Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Discussion

Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists

Gabriela Aguileta [a,b], Guislaine Refrégier [a,b], Roxana Yockteng [c], Elisabeth Fournier [d], Tatiana Giraud [a,b,*]

[a] *Ecologie, Systématique et Evolution, Université Paris-Sud, F-91405 Orsay cedex, France*
[b] *Ecologie, Systématique et Evolution, CNRS F-91405 Orsay cedex, France*
[c] *UMR 7205, CNRS-MNHN, Origine, Structure et Evolution de la Biodiversité, Département Systématique et Evolution, 16 rue Buffon CP 39, 75005, Paris, France*
[d] *UMR BGPI, TA A 54/K, Campus International de Baillarguet, 34398 Montpellier cedex 5, France*

ARTICLE INFO

Article history:
Received 28 October 2008
Received in revised form 14 March 2009
Accepted 26 March 2009
Available online xxx

Keywords:
Avirulence genes
Toxins
Antigens
Effector
Elicitor
Resistance genes
Polymorphism
Genomic scan
Gene-for-gene
Parasite
Plasmodium

ABSTRACT

The ongoing coevolutionary struggle between hosts and pathogens, with hosts evolving to escape pathogen infection and pathogens evolving to escape host defences, can generate an ‘arms race’, i.e., the occurrence of recurrent selective sweeps that each favours a novel resistance or virulence allele that goes to fixation. Host–pathogen coevolution can alternatively lead to a ‘trench warfare’, i.e., balancing selection, maintaining certain alleles at loci involved in host–pathogen recognition over long time scales. Recently, technological and methodological progress has enabled detection of footprints of selection directly on genes, which can provide useful insights into the processes of coevolution. This knowledge can also have practical applications, for instance development of vaccines or drugs. Here we review the methods for detecting genes under positive selection using divergence data (i.e., the ratio of nonsynonymous to synonymous substitution rates, d_N/d_S). We also review methods for detecting selection using polymorphisms, such as methods based on F_{ST} measures, frequency spectrum, linkage disequilibrium and haplotype structure. In the second part, we review examples where targets of selection have been identified in pathogens using these tests. Genes under positive selection in pathogens have mostly been sought among viruses, bacteria and protists, because of their paramount importance for human health. Another focus is on fungal pathogens owing to their agronomic importance. We finally discuss promising directions in pathogen studies, such as detecting selection in non-coding regions.

1. Introduction

Hosts and pathogens are engaged in a never-ending struggle, hosts evolving to escape pathogen infection and pathogens evolving to escape host defences, as illustrated by the Red Queen tale. Debate exists, however, about whether this coevolution process unleashes an ‘arms race’, i.e., the occurrence of recurrent selective sweeps that each favours novel resistance and virulence alleles or leads to ‘trench warfare’, i.e., balancing selection maintaining stable and long standing polymorphism at loci involved in host–pathogen recognition (Bergelson et al., 2001; Chisholm et al., 2006; Tellier and Brown, 2007). Under the ‘trench warfare’ model, the same alleles are maintained over long time scales while in the ‘arms race’ model novel alleles are recurrently driven to fixation. A high diversity of phenotypes has in fact long been observed regarding host resistance or pathogen infection ability depending on host/pathogen genotypes (Thrall et al., 2001; Salvaudon et al., 2005, 2008). The spatial distribution of the host and pathogen phenotypes (resistance/infection ability) have provided valuable information on local adaptation or maladaptation (Kaltz et al., 1999; Burdon and Thrall, 2000; Laine, 2006; Sicard et al., 2007), and thereby on coevolutionary processes.

More recently, technological and methodological progress has allowed the detection of footprints of selection directly on genes. For instance, genes assumed to be important in host–pathogen coevolution can be amplified, sequenced and analysed by powerful methods to check whether they are indeed targets of selection, what evolutionary forces are at play (e.g., balancing selection or selective sweeps), what the strength of selection is, and whether selection has been acting in the more or less recent past (Nielsen, 2005; Tenaillon and Tiffin, 2007). Such information can provide useful insights into the processes of coevolution, for instance for determining which process, between trench warfare or arms race, is the most frequent (Holub, 2001; Chisholm et al., 2006; Tiffin and Moeller, 2006), or how often selective sweeps occur. This

[*] Corresponding author at: Ecologie, Systématique et Evolution, Bâtiment 360, Université Paris-Sud, F-91405 Orsay cedex, France. Tel.: +33 1 39 15 56 69.
E-mail address: Tatiana.Giraud@u-psud.fr (T. Giraud).

1567-1348/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2009.03.010

knowledge can also have practical applications, for instance in the development of vaccines or drugs that target conserved regions in pathogens, so that they can recognize target sites in spite of the fast evolution that characterizes pathogens (Polley et al., 2003).

Powerful methods have been developed for detecting targets of selection in the genome without any *a priori* knowledge (Meyers et al., 2003; Nielsen, 2005; Nielsen et al., 2005a; Tenaillon and Tiffin, 2008). Such a bottom-up approach detects the most rapidly evolving genes (e.g., those subject to positive selection) in host and pathogen genomes. These potentially correspond to the most important functions involved in the host–pathogen struggle or in pathogen specialization on new hosts. In pathogens, such genes may be involved in evading host defences or generating novel mechanisms of infection (e.g., the production of novel toxins) (Piertney and Oliver, 2006; Tiffin and Moeller, 2006). Other kinds of rapidly evolving genes not involved in host pathogen coevolution *sensu stricto* include those playing a role in reproductive isolation (Wu and Ting, 2004; Noor and Feder, 2006) or self-recognition (Wu et al., 1998); these can also be important for the understanding of host–pathogen interactions.

Here, we first briefly review methods for detecting genes under positive selection, using either divergence (i.e., variation representing substitutions between species) or polymorphism data. We then illustrate applications of such methods, using relevant studies where selected genes have been identified in pathogens, providing insight into host–pathogen coevolution. The most studied pathogens in this context are viruses, bacteria and protists (e.g., *Plasmodium*), because they are of paramount importance for human health and because they have small and fully sequenced genomes. Next we focus on fungal pathogens that establish peculiar molecular interactions with their hosts (Flor, 1942) and have dramatic agronomic and ecological consequences (Desprez-Loustau et al., 2007). Many whole genomes are also now available for extensive evolutionary analyses since fungi possess small genomes, for eukaryotes (e.g., Galagan et al., 2005; Robbertse et al., 2006; Cornell et al., 2007; Aguileta et al., 2008, 2009a). Finally we discuss future directions that may be interesting to follow in order to detect positive selection in pathogens, such as in regions controlling gene expression.

2. Methods for detecting positive selection

2.1. Methods for detecting positive selection using divergence data: looking for traces of past positive selection

2.1.1. The d_N/d_S ratio

Measuring the d_N/d_S ratio (number of nonsynonymous substitutions over number of synonymous substitutions) provides a direct way to measure the selective pressure acting on codons. Evaluating whether d_N/d_S is significantly higher than 1 constitutes a test for the action of positive selection. The early approaches were the so-called counting methods, based on pairwise sequence comparisons (Miyata and Yasunaga, 1980; Li et al., 1985; Nei and Gojobori, 1986). All such methods rely on the same procedure: (1) count the number of synonymous and nonsynonymous changes taking into account all possible evolutionary pathways between homologous (aligned) codons; (2) count the number of synonymous and nonsynonymous sites; (3) correct for multiple substitutions at the same site by using an explicit evolutionary model (e.g., Jukes Cantor). Counting methods average d_N and d_S over all the sites in the sequence and over the whole time period separating the sequences. Averaging however drastically reduces the power of these approaches because the selection signal is diluted. Positive selection does not occur at a constant rate through time but often happens episodically (Gillespie, 1991). Selection events can however be recurrent, as it is often the case in host–pathogen interactions, and can lead to diversifying selection. In addition, positive selection does not act homogeneously along genes, but rather acts more strongly at some sites (i.e., usually functionally relevant sites). More powerful approaches estimate the d_N/d_S ratio using maximum likelihood and Bayesian methods within a probabilistic framework (Box 1). These methods allow estimation of different ratios at different sites in the molecule and/or in different lineages in the phylogeny (e.g., Yang, 1998; Yang et al., 2000; Yang and Nielsen, 2002; Bielawski and Yang, 2004).

Box 1. The d_N/d_S ratio estimated by maximum likelihood

An accurate way to estimate the d_N/d_S ratio is by using maximum likelihood (Yang and Nielsen, 2000). It has the advantages that hypotheses can be easily tested using the likelihood ratio test (LRT) and that it can be directly applied to multiple sequence alignments. A continuous-time Markov process of substitution can be implemented along the phylogeny relating the analyzed sequences (Goldman and Yang, 1994; Muse and Gaut, 1994). At the DNA level, each column in an alignment can be thought of as arising from a continuous-time Markov process running over the tree with a state space given by the four different nucleotides (i.e., a 4x4 matrix). To incorporate the codon level of complexity, the state space can be formulated into triplets with a coding meaning (i.e., a 61x61 matrix), with stop codons not allowed. A Markov process is implemented to obtain the parameters modeling the substitution process of each codon transition, such as the codon frequencies, topology, branch lengths, the rate of substitution and selective pressure, based on the formalization of a probabilistic model. There are subtle differences in the parameterization of the substitution models proposed by Goldman and Yang (1994) and Muse and Gaut (1994). In the latter the rates of substitution are proportional to the *target nucleotide* that is part of the codon, whereas in the former the rates are proportional to the *target codon* directly. Rodrigue et al. (2008) argue that in order to construct models that better reflect the causative factors of the substitution process in protein-coding sequences, modeling should incorporate parameters dealing with amino-acid or codon propensities, such as those recently implemented (Thorne, 2007; Thorne et al., 2007; Yang and Nielsen, 2008). A comparison of models using Bayes factors showed that a model explicitly accounting for uneven codon preferences, and simultaneously incorporating a global background of nucleotide propensities (i.e., as in the Muse and Gaut approach) and heterogeneous nonsynonymous substitution rates, outperform simpler models and match the best performing ones (Rodrigue et al., 2008).

2.1.2. Variable d_N/d_S ratios among different sites in a sequence

The first codon models that went beyond pairwise comparisons of the d_N/d_S ratio (also called ω) and were applicable to multiple sequence alignments were the maximum likelihood methods developed by Goldman and Yang (1994) and Muse and Gaut (1994). These models allow the estimation of important parameters that take into account the complexity of the codon substitution process. They also have the advantage of averaging over all the possible codon states at internal nodes, instead of requiring the inference of ancestral states. Different models each assume a specific distribution of (, described by a statistical distribution (e.g., beta, normal, gamma, general discrete distribution). The parameters are estimated in the chosen models by maximizing the likelihood function. LRTs (Likelihood Ratio Tests) are then used to compare two competing, nested models. One is the null model that assumes no selection ((≤ 1) and the other is the alternative model that allows for positive selection (where (can be >1). LRTs have been shown to be accurate in simulations (Anisimova et al., 2001). Their power however depends on the sequence divergence of the data and the sample size. They can be very conservative if there is not enough divergence in the sequence alignment. In the widely used PAML package (Yang, 1997; Yang, 2007), extensive evaluations by simulation and real data have resulted in a subset of recommended models: M0, M1a, M2a, M3, M7 and M8. These models form pairs (e.g., M1a vs M2a and M7 vs M8) that use particular statistical distributions of ω and can be compared by means of a LRT.

LRTs evaluate if a model assuming positive selection fits the data better than a null model without positive selection, but do not detect particular sites subject to an increased nonsynonymous rate. The Bayesian method can be employed to infer which sites in the alignment are under positive selection. This method is used to compute the posterior probability that each site belongs to a particular ω class. A site with a posterior probability ($\geq$0.95) of belonging to the class of sites with $\omega > 1$ has a high probability of being under positive selection. Note that for datasets with very similar and slowly evolving sequences, the Bayesian analysis may lack the power to detect which sites are under selection (Anisimova et al., 2002). As in any Bayesian analysis, the choice of parameter values describing the prior probabilities that are updated in order to derive the posterior probabilities has an influence on the final results. Nevertheless, if the signal is strong enough it will be able to "overwhelm" the prior, but this is rare in typical datasets (Yang et al., 2005). The original method described in Yang et al. (2000) implemented a naive empirical approach (NEB) in which maximum likelihood estimates were used as priors, but in this way the uncertainty associated with the estimation was not taken into account. Several Bayesian techniques were proposed to improve accuracy, including the Bayes Empirical Bayes (BEB) approach implemented in PAML (Huelsenbeck and Dyer, 2004; Yang et al., 2005; Aris-Brosou, 2006; Scheffler et al., 2006). In most practical situations however BEB appears to be most accurate and has the additional advantage of being fast (Yang et al., 2005).

