\times10^{-5}$ (95% CI: -4.97×10^{-4}, 5.45×10^{-4}), again favouring a scenario of hypermutation with almost no signature of gene conversion, that could be driving the extreme genomic diversification we see in N-domains of *VSG* genes.

When we allow for the combined action of mutation, selection, and gene conversion, the model fitting procedure estimates much lower mutation probabilities per bp per generation between 2.12×10^{-16} and 7.78×10^{-9}, and high conversion probabilities per bp per generation of the order of 2.2×10^{-2}, counterbalanced by strong selection against identity in the range between 0.06 (nB) and 0.09 (nA) per bp per generation. Such range of mutation rates

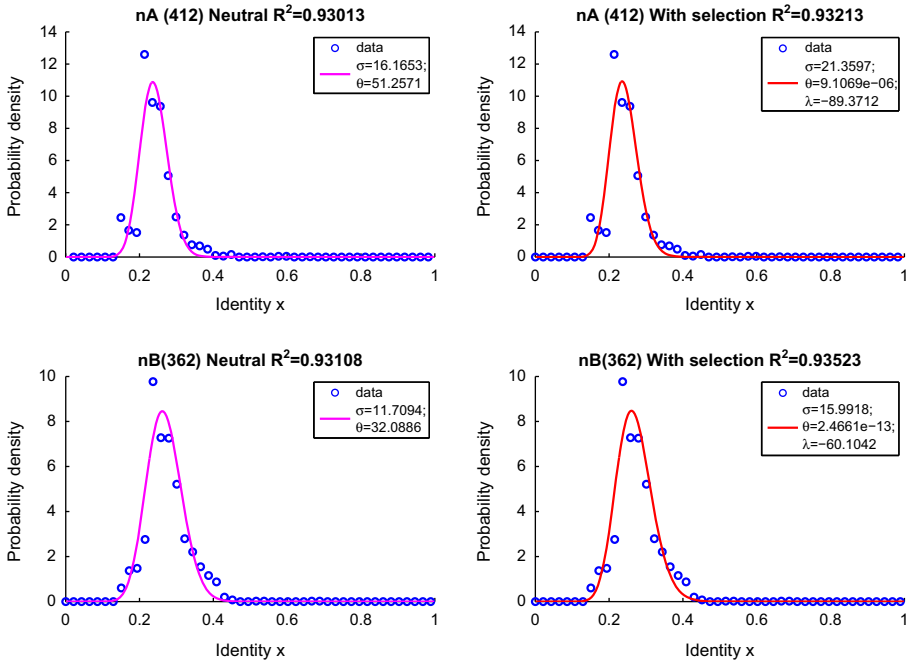

Fig. 7. The diffusion approximation fit for two N-domain subfamilies (*nA*, *nB*). The best fitting parameters were found by applying *lsqnonlin*, a nonlinear least squares optimization routine in MATLAB, to each empirical distribution (points): without selection (left panels) as specified by Eq. 7, and with selection (right panels) as specified by Eq. 9. The solid lines give the best diffusion approximations. R^2 denote the coefficients of determination for each model fit.

Table 2
Estimates of evolutionary parameters under the neutral and selection models for the two N-terminal domain types (nA and nB) of *VSG* genes in African trypanosomes. When we scale σ and θ derived from the fit of the diffusion approximation, by $1/L$, we can obtain the effective $0\rightarrow1$ and $1\rightarrow0$ transition probabilities per base-pair per generation: $\nu_{10}=\theta/L=\tau m$, $\nu_{01}=\sigma/L=\tau(c+m/3)$, where τ is the generation time of the parasite.

