The Control of Copy Number of IS*6110* in *Mycobacterium tuberculosis*

Mark M. Tanaka, Noah A. Rosenberg,† and Peter M. Small‡*

*School of Biotechnology & Biomolecular Sciences, University of New South Wales, Australia; †Molecular & Computational Biology, University of Southern California, Los Angeles, California; ‡Stanford University School of Medicine, Stanford, California

Insertion sequence (IS) elements are bacterial genes that are able to transpose to different locations in the genome. These elements are often used in molecular epidemiology as genetic markers that track the spread of pathogens. Transposable elements have frequently been described as "selfish DNA" because they facilitate their own transposition, causing damage when they insert into coding regions, while contributing little if anything to the bacterial host. According to this hypothesis, the expansion of copy number of insertion sequences is opposed by negative selection against high copy numbers. From an alternative point of view, we might expect IS elements to intrinsically regulate transposition within cells, thereby limiting damage to their bacterial host. Here, we report evidence that the copy number of IS*6110* in *Mycobacterium tuberculosis* is controlled by selection against the element. We first construct 12 different models of marker change resulting from a combination of possible transposition functions and selective regimes. We then compute the Akaike Information Criterion for each model to identify the models that best explain data consisting of serial isolates of *M. tuberculosis* genotyped with IS*6110*. We find that the best performing models all include selection against the accumulation of copies. Specifically, our analysis points to the interaction of separate copies of the element causing lethal effects. We discuss the implications of these findings for genome evolution and molecular epidemiology.

Introduction

Insertion sequence (IS) elements and other transposable elements (TEs) are of great interest to evolutionary biologists, geneticists, and molecular epidemiologists. The view that TEs are selfish (Doolittle and Sapienza 1980; Orgel and Crick 1980) has some support from studies that compare distributions of TE copy number with dynamical models in which TEs replicate within genomes and engender a fitness cost to the host (Charlesworth and Charlesworth 1983; Sawyer and Hartl 1986; Sawyer et al. 1987). The survival of TEs is better secured, however, if the proliferation of TE copies is negatively regulated, or is prevented from uncontrolled expansion, especially when the copy numbers are high within genomes. In this article, we will use the terms *regulation* to mean an intrinsic mechanism modifying the transposition rate and *control* to indicate any mechanism, including regulation and selection, that prevents copy numbers from undergoing unchecked expansion. Sawyer et al. (1987) used observed distributions of IS element copy numbers in *Escherichia coli* to demonstrate that some families appear to be regulated (IS*2*, IS*3*, IS*4*, IS*30*) while others do not (IS*1*, IS*5*). These results, however, depend on the assumption that the copy number distributions are at equilibrium, which may not hold for all elements (Tanaka et al. 2000).

In fact, many mechanisms of regulation have been described for various families of IS elements. One of the better-characterized elements in prokaryotes is IS*10*, in which several different mechanisms have been found to regulate transposition. These include the transcription of an antisense RNA that blocks translation of the transposase, *dam*-mediated methylation of the element, and the action of host factors IHF and HU (Kleckner et al. 1996). IS*3* and IS*911* are of particular interest here because they belong to the same group of IS elements as IS*6110* (Fayet et al. 1990; McAdam et al. 1990). The IS*3* transposase is inhibited by proteins (OrfA and OrfB) that are alternatively expressed through a single-base frameshift during translation of the IS*3* message (Sekine, Eisaki, and Ohtsubo 1994). Similarly, it has been shown that IS*911* produces a repressor that competes with the transposase (Haren et al. 2000). Additionally, IS*911* makes use of alternative promoters (P_{IRL} and P_{junc}) to modulate transposition rates: the stronger promoter (P_{junc}) is formed only transiently during transposition (Duval-Valentin et al. 2001). Taken together, these studies highlight the great diversity of mechanisms of regulation operating on IS elements. As it is likely that additional mechanisms have yet to be discovered, it is difficult to determine *a priori* which, if any, regulation mechanisms exist for a given element.

Although IS*6110* is widely used as a genetic marker in the molecular epidemiology of tuberculosis, the details of transposition in IS*6110* are poorly understood. However, some progress has been made by conducting manipulative experiments of IS*6110* in the related species *M. smegmatis*. We now know that particular insertions can alter the expression of nearby genes (Safi et al. 2004), and that transposition rates depend on the genetic background and environmental factors: they are stimulated by the presence of nearby promoters (Wall et al. 1999) and by microaerobic exposure (Ghanekar et al. 1999).