2.1.3. Variable d_N/d_S ratios among different lineages in a phylogeny

A sensible hypothesis that can be tested is whether selection has affected particular lineages in a phylogeny. Messier and Stewart (1997) were the first to formally test such an hypothesis. They used parsimony and maximum likelihood methods to reconstruct the ancestral lysozyme gene sequences of a group of primates including colobine monkeys, who possess the capacity for foregut fermentation. The authors compared ancestral and extant sequences and estimated the pairwise d_N/d_S differences. They found signals of positive selection occurring in the lineage leading to the common ancestors of colobine monkeys on the one hand, and of modern hominoid lysozymes on the other hand. Other studies also used ancestral state reconstruction to test for positive selection (Crandall and Hillis, 1997; Zhang et al., 1997; Williams et al., 2006; Pollock and Chang, 2007). However, it is not appropriate to use inferred ancestral sequences as observed data, because biases may be introduced by not taking the uncertainty of the inference into account.

The ω ratio can be estimated by using the maximum likelihood framework, which accounts for all possible ancestral sequences. Codon models developed by Yang (1998) assume an explicit model of codon evolution that accommodates important parameters, such as the transition/transversion rate bias or the unequal codon usage (Box 1). Furthermore, they allow the estimation of independent ω ratios in different branches. The latter codon models are called lineage-specific models and they vary in the number of parameters estimated. The simplest lineage-specific model assumes a single ω parameter across the phylogeny (i.e., one-ratio model). The most flexible lineage-specific model is the free-ratios model that assumes a different ω value for each branch in the tree. Intermediate models can accommodate a variable number of ω ratios for a different number of lineages.

In the case of lineage-specific analyses, LRTs compare in these cases two nested models with differing in the numbers of specified branches for which the ω has to be estimated in order to test which model explains the data better. This approach is recommended if *a priori* knowledge is available about which lineages in a phylogeny are more likely to be under selection. If not, methods that do not require the pre-selection of branches should be used. Kosakovsky Pond and Frost (2005) have proposed a genetic algorithm that assigns ω ratios to branches at the same time as it progressively maximizes the fit of the model to the data. This approach, however, does not constitute a test for positive selection. It only provides estimates of selective pressure for each branch. In all cases, positive selection can only be inferred by testing the significance of $\omega > 1$ (Anisimova and Kosiol, 2009). Correction for multiple testing may be necessary to control for false positives (Anisimova and Yang, 2007).

2.1.4. Variable d_N/d_S ratios among different sites in a sequence and lineages in a phylogeny: branch-site codon models

The methods discussed so far, based on detecting an elevated d_N/d_S ratio, can be divided into two types: site and branch methods, depending on whether they measure selection pressure variability across specific sites or along particular lineages. Clearly, the most realistic expectation is that selection pressure varies both across sites and among branches in the phylogeny. Yang and Nielsen (2002) proposed a method that allows for varying selection pressure among sites, but with a subset of sites also changing along a branch or set of branches previously specified (foreground branches). A LRT is then used to compare a branch-site model with a null model where no sites are subject to positive selection. These models were later improved in order to reduce the rate of false positives (Yang et al., 2005; Zhang et al., 2005). Other modifications include the possibility of testing for clade-specific variation in selection pressure (Bielawski and Yang, 2004; Roth et al., 2007). Here again, when no *a priori* information is available about lineages where selection might have occurred, one may want to test many or all branches in the phylogeny and set them as foreground branches. In this case, a correction for multiple testing must be implemented (Anisimova and Yang, 2007). Guindon et al. (2004) extended the variation of selection pressure across sites and among branches to allow changes of selection regimes. In an interesting development, Kosiol et al. (2008) have used a Bayesian technique allowing the study of how selection patterns vary among different lineages. The latter method is actually very computationally expensive. It has consequently only been used so far in datasets with few lineages.

2.1.5. Other improvements to d_N/d_S-based approaches

A key component of d_N/d_S tests is the codon evolution model used. Several modifications have been proposed to make evolution models more realistic in their reflection of the substitution process. Many such adjustments involve partitioning sites according to specific site characteristics. Sites can be partitioned based on their variable conservation under a covarion-like process (Siltberg and Liberles, 2002), on whether they are close together in the sequence (Fares et al., 2002), or in contact in the tertiary structure (Suzuki, 2004; Berglund et al., 2005; Dutheil et al., 2005; Fares and Travers, 2006; Wagner, 2007). A very important improvement is to account for variation in the synonymous rate (d_S) (Fares et al., 2002; Kosakovsky Pond and Frost, 2005; Scheffler and Seoighe, 2005; Zhang et al., 2006; Mayrose et al., 2007). Weak selection on codon usage, causing small scale variations in d_S, does not seem to affect estimates of selection pressure (Yang and Nielsen, 2008). However, large scale variations in d_S will be of concern and need to be accounted for. For some genes (e.g., transmembrane proteins) the process of codon substitution can in fact be highly variable among sites. This can lead to false positives in tests for positive selection. A new likelihood-based method, called LiBaC, groups codon sites according to similarities in the underlying substitution process of evolution (Bao et al., 2008). Other parameters may also vary significantly among sites. It therefore appears important to test the best mixed-effect model (i.e., models partitioning sites into pre-specified classes), as it may have an impact on the likelihood estimation of parameters (Bao et al., 2007).

An important pitfall of d_N/d_S tests for identifying positive selection is recombination. Special attention needs to be paid to distinguishing recombination from positive selection events, as the former can be mistaken for the latter. The problem arises because most methods (e.g., PAML) assume all the sequences in the alignment are related by a single phylogeny, an assumption that is violated when there is recombination (i.e., recombining sites do not conform to the same topology as non-recombining sites) (Anisimova et al., 2003). In addition, recombination can affect the estimation of d_N/d_S (Shriner et al., 2003; Spencer and Coop, 2004). This problem is particularly important when analyzing species prone to high recombination rates, such as viruses (Scheffler et al., 2006). Wilson and McVean (2006) proposed a likelihood approximation to the coalescent process, simultaneously estimating recombination and selection parameters by reversible-jump MCMC.

2.1.6. Beyond the d_N/d_S approach: amino acid models

Models at the protein level can be more appropriate for rapidly evolving proteins or distantly related species (i.e., deep divergences) because codon-level models are susceptible to bias due to synonymous rate saturation (Anisimova and Liberles, 2007). Another advantage of the protein-level is that one can distinguish between conservative substitution and a substitution leading to a chemically different amino-acid. There is a wealth of methods that incorporate parameters related to physicochemical amino acid characteristics aiming at more realistic models of amino acid replacement (e.g., Miyamoto and Fitch, 1995; Gu, 1999, 2001, 2006; Penny et al., 2001; Pupko and Galtier, 2002; Siltberg and Liberles, 2002; Abhiman and Sonnhammer, 2005a,b; Sainudiin et al., 2005; Wong et al., 2006; Wang et al., 2007). Other methods estimate the ratio of radical and conservative substitutions (Hughes et al., 1990). Such an approach is however controversial (Smith, 2003). Another approach involves detecting shifts in evolutionary rate and mapping them in the corresponding 3D protein structure (Shapiro and Alm, 2008). In an interesting study, non-homogeneous and non-stationary models have been incorporated under a Bayesian framework for amino acid replacements (Blanquart and Lartillot, 2006, 2008).

2.2. Methods for detecting positive selection using polymorphism data: looking for traces of recent and ongoing positive selection

The methods presented above search for footprints of selection by examining changes between species or between alleles at the protein-coding level and comparing synonymous vs nonsynonymous substitutions. Signals of selection can also be sought by comparing allelic frequencies within species (Box 2). Though selection modifies allelic frequencies, the evolution of genomes is not a purely deterministic process. It is affected not only by different selective forces, but also by the stochastic nature of genetic sampling, which itself is impacted by the demography of populations (Nielsen, 2005). Any statistical method aiming to detect selection at the population level faces the challenge of teasing apart the effects of demography and selection, which can leave similar traces in the genome (Nielsen et al., 2007). Another important goal is to quantify the relative contribution of positive Darwinian selection and random genetic drift in determining levels of variability within species and divergence between species (Eyre-Walker, 2006).

Box 2. Different applications in pathogens of methods for detecting positive selection using polymorphism and divergence data

The main difference between the methods using polymorphism and divergence data is the time-scale. The two approaches are complementary: divergence level methods detect patterns of substitutions between species and are best suited for detecting selection that operates over relatively long time periods, while methods developed to analyze polymorphism data rely on variation patterns within the same species and are thus best suited for detecting recent selection. Divergence methods are generally quite robust to demographic assumptions and provide excellent tools for investigating fixed adaptive changes, as well as coevolutionary and compensatory evolution such as is expected to take place in long-term host–pathogen arms races. Population genetics methods are indicated for elucidating the dynamics of emergent pathogen-caused diseases, often associated with recent colonization and migration among populations.

The neutral theory (King and Jukes, 1969; Kimura, 1983) states that the observed genetic variation within species, as well as divergence between species, is caused by the random fixation of neutral mutations. The advantage of this formulation is that it provides an easy-to-test null hypothesis. Actually, most selection tests are based on rejecting neutrality and allow detection of advantageous mutations that are in the process of becoming, or have recently become, fixed in the population. These events can leave a characteristic signature in the genome that is called a selective sweep, which occurs when a selected mutation reduces the variability of linked neutral sites (Maynard Smith and Haigh, 1974; Kim and Stephan, 2002). Signatures left by selection include genetic differentiation among populations, an overrepresentation of low- and high-frequency alleles vs those at intermediate frequency, increased linkage disequilibrium (LD) between the locus under selection and neutral loci, and a strong decrease of variability around an advantageous mutation (haplotypes in the vicinity of a sweep tend to become homozygous). Below we discuss methods that identify different types of positive selection signatures in the genome using polymorphism.

2.2.1. Population differentiation (F_{ST} measures): the degree of differentiation

Recent positive selection, such as that experienced during an arms race, can generate genetic differentiation between populations at loci under selection in a heterogeneous environment. This is more likely to happen in subdivided populations with restricted gene flow (Vitalis et al., 2001; Kayser et al., 2003). Therefore, extreme differentiation between populations at a given locus relative to other loci may be evidence for positive selection. The first proposed neutrality test rejects a neutral model for a locus if it shows larger than expected differentiation among populations (Lewontin and Krakauer, 1973). F_{ST} is the most frequently used measure of population differentiation and different methods have been proposed to test whether this statistics is significantly different from zero or from neutral expectations (Akey et al., 2002).

More recently, complex statistical models have been proposed that incorporate demographic scenarios, allowing different loci to have diverse demographic parameters or selective histories. Bayesian framework was developed for F_{ST} models (Beaumont and Nichols, 1996; Beaumont and Balding, 2004). Using such models, statistics derived from data are compared with distributions of summary statistics obtained by coalescence simulations

(Tenaillon et al., 2004). Alternatively, competing models with different demographic and selection parameters can be tested (Beaumont et al., 2002; Fagundes et al., 2007).

2.2.2. Site Frequency Spectrum: an increase in low-frequency alleles

A selective sweep will cause a skewed allele frequency distribution towards more alleles that are each present in low frequencies (Braverman et al., 1995). Tests based on signatures in the frequency spectrum compare the allele frequencies of individual segregating loci (i.e., frequencies of point mutations at different sites) with the expected neutral-site frequencies. These methods are among the most widely used for the analysis of multi-locus data (e.g., SNPs). In these tests, summary statistics of candidate loci are compared with neutral expectations to detect outliers that may indicate positive selection. Site-frequency spectra (SFS) can be analyzed using the traditional summary statistics: Tajima's D (Tajima, 1989), Fu and Li's D (Fu and Li, 1993) and Fay and Wu's H test (Fay and Wu, 2000). More powerful methods have also been developed to analyze the whole SFS without the use of summary statistics (Kim and Stephan, 2002).

A selective sweep affects the SFS, resulting in a deficiency of alleles segregating at intermediate frequencies, immediately after the selective sweep, and an increase in those alleles during the sweep. Tajima's D test compares two nucleotide diversity estimates, the theta estimates, based on the average pairwise differences in nucleotide composition, and the total number of segregating sites (Tajima, 1989). This difference is compared with the expectations under the neutral model in order to reject or accept the alternative hypothesis. The D statistics detects an excess or deficiency of mutations at intermediate frequencies compared to new mutations segregating at low or high frequencies (Tajima, 1989). Fu and Li's statistics establishes the polarity of the mutations by introducing an outgroup (Fu and Li, 1993). Fay and Wu's H test (Fay and Wu, 2000) gives a higher weight to high-frequency derived mutations. An interesting idea would be to perform these tests on nonsynonymous and synonymous sites separately (Anisimova, personal communication).

Selective sweeps have a complex spatial pattern that can be analyzed in order to predict the location of the selected locus and the strength of selection. Kim and Stephan (2002) developed a method to detect selective sweeps and estimate their parameters (location, intensity). Its principle is to perform a likelihood-ratio test of selective sweep vs neutrality, based on the complete SFS and on the spatial distribution of diversity along a chromosome. Later developments refined this original approach to account for the confounding effects of demography (Jensen et al., 2005; Nielsen et al., 2005b).