Parameter	*nA*	*nB*	**nA+nB**
Neutral model			
θ	51.2571	32.0886	37.7868
95% CI	[49.6269, 52.8874]	[31.0845, 33.0926]	[36.8698, 38.7038]
σ	16.1653	11.7094	12.6242
95% CI	[15.6560, 16.6746]	[11.3474, 12.0713]	[12.3214, 12.9269]
Model with selection			
θ	9.11×10^{-6}	2.5×10^{-13}	1.1019×10^{-13}
95% CI	[−48.7994, 48.7995]	[−21.6118, 21.6118]	[−22.1915, 22.1915]
σ	21.3597	15.9918	16.8500
95% CI	[16.4515, 26.2679]	[13.0462, 18.9374]	[14.3295, 19.3706]
λ	−89.3712	−60.1042	−67.5652
95% CI	[−174.2099, −4.5326]	[−100.7529, −19.4555]	[−107.3483, −27.7821]

per base pair seems comparable with the observed spontaneous mutation rates per base pair in specific loci of higher eukaryotes (Drake et al., 1998), although what we find is closer to the lower end of the magnitude spectrum. Notice that in this crude approach, a uniform selective pressure across all sites along the sequences analyzed is assumed and no information about the underlying DNA or amino-acid composition is used. Other statistical techniques such as the ratio of nonsynonymous/synonymous substitution rates might be more suitable for detecting with confidence purifying, neutral or positive selection. However, with the approaches described above, simple and quick evaluation of different hypotheses concerning the evolutionary dynamics of multigene families can be performed as a first step, which can be subsequently validated with other more complex datasets from multiple genomes and more sophisticated statistical tools.

4. Discussion

We have presented a new modelling framework for the study of global genetic diversification in a gene family shaped primarily by mutation, gene conversion and genetic drift that integrates simulation of discrete genetic events with population genetics models, such as the Wright–Fisher asymmetric mutation model and the diffusion approximation. Current approaches in the study of multi-gene families typically focus on small family sizes and treat genetic events in terms of per-site average rates only. Here we have explored characteristics of evolutionary processes and their implications for genetic diversification in larger gene families and making explicit assumptions about gene conversion tract length, overall gene length, mutation effects and the time-scale of discrete event occurrence and distribution. The simulation of genetic events is likely to be closer to biological reality, but it is

more challenging analytically as the signature of identity evolution is confounded by gene conversion and correlations between gene pairs. Under suitable parameter regimes, however, including long gene sequences, many genes, and small relative conversion lengths, we should be able to infer the per-site mutation and gene conversion rates from the stationary pairwise identity distribution in the multi-gene family. The mean of such a distribution contains information about the ratio between the basic rates of these opposing processes, while the variance has information about their combined sum and genetic drift. The use of the diffusion approximation however is likely to bias towards small values of relative conversion tract length, l_c/L, which may or may not correspond to the truth of a given system; this could therefore bias estimates of gene conversion probability per nucleotide c regardless of gene length L. For trypanosome VSG N-domains, we have recently inferred in another study relatively short, geometrically distributed conversion tract lengths (Gjini et al., 2012), in the range of 14–80 bp, underlying local diversification of highly related sequences, thus the application of the diffusion approximation here to this global dataset is not unreasonable.

Although we have focused on the stationary distribution of identity that is maintained in a gene family at equilibrium, with the simulation framework proposed, it is possible to study in detail also temporal characteristics of the process of diversification, the interaction network between gene family members, emerging phylogenetic relationships shaped by mutation and gene conversion and the explicit sequence content in terms of nucleotides. The simulation model can be extended to study different types of genetic processes, such as identity-driven gene conversion events, where only gene pairs that are sufficiently similar can interact via gene conversion. Recombination processes dependent on the pairwise identity between the interacting sequences have been found to occur across many systems (Liskay et al., 1987; Datta et al., 1997; Majewski and Cohan, 1999; Chen et al., 2007). However, to appropriately include such processes in the diffusion approximation framework, requires assuming an explicit function when making the conversion probability c dependent on the current identity level, for example a linear function or an exponential, which would change significantly the current drift term. It is thus necessary to obtain empirical gene conversion data on the genes in consideration that can motivate the use of explicit functions for the sequence identity dependence before a model extension of this kind can be applied. Similarly, a different distribution of gene conversion tract lengths, such as the geometric distribution, could be used. Since the diffusion approximation is dependent only on the mean conversion length, such extension should not affect the final analytical equilibrium that we obtain. However, in some parameter regimes, as we have shown, the variance of the stationary distribution is more sensitive to the mean conversion length and its variance, in particular when conversion events are rarer than mutation events, and further work in this direction is needed.