In the epidemiological setting, it is important to know the rate at which the marker changes *in vivo* in order to make inferences about the speed of transmission of the infectious agent (Yeh et al. 1998; de Boer et al. 1999; Tanaka and Rosenberg 2001). It is also important to know if the rate of change is a function of other factors (e.g., Eilers et al. 2004). In the case of insertion sequences, if each copy of the element in a given genome acts independently, then the rate of change (the transposition rate, or more precisely, the substitution rate for IS element genotypic profiles) is a linear function of copy number (Rosenberg, Tsolaki, and Tanaka 2003; Tanaka and

Key Words: *Mycobacterium tuberculosis*, molecular epidemiogy, insertion sequence, transposition rate, regulation, Akaike information criterion.

E-mail: m.tanaka@unsw.edu.au.

Mol. Biol. Evol. 21(12):2195–2201. 2004
doi:10.1093/molbev/msh234
Advance Access publication August 18, 2004

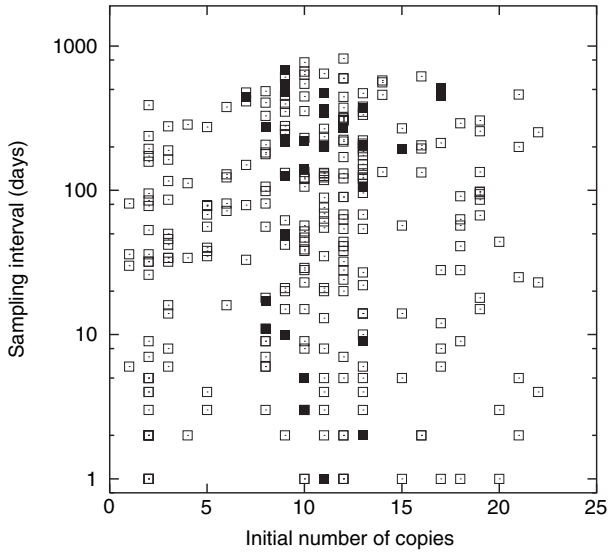

FIG. 1.—Serial samples for 303 time intervals. We plot time intervals against inital copy number for each pair of serial samples. Each pair of samples is represented by a point. Filled squares: intervals involving a change in genotype; open squares: no change in genotype.

Rosenberg 2001). Departures from independence will be reflected in departures from linearity in this relationship.

We offer a novel approach to the detection of control of IS elements, applicable to naturally occurring strains of pathogens. We statistically quantify the copy number control of the insertion sequence IS*6110* in *M. tuberculosis*. Our method is potentially applicable to other IS elements and prokaryotes, although we have not found data sets of the appropriate kind.

Methods

Data

The data set used here is the "Niemann+SF" data set described in Rosenberg, Tsolaki, and Tanaka (2003), consisting of 303 repeated isolates of tuberculosis from patients (see also Niemann, Richter, and Rusch-Gerdes [1999]). The isolates were typed using the standard protocol based on the element IS*6110* (van Embden et al. 1993). By keeping track of genotypes of isolates from repeated visits by the same patient, we have a record of changes in the fingerprint over time. These *serial isolates* allow the quantification of the rate of change of markers over time (we denote this parameter by θ). Our previous studies developed a maximum likelihood method for estimating the rate at which fingerprints change. Applying it to IS*6110* in *M. tuberculosis*, we found this rate to be around $\hat{\theta} = 0.0287$ changes per element per year (Rosenberg, Tsolaki, and Tanaka 2003).

In Figure 1, for the 303 serial isolates, we plot time intervals between repeated visits as a function of copy number. Serial samples that involve a change in the IS*6110* fingerprint are distinguished from those that do not involve a change. It appears from this plot that strains of intermediate copy number (say 7 to 17 copies) are less stable than strains of low and high copy number. The apparent variation in stability across copy number is likely due at least in part to the heavier sampling of strains of intermediate copy number, as depicted in Figure 4b of Rosenberg, Tsolaki, and Tanaka (2003). One of our goals is to explore this issue quantitatively in order to resolve this ambiguity.