2.2.3. Linkage Disequilibrium (LD) and haplotype structure: extended haplotype homozygosity and increased LD

Alleles at different loci will be more strongly associated in genomic regions under selection (Andolfatto et al., 1999). Following a selective sweep, the levels of linkage disequilibrium between selected and neutral loci (LD) will temporarily increase. The LD of two markers on the same side of the sweep will be high (Depaulis and Veuille, 1998; Kim and Nielsen, 2004), whereas the LD between two markers flanking the selective sweep will decrease (Kim and Nielsen, 2004; Stephan et al., 2006). Another consequence of a selective sweep, especially clear when the sweep is incomplete (when the polymorphism has not been fixed in the population yet), is that haplotypes carrying the beneficial mutation can rise to high frequencies and exhibit high homozygosity over large genomic regions. This creates a characteristic haplotype structure. In the case of balancing selection, LD will be reduced if the polymorphism is old, but may transiently increase at an earlier stage (Przeworski, 2002). These different signatures of selection can be identified, based on LD, by a range of different methods (Hudson et al., 1994; Kelly, 1997; Depaulis and Veuille, 1998; Andolfatto et al., 1999; Kim and Nielsen, 2004). The extended haplotype test (Sabeti et al., 2002) compares the increase in distinct haplotypes as a function of the distance from the selected locus. A later refinement of the method includes the comparison between haplotypes carrying the ancestral and derived SNP characters (Voight et al., 2006). Interestingly, selective sweeps can interfere with each other leading to less reduction in heterozygosity relative to the effect of a single sweep (Chevin and Billiard, 2008). Finally, Wang et al. (2006) proposed the linkage-disequilibrium decay test that uses exclusively homozygous SNP sites.

2.2.4. McDonald–Kreitman (MK) test: contrasting silent and replacement substitutions

The McDonald–Kreitman test (McDonald and Kreitman, 1991a,b) and its derivatives (Fay et al., 2001; Bustamante et al., 2002; Smith and Eyre-Walker, 2002; Sawyer et al., 2003; Bierne and Eyre-Walker, 2004) compare the number of synonymous and nonsynonymous substitutions, both within and between species. Positive selection increases the relative rate of nonsynonymous mutations compared to that of synonymous mutations. However, this effect is stronger at the inter-species than at the intra-species level. Adaptive mutations tend to contribute more to divergence than to polymorphism, when compared to neutral mutations (Eyre-Walker, 2006). Tests of selection generally assume that synonymous substitutions are neutral, and that nonsynonymous mutations are strongly deleterious, neutral, or strongly advantageous. In the absence of selection the level of divergence between populations and the level of variability within populations would be equal ($Dn/Ds = Pn/Ps$, according to the notation proposed by Eyre-Walker (2006). On the contrary, under the assumptions that most synonymous substitutions are neutral and that nonsynonymous mutations are strongly deleterious or strongly advantageous, the level of variability within populations under selection is expected to be larger than the divergence between species ($Dn/Ds < Pn/Ps$). Later modifications of the classical McDonald–Kreitman test take into account weak selection by introducing a parametric test that depends on the selection coefficient (Sawyer and Hartl, 1992), and estimate the number of adaptive substitutions under a Bayesian framework (Bustamante et al., 2001; Piganeau and Eyre-Walker, 2003; Sawyer et al., 2003). The MK-test and its derivatives are robust to demographic assumptions. However, the presence of slightly deleterious mutations (i.e., background selection) can have the effect of reducing their power to detect selection (Fay et al., 2001; Charlesworth and Eyre-Walker, 2008). It is possible that current estimates of positive selection based on MK and MK-based tests are actually underestimates and that positive selection is more pervasive than they would suggest (Charlesworth and Eyre-Walker, 2008).

2.2.5. Improvements and challenges

The different methods we have discussed here to identify positive selection make different assumptions, detect specific signatures, and have more or less power depending on the strength of selection, the recombination and mutation rates, the effective population size, gene density, whether selection acts on new or standing variation, and other factors (Teshima et al., 2006). It is desirable to separate the locus-specific effects (e.g., recombination, selection, evolutionary rate) from the demographic effects that act similarly throughout the genome (e.g., bottlenecks, population size expansion, inbreeding, founder effects). Given a large sample of loci from across the genome, genome-wide expectations of summary statistics may be compared with observations. Detected outlier genes are likely to be subject to selection. The challenges

Please cite this article in press as: Aguileta, G., et al., Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. Infect. Genet. Evol. (2009), doi:10.1016/j.meegid.2009.03.010

involved in estimating genome-wide effects and detecting outlier loci are many and not always trivial.

Genomic scans and the subsequent detection of outlier loci rely on sampling large numbers of markers from across the genome. Most methods assume that the loci analyzed are independent. However, given the very nature of the signature of selection and its relationship to linkage, this assumption is frequently not met (Stinchcombe and Hoekstra, 2008). A related statistical problem is caused by failing to account for the non-independence of observations. In this case, the tests are too liberal and may lead to a high number of false positives. Accounting for multiple testing may help reduce the number of false positives, a recognised problem in genomic scans (Biswas and Akey, 2006). Non-independent sampling is a frequent problem of sliding-window approaches in genomic scans although there are ways to correct for it (Ardell, 2004). An alternative may be to use methods that are especially powerful for the investigation of spatial patterns along sequences (i.e., selection signals). One such approach is to use methods based on hidden Markov models (HMMs) (Siepel and Haussler, 2004).

Teshima et al. (2006) and Stinchcombe and Hoekstra (2008) have indicated some weaknesses associated with genomic scans that detect outlier genes: (i) are the detected genes under selection, or are they linked to the target of selection instead? (ii) is there indeed a correspondence between the selected locus and a relevant phenotypic trait? Statistical concerns are also crucial (McVean and Spencer, 2006): (i) can all the selected genomic elements be detected, even if they do not leave the signals that current methods look for (i.e., are there false negatives)? (ii) how to minimize the number of false positives? (iii) what statistics best summarize the data? (iv) what models need to be used in model-based approaches? (v) how to determine the significance of the test? (vi) how to correct for ascertainment bias? In the case of tests based on allelic frequencies, wrongly assuming that the ancestral state of each single nucleotide polymorphism has a matching ortholog in the aligned species (i.e., ancestral misidentification under the parsimony criterion) may give false positives (Hernandez et al., 2007).

Finally, model-based approaches rely on the correct estimation of the parameters involved in population dynamics (i.e., Ne, rho, generation times, etc.) in order to differentiate the effects caused by selection and those that result from demography. Simplistic demographic models may lead to spurious conclusions relative to mutation rates and selection patterns (Patterson et al., 2006; but see Burgess and Yang, 2008; Wakeley, 2008). Important developments in this field promise to provide powerful new approaches based on coalescence theory, approximate Bayesian and composite likelihood methods for estimating demographic parameters and constructing more realistic models (Beaumont et al., 2002; Rannala and Yang, 2003; Hey and Nielsen, 2004; Burgess and Yang, 2008; Jensen et al., 2008). However, a better understanding of the ecology of populations is as important as model improvement.

Recently, new sequencing techniques have generated large quantities of multi-locus data sampled from across genomes. In parallel, there has been a spectacular development of powerful analytical tools to detect selection, which has enabled genome-wide searches of regions that are subject to different selective forces. Genomic scans do not require previous knowledge of genes or genomic regions and may result in the discovery of unanticipated interesting regions. Often the regions found to be under selection are not annotated and little insight can be gained without a careful functional characterization. In the coming years we anticipate more effort in the experimental validation of regions predicted to be under selection, using both polymorphism and divergence data (Yang et al., 2005; MacCallum and Hill, 2006).

Box 3. Software available for the detection of positive selection

For polymorphism data:
For an excellent review see Excoffier and Heckel, 2006.
Neutrality tests:
Arlequin (http://lgb.unige.ch/arlequin/)
FDIST2 (http://www.rubic.rdg.ac.uk/∼mab/software.html)
DnaSP (http://www.ub.edu/dnasp/)
MEGA (http://www.megasoftware.net/)

Linkage disequilibrium:
Arlequin (http://lgb.unige.ch/arlequin/)
FSTAT (http://www2.unil.ch/popgen/softwares/fstat.htm)
Genepop (http://genepop.curtin.edu.au/)
Genetix (http://www.genetix.univ-montp2.fr/genetix/genetix.htm)
Structure (http://pritch.bsd.uchicago.edu/structure.html)

Demographic history (used when explicit demographic parameters are required for testing selection models using polymorphism data):
LAMARC (http://evolution.genetics.washington.edu/lamarc/index.html)
Migrate (http://popgen.scs.fsu.edu/Migrate-n.html)
IM (http://lifesci.rutgers.edu/∼heylab/HeylabSoftware.htm#IM)
GENETIX (http://www.genetix.univ-montp2.fr/genetix/genetix.htm)
BATWING (http://www.mas.ncl.ac.uk/∼nijw/)
Genepop (http://genepop.curtin.edu.au/)
SPAGeDi (http://www.ulb.ac.be/sciences/ecoevol/spagedi.html)
BAPS (http://web.abo.fi/fak/mnf//mate/jc/software/baps.html)
MCMCcoal (http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html)
Coalesce (http://evolution.genetics.washington.edu/lamarc/coalesce.html)
Fluctuate (http://evolution.genetics.washington.edu/lamarc/fluctuate.html)

For divergence data:
For excellent reviews see Anisimova and Liberles (2007) and Anisimova and Kosiol (2009) for codon models in particular.
PAML http://abacus.gene.ucl.ac.uk/software/paml.html
HyPHy http://www.hyphy.org/
Datamonkey A server for detecting positive selection http://www.datamonkey.org/
LiBaC http://www.bielawski.info
Selecton http://selecton.bioinfo.tau.ac.il/overview.html

3. Evidence of positive selection in pathogens

Because of the coevolutionary arms race between hosts and their pathogens, genes involved in their interaction are expected to evolve under positive selection. Looking for genes under positive selection in pathogens has therefore triggered interest in elucidating the evolutionary dynamics of coevolution. Identifying rapidly evolving genes is also useful for developing vaccines or drugs that target conserved regions. This may guarantee that they will remain good targets after months or years. Most of the studies focusing on detecting positive selection in pathogens have used inter-species methods on candidate genes. A few have used genomic scans or methods at the within species level data (Box 3).

3.1. Evidence of positive selection in viruses, protists and bacteria

The most studied pathogens regarding positive selection are viruses, protists (*Plasmodium*) and bacteria, because they have

Please cite this article in press as: Aguileta, G., et al., Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. Infect. Genet. Evol. (2009), doi:10.1016/j.meegid.2009.03.010

dramatic effects on human health and because they have small and available sequenced genomes. Most of the studies aiming at detecting genes under positive selection in pathogens targeted specific genes that were likely candidates to be under positive selection due to their functional relevance, such as those coding for antigen proteins, as they can be recognized by host immune systems. This strategy consists of sequencing the gene in many individuals and analyzing the dataset with a battery of tests. Such studies have generally focused on the pathogen proteins that are predicted to be secreted or exposed at the pathogen surface. The expectation is that interacting proteins will be the target of host immune or defense systems and that new, rare variants of these proteins will confer high fitness to the pathogens, because they will thereby escape host recognition. This will result in positive selection acting on these proteins.

Plasmodium falciparum is one of the most commonly studied pathogens. Hughes (1992) identified 3 regions out of 11 in the transmembrane protein merozoite surface antigen-1 (MSA-1) in which the d_N/d_S ratio exceeded 1 and were thus potentially evolving under positive selection. An F_{ST} comparison on a large dataset showed that one of the regions of surface antigen MSP-1 evolved under balancing selection (Conway et al., 2000), suggesting selection for rare alleles to avoid host recognition. Similarly, Tajima's D indicated that a third gene involved in erythrocyte binding was also subject to balancing selection (Verra et al., 2006). The MSP-1 protein was also studied at the genus level. Using a McDonald Kreitman test, Putaporntip et al. (2008) identified 2 regions of MSP-1 with high diversity among several *Plasmodium* species, reinforcing the idea that regions in critical positions in the protein, probably because of large exposure on the surface, are often subjected to positive selection. Strong positive selection has also been detected in the highly immunogenic protein RhopH of *P. falciparum* (Iriko et al., 2008). In the closely related pathogen *P. vivax,* two domains in an erythrocyte binding protein were identified as being under positive selection (Gosi et al., 2008). This suggests that selection on these malaria protists acted on their ability to escape host recognition. This also implies that these regions are unsuitable as medical targets unless a multivaccine strategy is adopted, however this is only feasible with a restricted number of variants (Polley et al., 2003).