Despite the mathematical techniques employed in this study being classical, the application of this theory to the *VSG* antigenic archive of African trypanosomes is novel. By using first an explicit simulation approach for visualizing and intuitively understanding the occurrence of genetic processes in a multigene family, and then applying a diffusion approximation to the distribution of gene pairs at various identity levels, assumed to be stationary, we are able to quantify the evolutionary rates of mutation and gene conversion for N-terminal *VSG* domains, consisting mainly of pseudogenes. Invoking neutral molecular evolution at first (Kimura, 1985), we were able to capture the distribution of pairwise genetic identity in the *VSG* antigen archive by a simple difference in mutation and gene conversion rates. Our estimates under this model support very high probabilities of mutation per base pair per generation within genome in the absence of selection, in the range 2.6×10^{-2}–4.2×10^{-2}, and very little sign of homogenizing gene conversion. Although the diffusion approximation appears not to detect signature of gene conversion in terms of per-site probabilities, there exists a possibility that random gene conversion of considerable lengths may be occurring at high frequency and thereby spreading mutations sufficiently fast across all genes, so that the homogenization signal is lost over long timescales. What remains over long timescales is the signature of an elevated mutation rate on *VSG* N-terminal domains, making it a hypervariable genomic region. This is likely to have critical implications for antigenic variation within hosts, especially in the chronic stages of infection, by providing ready genetic and antigenic diversity available for the expression of mosaic genes (Marcello and Barry, 2007b). It is interesting to notice that the range of mutation rate estimated for N-domains under the neutral model very closely matches that of spontaneous antigen switching during trypanosome infections, found to be in the order of 10^{-2} per cell division (Turner and Barry, 1989).

When considering the possibility that selection may be acting to favour hypervariability between gene sequences in the VSG archive, our estimates of spontaneous mutation rates are much lower, at most in the order of 7.8×10^{-9} per bp per generation (nA N-domains), but they are compensated by very high negative selection for identity ($s=0.06$) that maintains the stationary distribution shifted towards the low-identity spectrum. This model can also explain the observed variability in N-domains data, but such strong selection on genomic regions that consist mainly of pseudogenes may be questionable, and also harder to reconcile perhaps with the observation that any one *VSG* gene is expressed rarely and for little time within a host, thus the selection pressure on these genes is expected to be minimal (Barry et al., 2012). However, diversifying selection has been found in epitope-coding genes across many pathogen systems, most notoriously in viruses (Paolo et al., 1999; Fares et al., 2001; Haydon et al., 2001; Anisimova and Yang, 2004), but also in other protozoa such as *Plasmodium falciparum* (Hughes, 1992; Polley and Conway, 2001; Baum et al., 2003). Thus, in light of this evidence, finding signatures of selection in antigen genes of African trypanosomes would not be that surprising. However, more thorough analyses and polymorphism data from field isolates are needed to ultimately ascertain the mode and action of selection in this multi-gene system and to tease apart better the neutral evolution from the diversifying selection hypotheses. In reality, members of a gene family are subjected to a considerably more complex set of processes than the ones covered by our models, thus the estimation process proposed here should be taken with care and augmented with different methodologies when available or possible.

Throughout our models and analyses, an important assumption was that the rates of gene conversion and mutation are constant. In fact, across organisms, rates of spontaneous mutation vary widely and specific evolutionary forces are involved in shaping them (Drake et al., 1998). An interesting question that arises also for trypanosomes is whether these evolutionary rates are subject to selection and to what extent they are adaptive. Mutation rates must depend on an evolutionary compromise between the need to create diversity – a basis for adaptive evolution as in the case of surface protein genes – and the requirement to preserve core or essential cellular functions, as in the case of genes encoding nuclear proteins. A challenging avenue for further exploration is to link such genetic processes at the DNA level to the explicit functions they encode, such as antigenicity, attachment properties, growth regulation, motility, or virulence, and finally to the fitness effects they impose on an organism as a whole.