General Framework

Our general approach in assessing copy number control is to construct many possible candidate models to describe transposition over time. We will compare these models by using statistical information theory to measure the evidence supporting each model given the data set. This approach was also used by Calabrese and Durrett (2003) for the similar problem of investigating copy numbers for repeat motifs in microsatellite loci.

We start with a general model providing the probability of a fingerprint with a given copy number k changing within a patient in a given time period t. Let $\mathbf{p}$ be the set of parameters of a particular model.

We assume that negative selection results from lethal effects of transposition. We further assume that there is no cost to simply carrying the element, so that a mutation will be selectively neutral or advantageous in the cellular population within the host, provided it survived the transposition event. The negligible metabolic burden of carrying multiple copies of the element justifies this assumption. Let the probability that a mutant survives the effects of a transposition event be $\sigma(k, \mathbf{p})$ and let the probability of a mutant reaching fixation given that it survives transposition be u. Then, if transposition follows a Poisson process with transposition rate $\alpha_g(k, \mathbf{p})$ per genome, substitution follows a marked Poisson process, and is therefore Poisson with rate $\alpha_g(k, \mathbf{p})\sigma(k, \mathbf{p})u$. Note that both the transposition and selection functions may depend on copy number k. Henceforth we omit the parameter u and let it be subsumed by the transposition function so that the change rate (θ, to be described in more detail later) describes the overall substitution process.

Analogously to the model of Rosenberg, Tsolaki, and Tanaka (2003) with "change resolution" and "frequent sampling," the probability w of a change being observed during time interval t is

$$w(k, t, \mathbf{p}) = 1 - e^{-\alpha_g(k,\mathbf{p})\sigma(k,\mathbf{p})t}. \qquad (1)$$

If $\sigma(k, \mathbf{p}) = 1$, then there is no selection against carrying the insertion sequence element. If $\sigma(k, \mathbf{p}) = 1$ and $\alpha(k, \mathbf{p}) = \theta k$, the transposition model is the same as the linear model in Tanaka and Rosenberg (2001) and Rosenberg, Tsolaki, and Tanaka (2003). We will explore particular forms of these two functions in the next section.

Letting G_i indicate whether the ith sample in the data corresponds to a changed fingerprint, the likelihood of the parameters given the data is

$$\mathrm{Lik}(\mathbf{p}) = \prod_{i:G_i=\text{change}} w(k_i, t_i, \mathbf{p}) \times \prod_{i:G_i=\text{no change}} (1 - w(k_i, t_i, \mathbf{p})). \qquad (2)$$

This likelihood function is maximized to find estimates of the parameters in each model. In addition, we compute the Akaike Information Criterion (AIC) value, which is given by $AIC = -2\ln(\mathrm{Lik}(\hat{\mathbf{p}})) + 2n$, where n is the number of