Many other proteins exposed at the surface of other pathogens were found to evolve under positive selection pressure in pathogens, for instance the transmembrane protein PorB of *Meningococcus* (Urwin et al., 2002), capsid genes of the Canine parvovirus (CPV) (Shackelton et al., 2005), siderophore locus Pyoverdin involved in ferrous ion uptake of the opportunistic human pathogen *Pseudomonas aeruginosa* (Smith et al., 2005), *hrpA* pilin, and an anchoring protein of the bacterial plant pathogen *Pseudomonas syringae* (Guttman et al., 2006). At a larger scale, surface proteins in bacteria consistently departed from neutral evolution, as compared to other classes of proteins (Hughes, 2005). These findings indicate that the corresponding proteins (or protein domains) are probably involved in the evolutionary arms race between hosts and pathogens. As a consequence, they are too rapidly evolving to be the target of control policies, either through vaccines in the case of human or livestock pathogens, or through creating resistant plant varieties in the case of crop pathogens. Noticeably, positive selection has also been detected in pathogens multiplying within a single host individual to escape the immune system. The *nef* gene from HIV for instance, likely to be involved in the regulation of viral transcription, has been shown to carry signatures of positive selection in a patient within 30 months of infection (Zanotto et al., 1999).

The pattern of rapid evolution was sometimes found to be specific to a lineage. For instance, the accessory proteins and *env* gene of the immunodeficiency virus evolved under positive selection in the virus lineage that is pathogenic to humans, but not in the lineage that is pathogenic in chimps (Lemey et al., 2005; Soares et al., 2008). This observation suggests that the nature of the interaction between host and pathogen has changed in some lineages, possibly because of modifications in pathogen life history traits. As a matter of fact, HIV propagates as a virion after the death of the infected cells, while related PTLV, the primate T-cell lymphotropic virus, mainly infects new cells via cell-to-cell contacts, which prevents its exposure to the host immune system.

An interesting way of classifying mutations is to distinguish not only synonymous vs non-synonymous ones but also conservative vs radical replacements among the non-synonymous changes (Hughes et al., 1990). These approaches were developed in particular for microbes against which vaccine design is desirable (e.g., *Plasmodium*). For instance, Jongwutiwes et al. (2008) have characterized the sequence of a protein against which immunity appears. They showed that non-synonymous mutations were frequent, but that they involved conservative replacements. Thus, a generic vaccine based on the consensus sequence may still be efficient.

Not all the proteins exposed on microbe surfaces reveal traces of positive selection. For instance, the envelope genes of dengue virus type I and HIV-1 evolve under strong purifying selection and occasional recombination (Goncalvez et al., 2002; Edwards et al., 2006) and not under positive selection as initially thought (Nielsen and Yang, 1998). These findings suggest that either the corresponding proteins are not immunogenic or that the immunity developed through those unchanging proteins does not impact virulence. Vaccines against these molecules may thus be efficient if they help impeding pathogen proliferation upon first exposure.

Positive selection has been identified in pathogens not only in relationship to the continual exposure to the same host species but also associated to host shifts. Greater mobility linked to human activity allows pathogens to encounter potential new hosts and can thus foster the emergence of new epidemics when pathogens manage to adapt to these novel hosts. Interestingly, by performing an extensive phylogenetic and molecular evolution analysis on complete sequences of all Canine distemper virus (CDV), both on dogs and non-dogs, McCarthy et al. (2007) showed that two mutations in the haemagglutinin molecule were recurrently present in viruses infecting non-dogs. This suggested that the new variants were selected during the host shift. A relationship was similarly detected between a host range expansion of the Canine Parvovirus (CPV) and a signal of positive selection at several codons of the capsid genes according to site-to-site detection analyses (Shackelton et al., 2005).

Duplicated genes are often impacted by positive selection (Ohno, 1970; Taylor and Raes, 2004; Crow and Wagner, 2006). Gene duplication arises at very high rates, possibly as frequently as individual nucleotide mutations (Lynch and Conery, 2000). Duplicates experience a brief period of relaxed purifying selection facilitating their evolutionary divergence after duplication. Their redundancy permits the accumulation of even highly deleterious mutations, but that can ultimately give rise to new functions (Gu, 2003; Crow and Wagner, 2006). Thus, a majority of duplicates are silenced and become pseudogenes (Lynch and Conery, 2000), following the so-called birth and death evolution (Nei, 2007). Duplicated genes may be retained if they provide an advantage, due to an increased gene dosage effect or increased functional diversification (Cornell et al., 2007). Duplicated copies can evolve a new function by a process known as neofunctionalization. An alternative fate is subfunctionalization in which the function of the ancestral gene is partitioned between the two daughter copies (Prince and Pickett, 2002). In both cases, positive selection is often responsible for functional divergence (Ohno, 1970; Tanaka and Nei, 1989; Zhang et al., 2002). The pathogenicity of *Helicobacter* *pilori* has for instance been related to an increase in the number of secreted proteins from the 'Sel1-like repeat' (SLR) gene family, many copies of which have undergone positive selection (Ogura et al., 2007). By comparing the complete genome sequences of the pathogen protists *Leishmania major*, *L. infantum* and L. *braziliensis*, Devault and Banuls (2008) have shown that some genes within the PSA family (promastigote surface antigen) evolved under positive selection. PSA is one of the major classes of membrane proteins present at the surface of the protists *Leishmania* and their leucine rich repeats are suggestive of their involvement in pathogen-to-host physical interactions.

Medical practices have become a new source of evolutionary selection pressure on pathogens. By analyzing 13 strains of *P. falciparum* collected in 6 different South East Asian countries having different antimalarial drug treatment policies, Anderson et al. (2005) showed that genes involved in drug resistance carried high proportions of non-synonymous mutations and showed a significant geographic structure in contrast to housekeeping genes. A positively selected site of the Classical Swine Fever virus (CSF) has also been linked to the failure of the most commonly used vaccine to confer immunity (Tang and Zhang, 2007).

In studies of positive selection it is important to exclude artifacts due to bottlenecks or recombination (Jensen et al., 2007). The methods described above for divergence data are designed to detect codons with an abnormally high proportion of non-synonymous mutations. However, there is always some risk that the codons identified using bioinformatic analyses are mere artifacts. A small number of functional studies have been dedicated to validate the tests for positive selection. For instance, Anisimova and Yang (2004) showed that all positively selected sites in the hepatitis D antigen lie in antigenic domains, and not in regions that are not targeted by selection. In a virus contributing to the virulence of its parasitoid host, Dupas et al. (2008) established a relationship between specific viral alleles with a high d_N/d_S ratio and the presence of the parasitoid's host for which virulence was required.

More recently, studies have taken advantage of the availability of whole genomes to identify genes under positive selection without *a priori* gene candidates. This approach becomes increasingly feasible as more complete genome sequences are available. Interestingly, the proportion of genes potentially subjected to positive selection and its magnitude differed across taxa, ranging between ca. 1% for *E. coli* and the protist *Cryptosporidium* (Chen et al., 2006; Ge et al., 2008) to more than 10% for *Streptococcus* and humans (Nielsen et al., 2005a). In an example, Anisimova et al. (2007) detected 136 cases of positive Darwinian selection out of 1730 gene clusters among 12 strains from five *Streptococcus* species. The functions of the corresponding genes are diverse and many of them are classified as "hypothetical proteins". This suggests that many of the proteins involved in host–pathogen interactions and which potentially are under high selection pressure are still unknown. Furthermore, the genes specific to pathogenic species were not more often subjected to positive selection than genes shared with non-pathogenic species. This suggests that many genes involved in host parasite interactions also have a function in non-pathogenic strains (Anisimova et al., 2007). Similarly, several genes shared by pathogenic and non-pathogenic *Escherichia coli* were found to be subjected to positive selection in pathogenic strains, but these genes represented only ca. 1% of the total number of sequences analyzed (Chen et al., 2006).

Another promising approach consists in hybridizing hundreds of thousands of oligomer probes, instead of sequencing whole genomes, to detect genes with high allelic variability (Kidgell et al., 2006). This allows screening large numbers of strains for a large portion of the genome. Applied to *P. falciparum*, this approach confirmed that sequences with higher variability were the drug targets and the immunogens, suggesting that the immune system and drug use are driving pathogen evolution (Kidgell et al., 2006).

3.2. Evidence of positive selection in pathogenic fungi

Fungal pathogens have also been screened for genes under positive selection, because they are responsible for devastating crop diseases, as well as some human diseases. Compared to other eukaryotes, their genomes are small and are therefore easy to sequence (Galagan et al., 2005; Aguileta et al., 2009a). Many fungal genomes are in fact available for extensive evolutionary analyses (e.g., Robbertse et al., 2006; Cornell et al., 2007; Aguileta et al., 2008). The two major groups among the true fungi are the Ascomycota, including the yeasts and filamentous fungi, with several important model species (e.g., *Saccharomyces cerevisiae*, *Neurospora crassa*) and the Basidiomycota, including the conspicuous mushrooms, the rusts and the smuts. Ascomycota and Basidiomycota have been resolved as sister taxa of the Dicaryota (James et al., 2006). Oomycota have long been included within fungi, but were recently recognized to be closely related to brown algae (James et al., 2006). They however share many morphological, physiological and pathogenic processes with true fungi, and we will hereafter use the term "fungi" *sensu lato*. Many ascomycetes, basidiomycetes and oomycetes are plant pathogens and can lead to severe economic losses due to infected crops (Berbee, 2001). Some fungi are animal pathogens (Roy et al., 2006), causing severe diseases in immuno-compromised humans, e.g., *Aspergillus fumigatus*, *Coccidioides immitis*, *Cryptococcus neoformans* and *Candida albicans* (Brakhage, 2005; Segal, 2005; Warnock, 2006).

The genes involved in host–pathogen recognition have been studied for a long time in fungi (Flor, 1942) and many appear to follow different rules than those in other pathogens like viruses and bacteria. This is the case for the most studied plant-fungal pathogen mode of interaction, the "gene-for-gene" (GFG) mechanism of resistance (Flor, 1942). Under the GFG model of interaction, the fungus produces an elicitor, encoded by an avirulence gene, which directly or indirectly triggers a defence reaction in the host, characterized by a hypersensitive (HR) response. In order to avoid recognition by the host R locus, avirulence genes are expected to evolve towards loss of function (through mutations leading to stop codons, insertions of transposable elements within the gene, gene deletion, or protein modification of genes coding for an essential function). Fungi also secrete effector proteins, i.e., proteins that directly affect host integrity, such as toxins or surface degrading enzymes. Such effectors are thought to be directly recognized by host proteins devoted to their neutralization (Stahl and Bishop, 2000). They are therefore also expected to evolve under positive selection in order to avoid recognition by the interactor host protein.

Most studies looking for positive selection in fungal pathogens aimed to find molecular footprints of the evolutionary arms race between fungal pathogens and their hosts, in particular on avirulence genes, elicitors and effectors. Necrosis inducing proteins, also called phytotoxins, are small secreted peptides that induce cellular death in their host. The scr74 gene family encodes such proteins in *Phytophthora infestans*, an oomycete causing potato late blight, a disease responsible for the Irish famine in the XIXth century. Liu et al. (2005) have characterized 2–5 copies per isolate in a collection of 12 isolates, representing 10 different alleles in total. This high level of variability was shown to be due to recombination between paralogs and to positive selection, as evidenced by a maximum likelihood (ML) codon-specific estimation of the d_N/d_S ratio. Positive selection was also detected in both the paralogous genes NEP1 and NEP2 encoding two necrosis- and ethylene-inducing proteins in several *Botrytis* species (Staats et al., 2007), including *Botrytis cinerea*, the agent of grey mold on numerous ornamental and agronomic plants. Residues under positive selection were however not the same in NEP1 and NEP2. Another necrosis- inducing protein, NIP1, was analyzed in the barley pathogen *Rhynchosporium secalis* (Schürch et al., 2004). The originality of this system is that, in contrast to the two other paralogs of this family, NIP2 and NIP3, NIP1 acts both as a toxin that induces an H+-ATPase in the host, and as an effector interacting with the product of the host Rrs1 resistance gene, following a gene-for-gene relationship. Some residues of NIP1 were shown to evolve under positive selection, especially in the region coding for the mature protein, where the polymorphism was 4 times higher than in introns and in flanking regions. Furthermore, many populations exhibited high frequencies of NIP1 deletions, whereas NIP2 and NIP3 were present in almost all isolates. These results may be interpreted as a balance between the different evolutionary mechanisms associated with the two roles of the proteins: selection in favour of a gain of virulence (i.e., the loss of the locus) on the one hand, and positive diversifying selection acting on some residues in order to escape recognition by a (still unknown) plant detoxifier while retaining its toxic function, on the other hand.

Another example of a toxic secondary metabolite produced by a fungal pathogen is the host-specific toxin SnToxA, secreted by the wheat pathogens *Pyrenophora tritici-repentis* and *Phaeosphaeria nodorum*. This gene is thought to have been acquired by *P. tritici-repentis* from *P. nodorum* through horizontal gene transfer (Friesen et al., 2006). The evolutionary pattern and geographic distribution of this gene were studied in worldwide populations (Stukenbrock and McDonald, 2007). The site-specific estimation of the d_N/d_S ratio along the protein showed that 2 codons were under strong positive selection. The Fu and Li's and Tajima's statistics showed a significant departure from the expected frequency distribution of haplotypes in populations under neutral evolution. Some of the haplotypes exhibited stop codons within the ORF, while some isolates entirely lacked the SnToxA gene.