Future models considering identity and differences in the gene family at the amino-acid level can help resolve such questions.

Are there any environmental cues within hosts that trigger higher mutation or alternatively higher conversion rates in the *VSG* genes of trypanosomes? Are there any structural genomic constraints that limit the generation of diversity, and if yes, what are their features? Understanding these aspects of *VSG* archive diversification could prove crucial in the design of control strategies, for example drugs that interfere with the capacity of the pathogen to mutate or diversify. In general, despite advances driven by molecular biology and genomics, there is a need to gain a deeper understanding of key mechanisms that may facilitate generation of diversity across biological systems and scales. The characterisation of parasite surface protein families or other multigene families on the basis of their capacity to generate variation is important, and necessitates quantitative models for explaining the role of the genetic processes involved. As illustrated in this paper, mathematical frameworks may be implemented to explore, visualise and estimate the dynamic diversification capacity of gene families.

Acknowledgements

The authors would like to thank Nick Savill for useful feedback at earlier stages of the manuscript. This work was funded by the Wellcome Trust (Grant number 055558). The Wellcome Trust Centre for Molecular Parasitology is supported by core funding from the Wellcome Trust (Grant number 085349). EG was supported by the Kelvin-Smith PhD fellowship scheme by the University of Glasgow (2007–2011).

Appendix A. Dynamics of mean pairwise identity

A.1. As a function of stochastic events

The expected changes in nucleotide identity between any two genes can be approximated as follows. Denote the current number of identical sites shared by a gene pair after T events by $n(T)$. An identical nucleotide can be gained with per-site probability $1-n(T)/L$, or lost with per-site probability $n(T)/L$. Gene conversion and point mutation happen with respective probabilities $\gamma/(\gamma+\mu)$ and $\mu/(\gamma+\mu)$. Note that a gene conversion involves l_c nucleotides out of L, whereas point mutation in either member of the pair affects only one site, with 1/3 probability of change to the same type as in the other gene. Thus the expected number of identical sites in a gene pair after $T+1$ events is given by

$$n(T+1)=n(T)+\left[\frac{2\gamma l_c}{(\gamma+\mu)(N-1)}+\frac{2\mu}{3(\gamma+\mu)}\right]\left[1-\frac{n(T)}{L}\right]-\frac{2\mu}{(\gamma+\mu)}\frac{n(T)}{L}. \tag{A.1}$$

With initial condition $n(0)=n_0$, the solution of the above recurrence relation is

$$n(T)=\left[1-\frac{2}{L(\gamma+\mu)}\left(\frac{\gamma l_c}{N-1}+\frac{4\mu}{3}\right)\right]^T\left[n_0-L\frac{\frac{\gamma l_c}{N-1}+\frac{\mu}{3}}{\frac{\gamma l_c}{N-1}+\frac{4\mu}{3}}\right]+L\frac{\frac{\gamma l_c}{N-1}+\frac{\mu}{3}}{\frac{\gamma l_c}{N-1}+\frac{4\mu}{3}} \tag{A.2}$$

for the expected number of identical nucleotides between two genes after T stochastic events. The mean pairwise identity follows easily from (A.2) from the ratio $n(T)/L$.

A.2. As a function of continuous time

Pairwise identity can be described as a function of *continuous time*, $t\in\mathbb{R}$. Consider how $n(t+\Delta t)$ depends on $n(t)$, where Δt is an infinitesimal time step. Recall that γ and μ are event rates per gene per unit of time. Thus in the time interval $[t, t+\Delta t]$, we expect $\gamma\Delta t$ conversion events and $\mu\Delta t$ mutation events on average. After conversion and mutation have occurred, the expected number of identical sites between two genes is

$$n(t+\Delta t)=n(t)+\Delta t\left[\frac{2\gamma l_c}{N-1}+\frac{2\mu}{3}\right]\left[1-\frac{n(t)}{L}\right]-2\mu\Delta t\frac{n(t)}{L}. \tag{A.3}$$

We can substitute $c=2\gamma l_c/(N-1)$, $m=2\mu$. Then, rearranging, dividing both sides by Δt and taking the limit $\Delta t\to 0$ gives a differential equation whose solution is

$$n(t)=n(0)\frac{c+m/3+me^{-(c+4m/3/L)t}}{c+4m/3}. \tag{A.4}$$

Dividing $n(t)$ by L yields mean pairwise identity in continuous time $h(t)$.