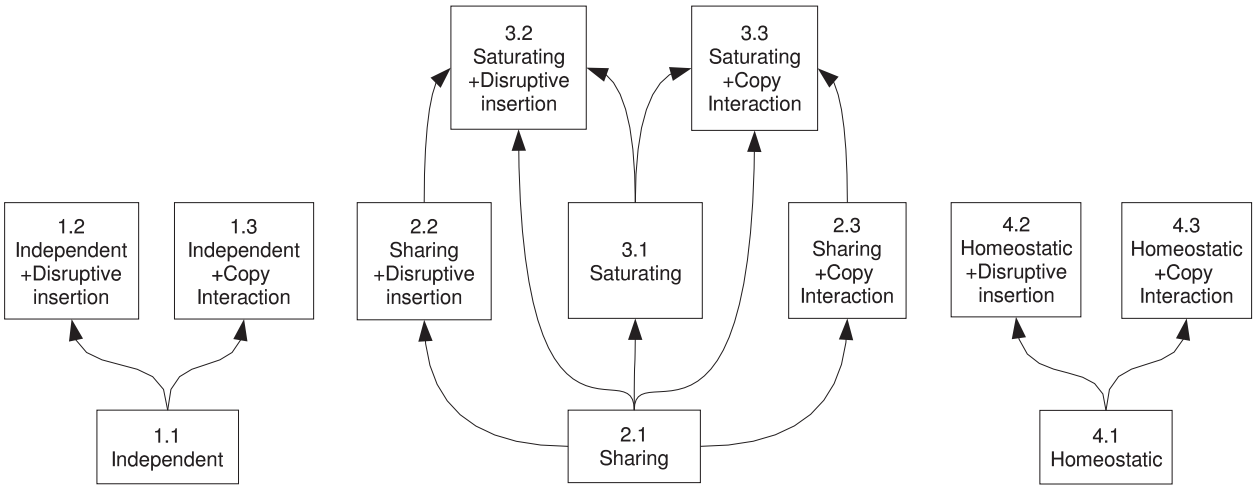

FIG. 2.—The relationships among the various models. Arrows point to more general models; the model at the base of an arrow is nested in the model at the tip. Models involving three parameters appear in the top row, those involving two parameters are in the middle row, and those involving a single parameter are in the bottom row. The models are also labeled by numbers reflecting the transposition and selection components.

parameters in the model, and $\hat{\mathbf{p}}$ is the set of maximum likelihood estimates for the model (Burnham and Anderson 2002). These AIC values enable us to select among competing candidate models. Lower values of AIC indicate superior models.

Selecting from a Set of Models

We now consider different ways to construct the transposition and selection functions and assess these models. We begin with the transposition rate function, $\alpha_g(k, \mathbf{p})$. Transposition is a relatively rare stochastic event depending on the amount of the transposase enzyme available as well as the number of copies of the element. For convenience, we distinguish the *genomic* transposition rate—the overall rate of transposition per genome—from the per-copy rate $\alpha(k, \mathbf{p})$; the two rates are related by $\alpha_g(k, \mathbf{p}) = k\alpha(k, \mathbf{p})$. For brevity, we will drop the arguments when discussing these functions.

Models of IS Element Change

We investigate four models of transposition. Each model may include or exclude selection against the element, and we explore two ways to model selection. The combinations of the transposition and selection models result in $4 \times 3 = 12$ different models of genetic marker change. We describe below the components of all of these models, and depict the relationships among all twelve models in Figure 2.

First, consider the following models of transposition.

1. **Independent.** The first model is the same one we used in previous papers (Tanaka and Rosenberg 2001; Rosenberg, Tsolaki, and Tanaka 2003). It assumes that each copy acts independently of others. Hence, α is set to a constant value θ events per copy per year and the overall genomic transposition rate is $\alpha_g = \theta k$.
2. **Sharing.** Suppose the transposase produced by each copy is available to catalyze transposition of any of the copies. If each copy produces θ units of the enzyme then $\alpha = \theta k$ and, assuming mass action, the genomic transposition rate is $\alpha_g = \theta k^2$. The parameter θ now has units of events per pair of copies per year.
3. **Saturating.** In this model the transposase is again shared, but the amount present is a saturating function of copy number, resulting from limited resources available to produce transposase and other proteins. In other words, the amount of transposase in a given cell approaches some constant limiting value as copy number increases, rather than increasing indefinitely. Taking the same approach as that used to model Michaelis-Menten kinetics, we let
$$\alpha = \frac{\theta k}{1 + rk}$$
where r is a constant associated with this saturating function. When $r = 0$ the model reduces to the transposase sharing model. The parameter θ is again measured in events per pair of copies per year.
4. **Homeostatic.** Suppose that there is perfect homeostasis with respect to transposition such that the genomic transposition rate is unaffected by copy number: $\alpha_g = \theta$. This implies that the concentration of transposase decreases with the reciprocal of copy number ($\alpha = \theta/k$). The parameter θ is then measured in units of events per year.

We now turn to the models of selection.

1. **No selection.** All strains are assigned equal fitness, regardless of their copy numbers. The absence of selection is established by setting $s = 0$ in either of the next two models, so that $\sigma = 1$.
2. **Disruptive Insertion.** If the fitness of the bacterial host is reduced by a factor $1 - s$ per copy of the element, then the selection function is
$$\sigma = (1 - s)^k.$$
We note that this function can alternatively be interpreted as a reduction in the transposition rate due to

Table 1
Comparison of Models by Means of the Akaike Information Criterion (AIC)