Fungal extracellular enzymes that are secreted by the pathogen to interact with the plant wall and facilitate the penetration in the host are also a potential target of adaptive evolution. A recent study showed that this was the case for endopolygalacturonase genes in *Botrytis cinerea* (Cettul et al., 2008), a gene family represented by at least six paralogs. Five of them, BcPG1 to 5, were sequenced in 35 *B. cinerea* isolates and 3 *B. fabae* isolates. Positive selection was found for BcPG1 and BcPG2, but not for the other paralogs, using site-specific estimation of d_N/d_S ratios and a McDonald and Kreitman test. Another extracellular enzyme very common in fungi is the laccase, involved in the degradation of polyphenolic compounds of the plant wall. Nine different alleles were recovered at the laccase locus from 33 isolates of the basidiomycete tree-pathogen *Heterobasidion annosum* (Asiegbu et al., 2004). However, positive selection was inferred based on the absolute count of the number of non-synonymous mutations along the gene. Adaptive evolution thus remains to be tested in this system using more refined and powerful tests.

Two recent studies focused on the RXLR effectors of oomycetes (Win et al., 2007; Jiang et al., 2008). Members of this highly duplicated family share little sequence similarity except a characteristic RXLR motif that defines a domain playing a role in the delivery of the effector protein into the host cell. Win et al. (2007) analyzed this family using the draft genome sequences of *Phytophthora sojae*, *P. ramorum* and *Hyaloperonospora parasitica*. ML analyses of protein evolution showed evidence of diversifying selection between paralogs, with positive selection targeting the C-terminal domain of the effector. Further, the RXLR family PsPGG20 included two paralogs that diverged under positive selection and differed in function, showing that they underwent a neo-functionalization process. Jiang et al. (2008) also analysed the full sequences of *P. sojae* and *P. ramorum* and showed that the vast majority of RXLR genes belonged to a superfamily of Avirulence-homolog (Avh) proteins characterized by a very high evolutionary rate, most of the positively selected residues being located at the C-terminus end.

The effectors in fungal plant pathogens have thus been shown to be under positive selection, as expected from a coevolutionary arms race. The elicitors in contrast seem to follow different evolutionary pathways to overcome host defences. In many cases where allelic variation at fungal avirulence loci has been studied, it was found that the means to escape host recognition was not positive selection, but simply loss of function (i.e., gene disruption or gene loss). Examples include the Avr1-Co39 locus in *Magnaporthe oryzae* (Tosa et al., 2005) and the AvrLm1 locus in *Leptosphaeria maculans* (Gout et al., 2007). In the latter case, among 460 virulent isolates analyzed from worldwide populations, 90% lacked the avirulence gene. A study conducted on eight effector and avirulence genes in *Cladosporium fulvum* (four known Avr genes and four Ecp genes encoding extracellular proteins) showed that many variants were non functional due to insertions of transposable elements in genes. Non-synonymous substitutions were observed in Avr4, but no formal test of positive selection was performed. An excess of non-synonymous mutations over synonymous ones was also found between several alleles of two other avirulence genes: the AvrL567 locus of *Melampsora lini* interacting with the highly polymorphic L locus of resistance in the flax *Linum usitatissimum* (Dodds et al., 2006), and the ATR13 avirulence locus of *Hyaloperonospora parasitica* interacting with the RPP13 resistance locus of the model plant *Arabidopsis thaliana* (Allen et al., 2004). However, both studies used a counting method of synonymous (S) and non-synonymous (NS) mutations reported to the number of S and NS sites over the entire gene, and used a *t*-test to assess a significant excess of NS mutations. Hence, it remains to be determined if fungal avirulence genes are really under positive selection using probabilistic approaches.

In fungal pathogens of animals, the molecular evolutionary arms race seems to resemble more that in bacterial or protist pathogens. This was the case for two proteins in the human fungal pathogen *Coccidioides immitis* and *C. posadii*. The Proline-Rich antigen PRA, a surface-located protein involved in spherule cell-wall morphogenesis, was shown to be the target of strong positive selection in the extracellular domain of the protein (Johannesson et al., 2004). Since very low levels of polymorphism were observed intraspecifically, the adaptive evolution of this protein is probably not a response to the extensive diversity maintained at the human MHC locus, but rather an evolution due to species-specific spherule morphogenesis. In a related paper, Johannesson et al. (2005) studied the spherule outer wall glycoprotein SOWgp in 8 isolates of *C. immitis* and 16 isolates of *C. posadii*. Besides its action as an adhesin, this protein, composed of tandemly repeated motifs, modulates the host-immune response. Haplotype frequency tests for selective neutrality, as well as site-specific estimation of the d_N/d_S ratio, were used to show that repeated motifs of the protein evolved under concerted evolution. Recombination analyses showed repeats homogeneization due to unequal crossing-over. Since the SOWgp protein is released within the host during pathogen proliferation, its homogenization may enhance the fungal ability to misdirect the immune response of the host.

Adaptive evolution was also studied in fungal genes or gene regions that are not involved in virulence. For example, targets of fungicides are under strong selection in natural pathogen populations. Many studies focused on the evolution in time and space of frequencies of resistance phenotypes in natural populations, but there are few examples of analyses of signatures of positive diversifying selection at the molecular level, probably because molecular targets of fungicides remains uncharacterized in many cases. Examples include the gene CYP51 in *Mycosphaerella graminicola* (Brunner et al., 2008), conferring the resistance against fungicides in fields, and the efflux pump genes CDR1 and CDR2 in *Candida albicans* (Holmes et al., 2008), conferring resistance to azole fungicides used in medicine.

Many fungi also secrete mycotoxins, a class of secondary metabolites whose production is generally the result of a biosynthesis cascade governed by gene clusters. The role, if any, of these toxic and carcinogenic polyketids in fungal virulence is largely debated. Nevertheless, they threaten animal and human health by food contamination. For example, *Fusarium* species produce several toxins, among which is the trichothecene family. In this system, isolates carrying different alleles produce different chemotypes (i.e., different toxins). A phylogenetic study conducted on the 8 genes of the TRI cluster governing the biosynthesis of these molecules in sibling species belonging to the *Fusarium graminearum* complex showed ancient trans-species polymorphism: polymorphic alleles governing the synthesis of the different toxins have been maintained across multiple speciation events, which is a signature of balancing selection (Ward et al., 2002). Moreover, the ML approach showed that genes TRI4 and TRI10 evolved by positive selection acting at a small number of residues. Significant signatures of branch-specific selection were observed in two other genes of the cluster, TRI3 and TRI11, diversifying selection being inferred for TRI11, and relaxation of selective constraint for TRI3. Similar results were obtained in a toxin gene cluster of *Aspergillus parasiticus* (Carbone et al., 2007). The diversity and evolution of intergenic regions of the aflatoxin gene cluster was analyzed in 21 *A. parasiticus* isolates. Several recombination breakpoints and haplotype blocks were identified, indicating that recombination between haplotypes was an important mechanism governing the evolution of this locus. Departure from neutrality was tested using Tajima's and Fu and Li's tests, and several analyses ruled out population subdivision and demographic events as possible explanations of the non-neutral pattern. Intergenic regions yielded genealogies with several clades separated by long branches, typical of balancing selection. The occurrence of balancing selection was further supported by a pattern of trans-species polymorphism obtained by sequencing isolates of other species of the section Flavi.

Genes controlling self-recognition are also expected to evolve under positive selection, balancing selection and to show trans-species polymorphism (Richman, 2000). In fact, positive selection has been detected in genes responsible for vegetative incompatibility in some fungi (Wu et al., 1998). Balancing selection and trans-species polymorphism has also been found at mating-type genes (controlling for sexual compatibility) in basidiomycetes (Devier et al., 2009). In *Neurospora* taxa, nonsynonymous and synonymous substitution rates of mating-type genes revealed that they evolved rapidly. Furthermore, the evolutionary trajectories were related to the mating systems of the taxa. Likelihood methods showed that positive selection acting on specific codons drove the diversity in heterothallic taxa (i.e., self incompatible at the haploid stage), while the rapid evolution was due to a lack of selective constraint among homothallic taxa (i.e., self compatible at the haploid stage) (Wik et al., 2008).

Duplication followed by neofunctionalization and subfunctionalization is a powerful means for evolutionary innovations, and often involves positive selection. The evolution of some genes, such as those coding for hydrophobins (*hyd*), which play an important role during the infection of host plants in ectomycorrhizal fungi, follows a combination of birth–death, neofunctionalization and subfunctionalization models (Rajashekar et al., 2007). Furthermore, parallel duplication of different sets of genes might allow distantly related species to adapt to the same niche (Hughes and Friedman, 2003). Two distantly related fungi, the *Saccharomyces cerevisae* yeast and the zygomycete *Rhizopus oryzae*, both found in environments associated with rotting fruit, present parallel duplications, possibly indicating adaptation to a similar ecological niche (Cornell et al., 2007). Byrne and Wolfe (2007) investigated the evolutionary consequences of the pattern of whole genome duplication (WGD), by analyzing gene loss and asymmetry in evolutionary rates within the complete genomes of *Saccharomyces cerevisae*, *Candida glabrata* and *S. castellii*, three yeast species that underwent a WGD. By comparing them to the non duplicated outgroup *Kluyveromyces lactis*, they showed that asymmetric rate of protein evolution was widespread among ohnologs (pairs of gene produced by the WGD), and that the faster-evolving copy was the same (the ortholog) in multiple post-WGD species. In general, when a copy was lost in a species, its ortholog was usually the faster evolving gene of a pair of ohnologs in the other species. These results are consistent with widespread neo-functionalization experienced by numerous ohnologs early after WGD.

The growing amount of full genome data now covers a large part of the fungal kingdom (Galagan et al., 2005; Aguileta et al., 2009a). This led several authors to search for homologs of effectors and elicitor proteins in a large number of fungal species (Patron et al., 2007). Some fungal proteins, that were proven to be involved solely in virulence and in any other processes, were shown to have a very limited or patchy phylogenetic distribution (Patron et al., 2007). Such a lack of some genes in many species distributed across the whole fungal phylogeny may be explained by (i) their rapid evolution that precludes the detection of orthologs as the phylogenetic distance increases, (ii) losses, that may be selected for, especially for avirulence genes, and (iii) horizontal gene transfers, a mechanism that allows the acquisition of new molecular weapons (van der Does and Rep, 2007). This view was corroborated by several recent whole genome studies. For example, Cai et al. (2006) analyzed the genome of seven Ascomycota with two animals as an outgroup. They defined different levels of lineage specificity (a measure of the breadth of the phylogenetic distribution of a gene): Eukaryote-core, Ascomycota-core, Euascomycetes-specific, Hemiascomycetes-specific, *Aspergillus*-specific and *Saccharomyces*-specific. Their results showed a robust positive correlation between lineage-specificity and rates of non-synonymous mutations: the patchier the phylogenetic distribution of a gene, the faster its protein evolution. Lineage specificity better explained this accelerated evolutionary rate than other predictors such as gene expression levels, gene essentiality, gene dispensability, and the number of protein-protein interactions. Thus, positive selection and accelerated evolution may accompany the emergence of orphan genes.

A recent study aimed at elucidating the functions involved in host specialization in pathogenic fungi, by identifying sequences that have evolved under positive selection among closely related pathogens specialized on different hosts. For this goal, ESTs were sequenced from each of four *Microbotryum* species, which are fungal pathogens responsible for anther smut disease on host plants in the Caryophyllaceae. Approximately 11% of the 372 predicted orthologous genes were found to evolve under positive selection, which suggests that a high proportion of genes are involved in host specialization and speciation in pathogens. Sequencing 16 of these genes in 9 additional *Microbotryum* species confirmed that they have indeed been rapidly evolving in the pathogen species specialized on different hosts. The genes detected to be under positive selection were putatively involved in nutrient uptake from the host, secondary metabolite synthesis and secretion, respiration under stressful conditions and stress response, hyphal growth and differentiation, and regulation of expression by other genes. Many of these genes had transmembrane domains and may therefore also be involved in pathogen recognition by the host. Such approach to detect genes under positive selection without *a priori* candidates thus provides insights on the percentage of genes involved in host specificity for biotrophic pathogens and their functions, some of which were not expected based upon prior studies. Such an approach should be interesting to apply to other pathogens with contrasted ecologies and should be feasible even for non-model organisms (Aguileta et al., submitted).