Appendix B. Supplementary material

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2013.10.001.

References

Anisimova, M., Yang, Z., 2004. Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? Journal of Molecular Evolution 59 (6), 815–826.
Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent segmental duplications in the human genome. Science 297 (5583), 1003–1007.
Barbet, A.F., Kamper, S., 1993. The importance of mosaic genes to trypanosome survival. Parasitology Today 9, 63–66.
Barry, J., 1997. The relative significance of mechanisms of antigenic variation in African trypanosomes. Parasitology Today 13, 212–218.
Barry, J.D., Hall, J.P., Plenderleith, L., 2012. Genome hyperevolution and the success of a parasite. Annals of the New York Academy of Sciences 1267 (1), 11–17.
Baum, J., Thomas, A.W., Conway, D.J., 2003. Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. Genetics 163 (4), 1327–1336.
Betran, E., Rozas, J., Navarro, A., Barbadilla, A., 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146, 89–99.
Borst, P., Rudenko, G., Blundell, P., Van Leeuwen, F., 1997. Mechanisms of antigenic variation in African trypanosomes. Behring Institute Mitteilungen 99, 1–15.
Carrington, M., Miller, N., Blum, M., Roditi, I., Wiley, D., Turner, M., 1991. Variant specific glycoprotein of *Trypanosoma brucei* consists of two domains each having an independently conserved pattern of cysteine residues. Journal of Molecular Biology 221, 823–835.
Chen, J., Cooper, D., Chuzhanova, N., Férec, C., Patrinos, G., 2007. Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics 8, 762–775.
Crow, J., Kimura, M., 1956. Some genetic problems in natural populations. In: Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability, vol. 4, pp. 1–22.
Crow, J., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper & Row, New York.
Datta, A., H.M., Lipsitch, M., Jinks-Robertson, S., 1997. Dual roles for dna sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc. Natl. Acad. Sci. USA 94, 9757–9762.
Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F., 1998. Rates of spontaneous mutation. Genetics 148 (4), 1667–1686.
Fares, M.A., Moya, A., Escarmís, C., Baranowski, E., Domingo, E., Barrio, E., 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (fmdv) subjected to experimental passage regimens. Molecular Biology and Evolution 18 (1), 10–21.
Fisher, R., 1922. On the dominance ratio. Proceedings of the Royal Society of Edinburgh 42, 321–431.
Fisher, R., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Gillespie, D., 1977. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry 81 (25), 2340–2361.