Model (label[a])	Number of parameters	AIC	Weights[b]
Sharing + Copy Interaction (2.3)	2	220.65	0.4452
Saturating + Copy Interaction (3.3)	3	222.66	0.1634
Sharing + Disruptive Insertion (2.2)	2	222.84	0.1491
Independent + Copy Interaction (1.3)	2	223.18	0.1256
Saturating + Disruptive Insertion (3.2)	3	224.84	0.0548
Independent + Disruptive Insertion (1.2)	2	225.65	0.0366
Homeostatic (4.1)	1	228.10	0.0107
Homeostatic + Copy Interaction (4.3)	2	229.28	0.0060
Homeostatic + Disruptive Insertion (4.2)	2	230.07	0.0040
Independent (1.1)	1	230.38	0.0034
Saturating (3.1)	2	232.38	0.0013
Sharing (2.1)	1	241.58	0.0000

[a] The models are labeled with numbers corresponding to the model components. See Figure 2.

[b] Akaike weights are computed using $W_i = \exp(-\Delta_i/2)/\sum_{j=1}^{12} \exp(-\Delta_j/2)$ where Δ_i is the difference between the ith AIC and the lowest AIC value, and the sum in the denominator is over all models being considered. This quantity can be interpreted as the weight of evidence in favor of each model within the set of considered models (Burnham and Anderson 2002).

a mechanism intrinsic to the element (or bacterial host) that senses the copy number of the genome. However, from a mechanistic point of view, this model is more appropriately viewed as corresponding to the deleterious effects of insertion.

3. **Copy Interaction.** We examine the possibility that separate copies of the element interact to lower fitness. This may be the result of separate copies interacting to produce lethal genomic rearrangements (Langley et al. 1988; Gray, Tanaka, and Sved 1996). Here, we model this scenario as follows:

$$\sigma = (1 - s)^{k(k-1)/2}.$$

Each pair of copies can produce the deleterious outcome.

The various combinations of the models will also be labeled according to the numbering of the components given above. For example, the model involving sharing of transposase with copy interaction is labeled **2.3**.

Model Selection and Statistics

As described earlier in the General Framework section, we use the various models established in the previous section to find the likelihood values given by equation (2). These likelihoods are then used to derive the AIC values for the different models. Table 1 shows the results of our model-selection analysis; using the estimates for the three best and two worst models, we plot the transposition function w in Figure 3. The model in which transposase is shared by all copies of the element in the genome, combined with negative selection against the element via copy interaction, best explains the data, with an Akaike weight of around 45%. Note that the top six models, with a total weight of ~97%, all include selection in some form. Although the best two models produce very similar results, as shown in Figure 3, Sharing + Copy Interaction has the lower AIC value because of its ability to explain the data more parsimoniously (with one less parameter).

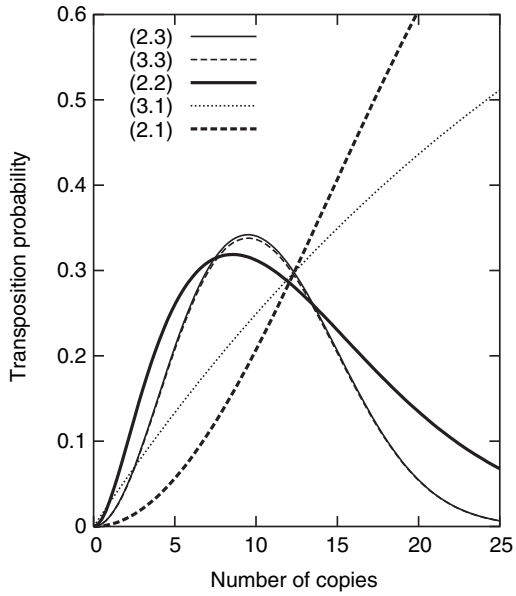

FIG. 3.—The probability of transposition $w(k, t, \mathbf{p})$ in a time period of $t = 1$ year, evaluated at the estimated values of the parameters, for the three best and two worst models. **(2.3)** Sharing + Copy Interaction (thin solid curve); **(3.3)** Saturating + Copy Interaction (thin dashed curve); **(2.2)** Sharing + Disruptive Insertion (thick solid curve); **(3.1)** Saturating (thin dotted curve); **(2.1)** Sharing (thick dashed curve). Note that curves 2.3 and 3.3 lie very close to each other.