4. Conclusion and future directions

The methods described above are powerful for detecting genes evolving by positive and balancing selection in pathogens. They will facilitate a better understanding of the evolutionary history of host–pathogen interactions with applications to the development of vaccines or resistant plant cultivars. However, these approaches are still restricted to some well-known human or crop pathogens. This is particularly true for studies aiming at detecting genes under positive selection without *a priori* candidate genes. We expect the use of the described methods to increase in the near future thanks to the growing availability of whole genome sequences of pathogens. Pathogens often have small genomes, which make them good candidates for fast whole genome scans for detecting genes under selection.

Another direction that would be interesting to explore is the existence and impact of mutation hotspots on positive selection in pathogens. Mutation frequencies indeed vary along a nucleotide sequence and some genome regions can present sites that evolve much faster than average (Benzer, 1961). Some mutational hotspots show an accelerated non-synonymous substitution rate, evidence that they may have a role in generating variability on which positive selection can act. In rice (*Oryza sativa*) for example, the leucine-rich repeat (LRR) receptor kinase, involved in pathogen recognition, has also hypervariable sites under positive selection. This variation is considered an important force in the evolution of the LRR domain, because it promotes instability in crosses with near relatives and generates novel resistance specificity (Sun et al., 2006). Mutation hotspots under positive selection can also be responsible for some types of human cancer. The transcription factor p53 for instance, which regulates cell cycle progression, repair and programmed cell death in mammals, evolves under strong positive selection during tumour progression (Glazko et al., 2004).

Another exciting direction is provided by the methods that have been developed very recently to detect selection in non coding regions, in particular in regulatory regions (Wong and Nielsen, 2004; Hahn, 2007; Egea et al., 2008). *Cis*-regulatory mutations are increasingly recognized as having a potentially significant effect on phenotypes (Borneman et al., 2007; Wray, 2007). These effects would include traits involved in host–pathogen interactions, and are therefore especially important to study in order to fully understand host–pathogen coevolution. By analogy to the d_N/d_S ratio, we can measure the ratio of substitutions per site in binding sites and intervening sites to detect *cis*-regulatory regions that have accumulated changes potentially having an effect on gene expression. A few studies have already provided evidence for selection in regulatory regions, almost exclusively in humans and *Drosophila*, in particular positive selection for sexual signalling and adaptation to local habitat (Hahn, 2007). Interestingly, balancing selection was also detected in regulatory regions of genes involved in host–pathogen interaction (Hahn, 2007). For instance, a single *cis*-regulatory mutation eliminates expression at a blood group locus that confers resistance to malaria and is maintained only in sub-Saharan Africa (Hamblin and Di Rienzo, 2000). Another example is given by modified binding sites which increase inducibility of expression of Interleukin-4 and lead to a faster immune response, and also appear to be under balancing selection in humans (Rockman et al., 2003). Analyses of complete fungal genomes, some belonging to pathogen species, have revealed cases of regulatory diversification (Gasch et al., 2004). Presumably, many other cases of regulatory evolution, directly linked to pathogenicity, exist in pathogens, which will be fascinating to discover.

Acknowledgements

We thank Michel Tibayrenc for his cordial invitation to write this review. We thank Maria Anisimova, Luis-Miguel Chevin, William Marande, Jacqui Shykoff and an anonymous referee for very useful comments on a previous version of this manuscript. We acknowledge the grants ANR 06-BLAN-0201 and ANR 07-BDIV-003. GA acknowledges post-doctoral grants from the CNRS and the University Paris Sud. We apologize to all those colleagues whose work we might have missed to cite in this article.

References

Abhiman, S., Sonnhammer, E.L.L., 2005a. FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res. 33, D197–D200.

Abhiman, S., Sonnhammer, E.L.L., 2005b. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Prot.-Struct. Funct. Bioinf. 60, 758–768.

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.-H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A., Giraud, T., 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst. Biol. 57, 613–627.

Aguileta, G., Hood, M., Refrégier, G., Giraud, T., 2009a. Genome evolution in pathogenic and symbiotic fungi. Adv. Bot. Res. 49, 151–193.

Aguileta, G., Lengelle, J., Marthey, S., Chiapello, H., Rodolphe, F., Gendrault-Jacquemard, A., Wincker, P., Dossat, C., Giraud, T. Genes under positive selection in pathogens: looking for the genes involved in host specialization. submitted.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., Shriver, M.D., 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12, 1805–1814.

Allen, R.L., Bittner-Eddy, P.D., Grenvitte-Briggs, L.J., Meitz, J.C., Rehmany, A.P., Rose, L.E., Beynon, J.L., 2004. Host–parasite coevolutionary conflict between *Arabidopsis* and downy mildew. Science 306, 1957–1960.

Anderson, T.J.C., Nair, S., Sudimack, D., Williams, J.T., Mayxay, M., Newton, P.N., Guthmann, J.-P., Smithuis, F.M., Hien, T.T., van den Broek, I.V.F., White, N.J., Nosten, F., 2005. Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. Mol. Biol. Evol. 22, 2362–2374.

Andolfatto, P., Wall, J.D., Kreitman, M., 1999. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. Genetics 153, 1297–1311.

Anisimova, M., Kosiol, C., 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol. Biol. Evol. 26, 255–271.

Anisimova, M., Liberles, D.A., 2007. The quest for natural selection in the age of comparative genomics. Heredity 99, 567–579.

Anisimova, M., Yang, Z.H., 2004. Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? J. Mol. Evol. 59, 815–826.

Anisimova, M., Yang, Z.H., 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol. Biol. Evol. 24, 1219–1228.

Anisimova, M., Bielawski, J.P., Yang, Z.H., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18, 1585–1592.

Anisimova, M., Bielawski, J.P., Yang, Z.H., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19, 950–958.

Anisimova, M., Nielsen, R., Yang, Z.H., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164, 1229–1236.

Anisimova, M., Bielawski, J., Dunn, K., Yang, Z., 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. BMC Evol. Biol. 7, 154.

Ardell, D.H., 2004. SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. Bioinformatics 20, 1986–1988.

Aris-Brosou, S., 2006. Identifying sites under positive selection with uncertain parameter estimates. Genome 49, 767–776.

Asiegbu, F.O., Abu, S., Stenlid, J., Johansson, M., 2004. Sequence polymorphism and molecular characterization of laccase genes of the conifer pathogen *Heterobasidion annosum*. Mycol. Res. 108, 136–148.

Bao, L., Gu, H., Dunn, K.A., Bielawski, J.P., 2007. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. BMC Evol. Biol. 7, S5.

Bao, L., Gu, H., Dunn, K.A., Bielawski, J.P., 2008. Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. Mol Biol Evol 25, 1995–2007.

Please cite this article in press as: Aguileta, G., et al., Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. Infect. Genet. Evol. (2009), doi:10.1016/j.meegid.2009.03.010

Beaumont, M.A., Balding, D.J., 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13, 969–980.

Beaumont, M.A., Nichols, R.A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B 263, 1619–1626.

Beaumont, M.A., Zhang, W.Y., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035.

Benzer, S., 1961. On the topography of the genetic fine structure. Proc. Natl. Acad. Sci. U.S.A. 47, 403–415.

Berbee, M.L., 2001. The phylogeny of plant and animal pathogens in the Ascomycota. Physiol. Mol. Plant Pathol. 59, 165–187.

Bergelson, J., Dwyer, G., Emerson, J., 2001. Models and data on plant–enemy coevolution. Annu. Rev. Genet. 35, 469–499.

Berglund, A.C., Wallner, B., Elofsson, A., Liberles, D.A., 2005. Tertiary windowing to detect positive diversifying selection. J. Mol. Evol. 60, 499–504.

Bielawski, J.P., Yang, Z.H., 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59, 121–132.

Bierne, N., Eyre-Walker, A., 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol. Biol. Evol. 21, 1350–1360.

Biswas, S., Akey, J.M., 2006. Genomic insights into positive selection. Trends Genet. 22, 437–446.

Blanquart, S., Lartillot, N., 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23, 2058–2071.

Blanquart, S., Lartillot, N., 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25, 842–858.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D.D., Yu, H.Y., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., Snyder, M., 2007. Divergence of transcription factor binding sites across related yeast species. Science 317, 815–819.

Brakhage, A.A., 2005. Systemic fungal infections caused by *Aspergillus* species: epidemiology, infection process and virulence determinants. Curr. Drug Targets 6, 875–886.

Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H., Stephan, W., 1995. The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. Genetics 140, 783–796.

Brunner, P.C., Stefanato, F.L., McDonald, B.A., 2008. Evolution of the CYP51 gene in *Mycosphaerella graminicola:* evidence for intragenic recombination and selective replacement. Mol. Plant Pathol. 9, 305–316.

Burdon, J., Thrall, P., 2000. Coevolution at multiple spatial scales: *Linum marginale-Melampsora lini*—from the individual to the species. Evol. Ecol. 14, 261–281.

Burgess, R., Yang, Z., 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25, 1979–1994.

Bustamante, C.D., Wakeley, J., Sawyer, S., Hartl, D.L., 2001. Directional selection and the site-frequency spectrum. Genetics 159, 1779–1788.

Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D., Hartl, D.L., 2002. The cost of inbreeding in *Arabidopsis*. Nature 416, 531–534.

Byrne, K.P., Wolfe, K.H., 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175, 1341–1350.

Cai, J.J., Woo, P.C.Y., Lau, S.K.P., Smith, D.K., Yuen, K.Y., 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota. J. Mol. Evol. 63, 1–11.

Carbone, I., Jakobek, J.L., Ramirez-Prado, J.H., Horn, B.W., 2007. Recombination, balancing selection and adaptive evolution in the aflatoxin gene cluster of *Aspergillus parasiticus*. Mol. Ecol. 16, 4401–4417.

Cettul, E., Rekab, D., Locci, R., Firrao, G., 2008. Evolutionary analysis of endopolygalacturonase-encoding genes of *Botrytis cinerea*. Mol. Plant Pathol. 5, 675–685.

Charlesworth, J., Eyre-Walker, A., 2008. The McDonald–Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. 25, 1007–1015.

Chen, S.L., Hung, C.S., Xu, J.A., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P., Armstrong, J.R., Fulton, R.S., Latreille, J.P., Spieth, J., Hooton, T.M., Mardis, E.R., Hultgren, S.J., Gordon, J.I., 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli:* a comparative genomics approach. Proc. Natl. Acad. Sci. U.S.A. 103, 5977–5982.

Chevin, L.M., Billiard, S., 2008. Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. Genetics 180, 301–316.

Chisholm, S., Coaker, G., Day, B., Staskawicz, B., 2006. Host–microbe interactions: shaping the evolution of the plant immune response. Cell 124, 803–814.

Conway, D.J., Cavanagh, D.R., Tanabe, K., Roper, C., Mikes, Z.S., Sakihama, N., Bojang, K.A., Oduola, A.M.J., Kremsner, P.G., Arnot, D.E., Greenwood, B.M., McBride, J.S., 2000. A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. Nat. Med. 6, 689–692.

Cornell, M.J., Alam, I., Soanes, D.M., Wong, H.M., Hedeler, C., Paton, N.W., Rattray, M., Hubbard, S.J., Talbot, N.J., Oliver, S.G., 2007. Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the Fungi. Genome Res. 17, 1809–1822.

Crandall, K.A., Hillis, D.M., 1997. Rhodopsin evolution in the dark. Nature 387, 667–668.

Crow, K.D., Wagner, G.P., 2006. What is the role of genome duplication in the evolution of complexity and diversity? Mol. Biol. Evol. 23, 887–892.

Depaulis, F., Veuille, M., 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15, 1788–1790.

Desprez-Loustau, M.R., C, Buee, M., Courtecuisse, R., Garbaye2, J., Suffert, F., Sache, I., Rizzo, D. 2007. The fungal dimension of biological invasions. Trends Ecol Evol 22, 472–480.

Devault, A., Banuls, A.-L., 2008. The promastigote surface antigen gene family of the *Leishmania* parasite: differential evolution by positive selection and recombination. BMC Evol. Biol. 8, 292.

Devier, B., Aguileta, G., Hood, M., Giraud, T., 2009. Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes. Genetics 181, 209–223.

Dodds, P., Lawrence, G., Catanzariti, A., Teh, T., Wang, C., Ayliffe, M., Kobe, B., Ellis, J., 2006. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. Proc. Natl. Acad. Sci. U.S.A. 103, 8888–8893.

Dupas, S., Wanjiru Gitau, C., Branca, A., Le Ru, P., Silvain, J., 2008. Evolution of a polydnavirus gene in relation to parasitoid–host species immune resistance. J. Hered. 99, 491–499.

Dutheil, J., Pupko, T., Jean-Marie, A., Galtier, N., 2005. A model-based approach for detecting coevolving positions in a molecule. Mol. Biol. Evol. 22, 1919–1928.

Edwards, C.T.T., Holmes, E.C., Pybus, O.G., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics 174, 1441–1453.

Egea, R., Casillas, S., Barbadilla, A., 2008. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 36, W157–162.

Excoffier, L., Heckel, G., 2006. Computer programs for population genetics data analysis: a survival guide. Nature Rev. Genet. 7, 745–758.