Gjini, E., Haydon, D., Barry, J., Cobbold, C., 2010. Critical interplay between parasite differentiation, host immunity, and antigenic variation in trypanosome infections. The American Naturalist 176 (4), 424–439.
Gjini, E., Haydon, D.T., Barry, J.D., Cobbold, C.A., 2012. The impact of mutation and gene conversion on the local diversification of antigen genes in African trypanosomes. Molecular Biology and Evolution 29 (11), 3321–3331.
Goldberg, S., 1950. Ph.D. thesis. Cornell University.
Griffiths, R., Watterson, G., 1990. The number of alleles in multigene families. Theoretical Population Biology 37 (1), 110–123.
Haydon, D.T., Bastos, A.D., Knowles, N.J., Samuel, A.R., 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. Genetics 157 (1), 7–15.
Hilliker, A., Harauz, A., Reaume, M., G.S., Clark, S., Chovnick, A., 1994. Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics 137(4), 1019–1026.
Hsia, R., Beals, T., Boothroyd, J., 1996. Use of chimeric recombinant polypeptides to analyse conformational, surface epitopes on trypanosome variant surface glycoproteins. Journal of Molecular Microbiology 19, 53–63.
Hughes, A.L., 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (msa-1) locus of *Plasmodium falciparum*. Molecular Biology and Evolution 9 (3), 381–393.
Innan, H., 2002. A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. Genetics 161 (2), 865–872.
Innan, H., 2009. Population genetic models of duplicated genes. Genetica 137 (1), 19–37.
Kimura, M., 1955. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbour Symposium on Quantitative Biology 20, 33–53.
Kimura, M., 1985. The Neutral Theory of Molecular Evolution. Cambridge University Press.
Kimura, M., et al., 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genetics Research 11, 247–269.
Kimura, M., Crow, J.F., 1964. The number of alleles that can be maintained in a finite population. Genetics 49 (4), 725.
Liskay, R., Letsou, A., Stachelek, J., 1987. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics 115, 161–167.
Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290 (5494), 1151–1155.
Lythgoe, K., M.L.J., Read, A., Barry, J., 2007. Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. Proceedings of the National Academy of Sciences of the United States of America 104(19), 8095–8100.
Majewski, J., Cohan, F., 1999. DNA sequence similarity requirements for interspecific recombination in bacillus. Genetics 153 (4), 1525–1533.
Mano, S., Innan, H., 2008. The evolutionary rate of duplicated genes under concerted evolution. Genetics 180, 493–505.
Mansai, S.P., Kado, T., Innan, H., 2011. The rate and tract length of gene conversion between duplicated genes. Genes 2 (2), 313–331.
Marcello, L., Barry, J., 2007a. Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. Genome Research 17 (9), 1344–1352.
Marcello, L., Barry, J., 2007b. From silent genes to noisy populations—dialogue between the genotype and phenotypes of antigenic variation. Eukaryotic Microbiology 54 (1), 14–17.
Miller, E., Allan, L., Turner, M., 1984. Topological analysis of antigenic determinants on a variant surface glycoprotein of *Trypanosoma brucei*. Journal of Molecular and Biochemical Parasitology 13, 67–81.
Morrison, L., Majiwa, P., Read, A., Barry, J., 2005. Probabilistic order in antigenic variation of *Trypanosoma brucei*. International Journal for Parasitology 35, 961–972.
Nagylaki, T., 1984. Evolution of a large population under gene conversion. Proceedings of the National Academy of Sciences of the United States of America 80, 5941–5945.
Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, Berlin.
Ohta, T., 1976. Simple model for treating evolution of multigene families. Nature 191, 74–76.
Ohta, T., 1982. Allelic and non-allelic homology of a supergene family. Proceedings of the National Academy of Sciences of the United States of America 79, 3251–3254.
Ohta, T., 1983. On the evolution of multigene families. Theoretical Population Biology 23, 216–240.
Ohta, T., 2010. Gene conversion and evolution of gene families: an overview. Genes 1 (3), 349–356.
Paolo, M.d.A., Kallas, E.G., de Souza, R.F., Holmes, E.C., 1999. Genealogical evidence for positive selection in the nef gene of hiv-1. Genetics 153 (3), 1077–1089.
Polley, S.D., Conway, D.J., 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics 158 (4), 1505–1512.
Song, G., Hsu, C.-H., Riemer, C., Miller, W., 2011. Evaluation of methods for detecting conversion events in gene clusters. BMC Bioinformatics 12 (Suppl 1), S45.
Tachida, H., Kuboyama, T., 1998. Evolution of multigene families by gene duplication: a haploid model. Genetics 149 (4), 2147–2158.
The MathWorks, Natick, MA, 2011. MATLAB R2011a.
Turner, C., Barry, J., 1989. High frequency of antigenic variation in *Trypanosoma brucei* rhodesiense infections. Parasitology 99 (01), 67–75.
Walsh, B., 1983. Role of biased gene conversion in one-locus neutral theory and genome evolution. Genetics 105, 461–468.
Walsh, J.B., 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion?. Genetics 117 (3), 543–557.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.
Wright, S., 1945. The differential equation of the distribution of gene frequencies. Proceedings of the National Academy of Sciences of the United States of America 31, 382–389.