Estimation of parameters

For the chosen model (Sharing + Copy Interaction), which carries the greatest weight of evidence among the models, the parameter estimates are: $\hat{\theta} = 0.0119$ per pair of copies per year and $\hat{s} = 0.0231$ per pair of copies. To estimate the uncertainty in these estimates, a standard method is to obtain a variance-covariance matrix by inverting the Fisher information matrix, evaluated at the estimates (see Appendix A for details). With this method, the approximate 95% confidence intervals for the estimates are: (0.00761, 0.0162) for $\hat{\theta}$ and (0.0186, 0.0276) for $\hat{s}$.

Hypothesis testing

Hypotheses regarding whether one of a pair of models has a "significantly better" fit than the other can readily be tested. Because there is some nesting of the models (as shown in Figure 2), we can use likelihood ratio tests (LRTs) to compare certain pairs of models. The test statistic values for LRTs can be extracted from the AIC values in Table 1. Table 2 shows the results of all 13 tests. All but four tests gave rise to significant refinements to the simpler model.

Discussion

Evidence for Selection Against IS*6110*

We have identified a single model, Sharing + Copy Interaction, which outperforms the others in the set of models we examined. Although the evidence in support of this particular model is high, it is not necessarily the "correct model" but one which best explains the data among the set of models. The Saturating and Homeostasis models describe different ways in which the element is

Table 2
Comparison of Models Using Likelihood Ratio Tests

Models (labels)	df[a]	LRT[b]	*p* Value
Independent vs Independent + Disruptive Insertion (1.1, 1.2)	1	6.73	0.0095
Independent vs Independent + Copy Interaction (1.1, 1.3)	1	9.20	0.0024
Sharing vs Saturating (2.1, 3.1)	1	11.20	0.00082
Sharing vs Sharing + Disruptive Insertion (2.1, 2.2)	1	20.74	$<10^{-4}$
Sharing vs Sharing + Copy Interaction (2.1, 2.3)	1	22.93	$<10^{-4}$
Sharing vs Saturating + Disruptive Insertion (2.1, 3.2)	2	20.74	$<10^{-4}$
Sharing vs Saturating + Copy Interaction (2.1, 3.3)	2	22.93	$<10^{-4}$
Saturating vs Saturating + Disruptive Insertion (3.1, 3.2)	1	9.54	0.0020
Saturating vs Saturating + Copy Interaction (3.1, 3.3)	1	11.73	0.00062
Sharing + Disruptive Insertion vs Saturating + Disruptive Insertion (2.2,3.2)	1	0.00	1
Sharing + Copy Interaction vs Saturating + Copy Interaction (2.3, 3.3)	1	0.00	1
Homeostatic vs Homeostatic + Disruptive Insertion (4.1, 4.2)	1	0.03	0.86
Homeostatic vs Homeostatic + Copy Interaction (4.1, 4.3)	1	0.83	0.36

[a] Degrees of freedom.

[b] The likelihood ratio test statistic, given by $\mathrm{LRT} = -2[\ln(\mathrm{Lik}(\hat{\mathbf{p}}_1)/\mathrm{Lik}(\hat{\mathbf{p}}_2))]$, where $\hat{\mathbf{p}}_1$ is the vector of maximum likelihood estimates under the more specific model, and $\hat{\mathbf{p}}_2$ is the corresponding vector for the general model (with more parameters than the former). This test statistic is distributed approximately as χ^2_n with degrees of freedom n, which is given by the difference in the number of parameters between the two models.

intrinsically regulated. Neither of these models, however, is as strongly supported by the data as the case in which the element drives its own proliferation by sharing the transposase molecules produced, with only selection against expansion holding copy number down. Interestingly, this supports the idea that transposable elements are "selfish" in that they spread in genomes despite the deleterious side-effects associated with this spread. We stress that we have not ruled out molecular mechanisms that regulate the spread of copies; if those mechanisms are present, their effects are not strong enough to be detected by the models we have considered.

Notably, the best six models in Table 1 all involve negative selection against the element. Even if copy interaction is not the actual cause of selection, it is likely that some form of negative selection is acting on the element. It is also interesting that the best two models both involve transposase sharing. This supports the hypothesis that separate copies of the element do not operate independently. In agreement with the findings of van der Spuy et al. (2003), the current evidence suggests that the linear model of transposition (Independent) should be replaced for IS*6110*; our analysis offers some alternatives.