Eyre-Walker, A., 2006. The genomic rate of adaptive evolution. Trends Ecol. Evol. 21, 569–575.

Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L., 2007. Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. U.S.A. 104, 17614–17619.

Fares, M.A., Travers, S.A.A., 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics 173, 9–23.

Fares, M.A., Elena, S.F., Ortiz, J., Moya, A., Barrio, E., 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J. Mol. Evol. 55, 509–521.

Fay, J.C., Wu, C.I., 2000. Hitchhiking under positive Darwinian selection. Genetics 155, 1405–1413.

Fay, J.C., Wyckoff, G.J., Wu, C.I., 2001. Positive and negative selection on the human genome. Genetics 158, 1227–1234.

Flor, H.H., 1942. Inheritance of pathogenicity in *Melampsoira lini*. Phytopathology 32, 653–669.

Friesen, T.L., Stukenbrock, E.H., Liu, Z.H., Meinhardt, S., Ling, H., Faris, J.D., Rasmussen, J.B., Solomon, P.S., McDonald, B.A., Oliver, R.P., 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat. Genet. 38, 953–956.

Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Galagan, J.E., Henn, M.R., Ma, L.J., Cuomo, C.A., Birren, B., 2005. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 15, 1620–1631.

Gasch, A.P., Moses, A.M., Chiang, D.Y., Fraser, H.B., Berardini, M., Eisen, M.B., 2004. Conservation and evolution of *cis*-regulatory systems in ascomycete fungi. PLoS Biol. 2, e398.

Ge, G., Cowen, L., Feng, X., Widmer, G., 2008. Protein coding gene nucleotide substitution pattern in the apicomplexan Protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. Comp. Funct. Genom. 879023.

Gillespie, J., 1991. The Causes of Molecular Evolution. Oxford University Press, Oxford.

Glazko, G.V., Koonin, E.V., Rogozin, I.B., 2004. Mutation hotspots in the p53 gene in tumors of different origin: correlation with evolutionary conservation and signs of positive selection. Biochimica et Biophysica Acta (BBA)—Gene Structure and Expression 1679, 95–106.

Goldman, N., Yang, Z.H., 1994. Codon-based model of nucleotide substitution for protein coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Goncalvez, A.P., Escalante, A.A., Pujol, F.H., Ludert, J.E., Tovar, D., Salas, R.A., Liprandi, F., 2002. Diversity and evolution of the envelope gene of dengue virus type 1. Virology 303, 110–119.

Gosi, P., Khusmith, S., Khalambaheti, T., Lanar, D., Schaecher, K., Fukuda, M., Miller, S., 2008. Polymorphism patterns in Duffy-binding protein among Thai *Plasmodium vivax* isolates. Malaria J. 7, 112.

Gout, L., Kuhn, M., Vincenot, L., Bernard-Samain, S., Cattolico, L., Barbetti, M., Moreno-Rico, O., Balesdent, M., Rouxel, T., 2007. Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environ. Microbiol. 9, 2978–2992.

Gu, X., 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. 16, 1664–1674.

Gu, X., 2001. A site-specific measure for rate difference after gene duplication or speciation. Mol. Biol. Evol. 18, 2327–2330.

Gu, X., 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. 19, 354–356.

Gu, X., 2006. A simple statistical method for estimating Type-II (Cluster-Specific) functional divergence of protein sequences. Mol. Biol. Evol. 23, 1937–1945.

Guindon, S., Rodrigo, A.G., Dyer, K.A., Huelsenbeck, J.P., 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl. Acad. Sci. U.S.A. 101, 12957–12962.

Guttman, D.S., Gropp, S.J., Morgan, R.L., Wang, P.W., 2006. Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. Mol. Biol. Evol. 23, 2342–2354.

Hahn, M.W., 2007. Detecting natural selection on *cis*-regulatory DNA. Genetica 129, 7–18.

Hamblin, M., Di Rienzo, A., 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. 66, 1669–1679.

Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24, 1792–1800.

Hey, J., Nielsen, R., 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D-persimilis. Genetics 167, 747–760.

Holmes, A., Lin, Y., Niimi, K., Lamping, E., Keniya, M., Niimi, M., Tanabe, K., Monk, B., Cannon, R., 2008. ABC transporter Cdr1p contributes more than Cdr2p to fluconazole efflux in fluconazole-resistant *Candida albicans* clinical isolates. Antimicrob. Agents Chemother 52, 3851–3862.

Holub, E.B., 2001. The arms race is ancient history in *Arabidopsis*, the wildflower. Nat. Rev. Genet. 2, 516–527.

Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowski, J., Ayala, F.J., 1994. Evidence for positive selection in the superoxide-dismutase (Sod) region of *Drosophila Melanogaster*. Genetics 136, 1329–1340.

Huelsenbeck, J.P., Dyer, K.A., 2004. Bayesian estimation of positively selected sites. J. Mol. Evol. 58, 661–672.

Hughes, A., 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. Mol. Biol. Evol. 9, 381–393.

Hughes, A.L., 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. Genetics 169, 533–538.

Hughes, A.L., Friedman, R., 2003. Parallel evolution by gene duplication in the genomes of two unicellular fungi. Genome Res. 13, 794–799.

Hughes, A.L., Ota, T., Nei, M., 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class-I Major-Histocompatibility-Complex molecules. Mol. Biol. Evol. 7, 515–524.

Iriko, H., Kaneko, O., Otsuki, H., Tsuboi, T., Su, X.Z., Tanabe, K., Torii, M., 2008. Diversity and evolution of the rhoph1/clag multigene family of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 158, 11–21.

James, T., et al., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443, 818–822.

Jensen, J.D., Kim, Y., DuMont, V.B., Aquadro, C.F., Bustamante, C.D., 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170, 1401–1410.

Jensen, J.D., Wong, A., Aquadro, C.F., 2007. Approaches for identifying targets of positive selection. Trends Genet. 23, 568–577.

Jensen, J.D., Thornton, K.R., Andolfatto, P., 2008. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. PLoS Genet. 4, e1000198.

Jiang, R.H.Y., Tripathy, S., Govers, F., Tyler, B.M., 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. Proc. Natl. Acad. Sci. U.S.A. 105, 4874–4879.

Johannesson, H., Vidal, P., Guarro, J., Herr, R.A., Cole, G.T., Taylor, J.W., 2004. Positive directional selection in the proline-rich antigen (PRA) gene among the human pathogenic fungi *Coccidioides immitis*, *C. posadasii* and their closest relatives. Mol. Biol. Evol. 21, 1134–1145.

Johannesson, H., Townsend, J.P., Hung, C.Y., Cole, G.T., Taylor, J.W., 2005. Concerted evolution in the repeats of an immunomodulating cell surface protein, SOWgp, of the human pathogenic fungi *Coccidioides immitis* and *C. posadasii*. Genetics 171, 109–117.

Jongwutiwes, S., Putaporntip, C., Karnchaisri, K., Seethamchai, S., Hongsrimuang, T., Kanbara, H., 2008. Positive selection on the *Plasmodium falciparum* sporozoite threonine-asparagine-rich protein: analysis of isolates mainly from low endemic areas. Gene 410, 139–146.

Kaltz, O., Gandon, S., Michalakis, Y., Shykoff, J., 1999. Local maladaptation of the plant pathogen *Microbotryum violaceum* to its host *Silene latifolia*: evidence from a cross-inoculation experiment. Evolution 53, 395–407.

Kayser, M., Brauer, S., Stoneking, M., 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. 20, 893–900.

Kelly, J.K., 1997. A test of neutrality based on interlocus associations. Genetics 146, 1197–1206.

Kidgell, C., Volkman, S.K., Daily, J., Borevitz, J.O., Plouffe, D., Zhou, Y.Y., Johnson, J.R., Le Roch, K.G., Sarr, O., Ndir, O., Mboup, S., Batalov, S., Wirth, D.F., Winzeler, E.A., 2006. A systematic map of genetic variation in *Plasmodium falciparum*. Plos Path. 2, 562–577.

Kim, Y., Nielsen, R., 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167, 1513–1524.

Kim, Y., Stephan, W., 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160, 765–777.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798.

Kosakovsky Pond, S.L., Frost, S.D.W., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., Siepel, A., 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4, e1000144.

Laine, A., 2006. Evolution of host resistance: Looking for coevolutionary hotspots at small spatial scales. Proc. R. Soc. Lond. B Biol. Sci. 273, 267–273.

Lemey, P., Derdelinckx, I., Rambaut, A., Van Laethem, K., Dumont, S., Vermeulen, S., Van Wijngaerden, E., Vandamme, A.-M., 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J. Virol. 79, 11981–11989.

Lewontin, R.C., Krakauer, J., 1973. Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics 74, 175–195.

Li, W.H., Wu, C.I., Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.

Liu, Z., Bos, J., Armstrong, M., Whisson, S., da Cunha, L., Torto-Alalibo, T., Win, J., Avrova, A., Wright, F., Birch, P., Kamoun, S., 2005. Patterns of diversifying selection in the Phytotoxin-like scr74 gene family of *Phytophtora infestans*. Mol. Biol. Evol. 22, 659–672.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–2115.

MacCallum, C., Hill, E., 2006. Being positive about selection. PLoS Biology 4.

Maynard Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23–35.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., Pupko, T., 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics 23, I319–I327.

McCarthy, A., Shaw, M., Goodman, S., 2007. Pathogen evolution and disease emergence in carnivores. Proc. R. Soc. B Lond. 274, 3165–3174.

McDonald, J.H., Kreitman, M., 1991a. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654.

McDonald, J.H., Kreitman, M., 1991b. Neutral mutation hypothesis test—reply. Nature 354, 116–1116.

McVean, G., Spencer, C.C.A., 2006. Scanning the human genome for signals of selection. Curr. Op. Genet. Dev. 16, 624–629.

Messier, W., Stewart, C.B., 1997. Episodic adaptive evolution of primate lysozymes. Nature 385, 151–154.

Meyers, B., Kozik, A., Griego, A., Kuang, H., Michelmore, R., 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15, 809–834.

Miyamoto, M.M., Fitch, W.M., 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12, 503–513.

Miyata, T., Yasunaga, T., 1980. Molecular evolution of messenger-RNA—a method for estimating evolutionary rates of synonymous and amino-acid substitutions from homologous nucleotide-sequences and its application. J. Mol. Evol. 16, 23–36.

Muse, S.V., Gaut, B.S., 1994. A Likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

Nei, M., 2007. The new mutation theory of phenotypic evolution. Proc. Natl. Acad. Sci. U.S.A. 104, 12235–12242.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nielsen, R., 2005. Molecular signatures of natural selection. Ann. Rev. Genet. 39, 197–218.

Nielsen, R., Yang, Z.H., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., Sninsky, J.J., Adams, M.D., Cargill, M., 2005a. A scan of positively selected genes in the genome of humans and chimpanzees. Plos Biol. 3, e170-.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005b. Genomic scans for selective sweeps using SNP data. Genome Res. 15, 1566–1575.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., Clark, A.G., 2007. Recent and ongoing selection in the human genome. Nat. Rev. Genet. 8, 857–868.

Noor, M.A.F., Feder, J.L., 2006. Speciation genetics: evolving approaches. Nat. Rev. Genet. 7, 851–861.

Ogura, M., Perez, J.C., Mittl, P.R.E., Lee, H.K., Dailide, G., Tan, S., Ito, Y., Secka, O., Dailidiene, D., Putty, K., Berg, D.E., Kalia, A., 2007. *Helicobacter pylori* evolution: lineage-specific adaptations in homologs of eukaryotic Sel1-like genes. Plos Comp. Biol. 3, 1455–1467.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, New York.

Patron, N.J., Waller, R.F., Cozijnsen, A.J., Straney, D.C., Gardiner, D.M., Nierman, W.C., Howlett, B.J., 2007. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. BMC Evol. Biol. 7, 174.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441, 1103–1108.

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53, 711–723.

Piertney, S., Oliver, M., 2006. The evolutionary ecology of the major histocompatibility complex. Heredity 96, 7–21.

Piganeau, G., Eyre-Walker, A., 2003. Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. U.S.A. 100, 10335–10340.

Polley, S.D., Chokejindachai, W., Conway, D.J., 2003. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. Genetics 165, 555–561.

Pollock, D.D., Chang, B.S.W., 2007. Dealing with uncertainty in ancestral sequence reconstruction: sampling from the posterior distribution. In: Liberles, D.A. (Ed.)., Ancestral Sequence Reconstruction. Oxford University Press, Oxford.

Prince, V.E., Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3, 827–837.

Przeworski, M., 2002. The signature of positive selection at randomly chosen loci. Genetics 160, 1179–1189.

Pupko, T., Galtier, N., 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. R. Soc. B Lond. 269, 1313–1316.