Transposition Rate as a Function of Copy Number

This study concerns transposition rates as functions of copy number, a topic of relevance to molecular epidemiology and molecular evolution. However, the body of research on molecular mechanisms regulating transposition rate does not often consider the effect of other copies. If the molecular mechanisms are to be effective in regulating the expansion of insertion sequences, they should preferentially lower transposition rates in high-copy strains. If molecular mechanisms simply slow down the transposition rate regardless of copy number, the uncontrolled proliferation of copies may be temporarily retarded, but in the long run, the copy numbers may increase to extreme levels.

It is also important to consider transposition rates changing as a function of genetic and environmental factors. For instance, there may be location effects—rates may depend on insertion position in the genome; transposition may occur as a "stress response." Another way to put this is that there could be heterogeneities in the rate over space and time. It is possible that the isolates of intermediate copy number in our sample represent a set of strains that are predisposed to change.

The examination of the mobility of IS elements as a function of copy number raises the question of what will happen to the population of insertion sequences in a species in the long term. Will IS*6110* go extinct in the long term or will it persist? The persistence of IS*6110* may be allowed by a balance between element replication and negative selection against copies, as suggested by our analysis. There may be occasional beneficial effects produced by the element. Although the element probably does not move between bacterial cells at a pace rapid enough to escape its destructive effects, the long-term rate of (possibly trans-specific) horizontal transfer may be sufficient to ensure survival (Bergstrom, Lipsitch, and Levin 2000). The extinction of IS*6110* is also a possible long-term outcome. There is no *a priori* reason to expect a family of IS elements to evolve strategies to create "safe" equilibrium distributions of copy number. The peaked distributions of IS*6110* copy number suggest that the dynamics of the element are out of equilibrium, which may reflect a transient presence of the element in *M. tuberculosis* (Tanaka et al. 2000).

It is not known whether elements other than IS*6110* follow the same process of copy number control suggested by our analysis, but the analysis could easily be adapted for other data sets. It should be possible to design experiments using well-characterised elements and host species (e.g., IS*3* or IS*10* in *E. coli*) to study a range of alternative models as done here.

Genomic Conflict: Something to Fight About?

Two main alternative views exist about the relationship between IS elements (and other mobile genes) and the rest of the genome, which can be discussed in terms of the metaphor of genomic conflict. First, insertion sequences might be *selfish*, implying that they replicate within genomes despite causing deleterious effects in the host genome. According to this metaphor, it is in the evolutionary interest of the insertion sequence to increase its replication rate, whereas in contrast, the genome should do the opposite—namely, down-regulate the rate of transposition. A second and opposing view is that the genome

is a well-coordinated system that has resolved most conflicts or inefficiencies. That is, insertion sequences have a role in the genome to produce beneficial effects aligned with the interests of the rest of the genome.

Although the results of this study favor the first view, the two views are not mutually exclusive. The evolution of mechanisms that regulate copy number effectively would benefit both host and element in the case of organisms that undergo little genetic exchange, such as *M. tuberculosis*. Furthermore, it is likely that insertion sequences, like all mutation rate modifiers, produce both beneficial and deleterious effects as they undergo transposition (Chao et al. 1983). In the context of pathogenic bacteria, an important example of the adaptive role of insertion sequences is their complicity in the acquisition of antibiotic resistance genes and virulence factors. As the workings of bacterial genomes are unraveled, we will need to assess the role of IS elements: how they affect genome organization and give rise to genetic innovation.

Molecular Epidemiology and IS Elements

Insertion sequences have been widely exploited for genotyping bacterial pathogens, many of which have little variation at individual nucleotides. The mycobacterial insertion sequence IS*6110* exhibits great variability in both copy number and genomic location (Hermans et al. 1990; McAdam et al. 1990; Stanley and Saunders 1996), making it a valuable tool for studying tuberculosis. IS*6110*-based genotyping is the most widely used marker for molecular epidemiologic studies that have provided fundamental insights into the contemporary transmission and pathogenesis of tuberculosis (Small et al. 1994).

In order to use any genetic marker rationally, however, we must know something about its underlying biology. For example, if a marker changes very slowly, clusters of identical genotypes overestimate the severity of disease transmission, whereas if it evolves very fast, clusters will differentiate quickly and an outbreak may be underestimated.