Putaporntip, C., Seethamchai, S., Suvannadhat, V., Hongsrimuanga, T., Sattabongkot, J., Jongwutiwes, S., 2008. Selective pressure on the merozoite surface protein-1 genes of *Plasmodium vivax*, *P. knowlesi* and *P. cynomolgi*. Asian Biomed. 2, 123–134.

Rajashekar, B., Samson, P., Johansson, T., Tunlid, A., 2007. Evolution of nucleotide sequences and expression patterns of hydrophobin genes in the ectomycorrhizal fungus *Paxillus involutus*. N. Phytol. 174, 399–411.

Rannala, B., Yang, Z.H., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.

Richman, A., 2000. Evolution of balanced genetic polymorphism. Mol. Ecol. 9, 1953–1963.

Robbertse, B., Reeves, J.B., Schoch, C.L., Spatafora, J.W., 2006. A phylogenomic analysis of the Ascomycota. Fung. Genet. Biol. 43, 715–725.

Rockman, M., Hahn, M., Soranzo, N., Goldstein, D., Wray, G., 2003. Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. 13, 2118–2123.

Rodrigue, N., Philippe, H., Lartillot, N., 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics 24, 56–62.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., Liberles, D.A., 2007. Evolution after gene duplication: Models, mechanisms, sequences, systems, and organisms. J. Exp. Zool. B Mol. Dev. Evol. 308B, 58–73.

Roy, H.E., Steinkraus, D.C., Eilenberg, J., Hajek, A.E., Pell, J.K., 2006. Bizarre interactions and endgames: entomopathogenic fungi and their arthropod hosts. Ann. Rev. Entomol. 51, 331–357.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837.

Sainudiin, R., Wong, W.S.W., Yogeeswaran, K., Nasrallah, J.B., Yang, Z.H., Nielsen, R., 2005. Detecting site-specific physicochemical selective pressures: applications to the class IHLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. J. Mol. Evol. 60, 315–326.

Salvaudon, S., Héraudet, V., Shykoff, J.A., 2005. Parasite-host fitness trade-offs change with parasite identity: gentotype-specific interactions in a plant–pathogen system. Evolution 59, 2518–2524.

Salvaudon, L., Giraud, T., Shykoff, J., 2008. Genetic diversity in natural populations: a fundamental component of plant-microbe interactions. Curr. Opin. Plant Biol. 11, 135–143.

Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. Genetics 132, 1161–1176.

Sawyer, S.A., Kulathinal, R.J., Bustamante, C.D., Hartl, D.L., 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J. Mol. Evol. 57, S154–S164.

Scheffler, K., Seoighe, C., 2005. A Bayesian model comparison approach to inferring positive selection. Mol. Biol. Evol. 22, 2531–2540.

Scheffler, K., Martin, D.P., Seoighe, C., 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics 22, 2493–2499.

Schürch, S., Linde, C.C., Knogge, W., Jackson, L.F., McDonald, B.A., 2004. Molecular population genetic analysis differentiates two virulence mechanisms of the fungal avirulence gene NIP1. Mol. Plant Mol. Int. 17, 1114–1125.

Segal, E., 2005. *Candida*, still number one—what do we know and where are we going from there? Mycoses 48, 3–11.

Shackelton, L.A., Parrish, C.R., Truyen, U., Holmes, E.C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U.S.A. 102, 379–384.

Shapiro, B.J., Alm, E.J., 2008. Comparing patterns of natural selection across species using selective signatures. Plos Genet. 4.

Shriner, D., Nickle, D.C., Jensen, M.A., Mullins, J.I., 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. 81 (2), 115–121.

Sicard, D., Pennings, P.S., Grandclement, C., Acosta, J., Kaltz, O., Shykoff, J.A., 2007. Specialization and local adaptation of a fungal parasite on two host plant species as revealed by two fitness traits. Evolution 61, 27–41.

Siepel, A., Haussler, D., 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. J. Comp. Biol. 11, 413–428.

Siltberg, J., Liberles, D.A., 2002. A simple covarion-based approach to analyse nucleotide substitution rates. J. Evol. Biol. 15, 588–594.

Smith, N.G.C., 2003. Are radical and conservative substitution rates useful statistics in molecular evolution? J. Mol. Evol. 57, 467–478.

Smith, N.G.C., Eyre-Walker, A., 2002. Adaptive protein evolution in *Drosophila*. Nature 415, 1022–1024.

Smith, E.E., Sims, E.H., Spencer, D.H., Kaul, R., Olson, M.V., 2005. Evidence for diversifying selection at the pyoverdine locus of *Pseudomonas aeruginosa*. J. Bacteriol. 187, 2138–2147.

Soares, A., Soares, M., Schrago, C., 2008. Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission. J. Mol. Evol. 66, 598–604.

Spencer, C.C.A., Coop, G., 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20 (18), 3673–3675.

Staats, M., van Baarlen, P., Schouten, A., van Kan, J.A.L., Bakker, F.T., 2007. Positive selection in phytotoxic protein-encoding genes of *Botrytis* species. Fung. Genet. Biol. 44, 52–63.

Stahl, E., Bishop, J., 2000. Plant–pathogen arms races at the molecular level. Curr. Opin. Plant Biol. 3, 299–304.

Stephan, W., Song, Y.S., Langley, C.H., 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172, 2647–2663.

Stinchcombe, J.R., Hoekstra, H.E., 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100, 158–170.

Stukenbrock, E.H., McDonald, B.A., 2007. Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. Mol. Plant Pathol. 8, 321–332.

Sun, X.L., Cao, Y.L., et al., 2006. Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. Plant Physiol. 140, 998–1008.

Suzuki, Y., 2004. New methods for detecting positive selection at single amino acid sites. J. Mol. Evol. 59, 11–19.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tanaka, T., Nei, M., 1989. Positive Darwinian selection observed at the variable-region genes of immunoglobulins. Mol. Biol. Evol. 6, 447–459.

Tang, F., Zhang, C., 2007. Evidence for positive selection on the E2 gene of bovine viral diarrhoea virus type 1. Virus Genes 35, 629–634.

Taylor, J.S., Raes, J., 2004. Duplication and divergence: the evolution of new genes and old ideas. Ann. Rev. Genet. 38, 615–643.

Tellier, A., Brown, J.K.M., 2007. Stability of genetic polymorphism in host–parasite interactions. Proc. R Soc. Lond. B Biol. Sci. 274, 809–817.

Tenaillon, M., Tiffin, P., 2007. The quest for adaptive evolution: a theoretical challenge in a maze of data. Curr. Opin. Plant Biol. 11, 1–6.

Tenaillon, M., Tiffin, P., 2008. The quest for adaptive evolution: a theoretical challenge in a maze of data. Curr. Opin. Plant Biol. 11, 1–6.

Tenaillon, M.I., U'Ren, J., Tenaillon, O., Gaut, B.S., 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. 21, 1214–1225.

Teshima, K.M., Coop, G., Przeworski, M., 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16, 702–712.

Thorne, J.L., 2007. Protein evolution constraints and model-based techniques to study them. Curr. Opin. Struct. Biol. 17, 337–341.

Thorne, J.L., Choi, S.C., Yu, J., Higgs, P.G., Kishino, H., 2007. Population genetics without intraspecific data. Mol. Biol. Evol. 24, 1667–1677.

Thrall, P., Burdon, J., Young, A., 2001. Variation in resistance and virulence among demes of a plant host–pathogen metapopulation. J. Ecol. 89, 736–748.

Tiffin, P., Moeller, D.A., 2006. Molecular evolution of plant immune system genes. Trends Genet. 22, 662–670.

Tosa, Y., Osue, J., Eto, Y., Oh, H., Nakayashiki, H., Mayama, S., Leong, S., 2005. Evolution of an avirulence gene, AVR1-CO39, concomitant with the evolution and differentiation of *Magnaporthe oryzae*. Mol. Plant. Microbe Interact. 18, 1148–1160.

Urwin, R., Holmes, E.C., Fox, A.J., Derrick, J.P., Maiden, M.C.J., 2002. Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. Mol. Biol. Evol. 19, 1686–1694.

van der Does, H., Rep, M., 2007. Virulence genes and the evolution of host specificity in plant–pathogenic fungi. Mol. Plant Microbe Interact. 20, 1175–1182.

Verra, F., Chokejindachai, W., Weedall, G.D., Polley, S.D., Mwangi, T.W., Marsh, K., Conway, D.J., 2006. Contrasting signatures of selection on the *Plasmodium falciparum* erythrocyte binding antigen gene family. Mol. Biochem. Parasitol. 149, 182–190.

Vitalis, R., Dawson, K., Boursot, P., 2001. Interpretation of variation across marker loci as evidence of selection. Genetics 158, 1811–1823.

Voight, B.F., Kudaravalli, S., Wen, X.Q., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. Plos Biol. 4, 446–458.

Wagner, A., 2007. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics 176, 2451–2463.

Wakeley, J., 2008. Complex speciation of humans and chimpanzees. Nature 452, E3–E4.

Wang, E.T., Kodama, G., Baidi, P., Moyzis, R.K., 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc. Natl. Acad. Sci. U.S.A. 103, 135–140.

Wang, H.C., Spencer, M., Susko, E., Roger, A.J., 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24, 294–305.

Ward, T.J., Bielawski, J.P., Kistler, H.C., Sullivan, E., O'Donnell, K., 2002. Ancestral polymorphism and adaptative evolution in the trichothecene mycotoxin gene cluster of phytopathogenic *Fusarium*. Proc. Natl. Acad. Sci. U.S.A. 99, 9278–9283.

Warnock, D.W., 2006. Fungal diseases: an evolving public health challenge. Med. Mycol. 44, 697–705.

Wik, L., Karlsson, M., Johannesson, H., 2008. The evolutionary trajectory of the Mating-Type (mat) genes in *Neurospora* relates to reproductive behavior of Taxa. BMC Evol. Biol. 8, 109.

Williams, P.D., Pollock, D.D., Blackburne, B.P., Goldstein, R.A., 2006. Assessing the accuracy of ancestral protein reconstruction methods. Plos Comp. Biol. 2, 598–605.

Wilson, D.J., McVean, G., 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172, 1411–1425.

Win, J., Morgan, W., Bos, J., Krasileva, K.V., Cano, L.M., Chaparro-Garcia, A., Ammar, R., Staskawicz, B.J., Kamoun, S., 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. Plant Cell 19, 2349–2369.

Wong, W.S.W., Nielsen, R., 2004. Detecting selection in noncoding regions of nucleotide sequences. Genetics 167, 949–958.

Wong, W.S.W., Sainudiin, R., Nielsen, R., 2006. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinf. 7, 148.

Wray, G.A., 2007. The evolutionary significance of *cis*-regulatory mutations. Nat. Rev. Genet. 8, 206–216.

Wu, C., Ting, C., 2004. Genes and speciation. Nat. Rev. Genet. 5, 114–122.

Wu, J., Saupe, S.J., Glass, N.L., 1998. Evidence for balancing selection operating at the *het-c* heterokaryon incompatibility locus in a group of filamentous fungi. Proc. Natl. Acad. Sci. U.S.A. 95, 12398–12403.

Yang, Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comp. Appl. Biosci. 13, 555–556.

Yang, Z.H., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573.

Yang, Z.H., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yang, Z.H., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Yang, Z.H., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19, 908–917.

Yang, Z.H., Nielsen, R., 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25, 568–579.

Yang, Z.H., Nielsen, R., Goldman, N., Pedersen, A.M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

Yang, Z.H., Wong, W.S.W., Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.

Zanotto, P.M.D., 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, 1077–1089.

Zhang, J.Z., Kumar, S., Nei, M., 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol. Biol. Evol. 14, 1335–1338.

Zhang, J.Z., Zhang, Y.P., Rosenberg, H.F., 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30, 411–415.

Zhang, J.Z., Nielsen, R., Yang, Z.H., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479.

Zhang, Z., Li, J., Yu, J., 2006. Computing Ka and Ks with a consideration of unequal transitional substitutions. BMC Evol. Biol. 6, 44.

Salvaudon et al., 2008)",13,"Glossary (adapted from Salvaudon et al., 2008)",5,0,1,0,280pt,240pt,0mm,0mm>Glossary (adapted from Salvaudon et al., 2008)

Paralogs:: Genes that have diverged after a gene duplication event, in contrast to orthologs, which have diverged after a speciation event.

Balancing selection:: Evolutionary process which maintains genetic polymorphism within a population by frequency dependent selection (advantage of rare alleles) or overdominance (heretozygote advantage).

Diversifying selection:: Evolutionary process that favors divergent phenotypes.

Directional selection:: Evolutionary process which favors a single and extreme phenotype in an environment.

Positive selection:: At the molecular level, evolutionary process which favors nonsynonymous substitutions to change or to optimize the function of the protein.

Purifying selection *(or stabilizing selection)*:: At the molecular level, evolutionary process which acts against nonsynonymous deleterious substitutions.

Neutral evolution:: Evolutionary process which neither favors nor selects against changes.

Relaxed constraints:: Decreased selective pressure to retain a given sequence and/or function.