In the analysis of clusters of IS*6110*-based genotypes, it is important to recognise that strains with different copy numbers evolve at different rates. This study demonstrates statistically that strains with intermediate copy numbers (7–17) are substantially less stable than strains of low and high copy numbers. Thus, for intermediate copy numbers, more permissive definitions of clusters might be used.

Insertion Sequences and Error Catastrophe

We tentatively raise an intriguing medical implication following from an understanding of IS element control. Our analysis demonstrates the presence of negative selection against IS*6110* increasing with copy number. If it is possible to increase the rate of transposition, sufficient damage may be caused to the genome to lead to the demise of the bacterial host. Hence, it may be possible to develop a drug treatment that targets IS*6110* by interfering with its regulation within *M. tuberculosis*. One advantage of such a drug for tuberculosis would be its specificity to bacterial transposition. A potential difficulty, as with many antibacterials, is the probable evolution of resistance.

Although a treatment of this kind is not likely to be soon attainable, there is a precedent for this idea in antiviral therapy. Ribavirin works by elevating the mutation rate beyond the "catastrophe threshold" such that the viral population is no longer viable (Crotty et al. 2000). Also related is the phenomenon of hybrid dysgenesis in Drosophila caused by P elements (Kidwell, Kidwell, and Sved 1977), in which the removal of transposase repression leads to elevated transposition rates and consequently to deleterious effects to the genome.

Acknowledgments

We thank Dmitri Petrov for making important suggestions and William Dunsmuir for helpful discussions. This work was supported by a Faculty Research Grant from the University of New South Wales to M.M.T. and by a National Science Foundation (NSF) Postdoctoral Fellowship in Biological Informatics to N.A.R. P.M.S. was supported by National Institutes of Health (NIH) Grant AI34238 and Wellcome Trust grant 176W009.

Appendix A: Standard Errors

Here, we outline how standard errors are computed for the general model. The Fisher information matrix is additive when the data points are independent from each other. We shall thus begin with the information from each sampled interval. First, define $l_i = \ln(\mathrm{Lik}_i(\mathbf{p}))$, which is the log-likelihood for the ith interval, where $\mathbf{p}$ is the vector of parameters. In other words,

$$l_i = \ln(\mathrm{Lik}_i(\mathbf{p})) = \begin{cases} \ln(w_i) & \text{if } G_i = \text{change} \\ \ln(1 - w_i) & \text{if } G_i = \text{no change,} \end{cases}$$

where $w_i = w(k_i, t_i, \mathbf{p})$.

The (a, b)th element of the information matrix for the single observation is given by

$$I_{iab} = -E\left[\frac{\partial^2 l_i}{\partial p_a \partial p_b}\right], \tag{3}$$

where E[•] represents the expectation and p_j is the jth parameter. Defining $\beta_i = \alpha_g(k_i, \mathbf{p})\sigma(k_i, \mathbf{p})t$, equation (3) simplifies to

$$I_{iab} = \left(\frac{\partial w_i}{\partial p_a}\right)\left(\frac{\partial w_i}{\partial p_b}\right)\frac{1}{w_i(1 - w_i)} \tag{4}$$

$$= \left(\frac{\partial \beta_i}{\partial p_a}\right)\left(\frac{\partial \beta_i}{\partial p_b}\right)\left(\frac{e^{-\beta_i}}{1 - e^{-\beta_i}}\right). \tag{5}$$

The Fisher information matrix I is computed by constructing the matrix with elements (a, b) given by the sum over all data points:

$$I_{ab} = \sum_i I_{iab}\big|_{\mathbf{p}=\hat{p}}. \tag{6}$$

The variance-covariance matrix is the inverse of this matrix, which can readily be evaluated numerically.

Literature Cited

Bergstrom, C. T., M. Lipsitch, and B. R. Levin. 2000. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. Genetics **155**:1505–1519.

Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Calabrese, P., and R. Durrett. 2003. Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. Mol. Biol. Evol. **20**:715–725.

Chao, L., C. Vargas, B. B. Spear, and E. C. Cox. 1983. Transposable elements as mutator genes in evolution. Nature **303**:633–635.

Charlesworth, B., and D. Charlesworth. 1983. The population dynamics of transposable elements. Genet. Res. **42**:1–27.

Crotty, S., D. Maag, J. Arnold, W. Zhong, J. Lau, Z. Hong, R. Andino, and C. Cameron. 2000. The broad-spectrum antiviral ribonucleoside ribavirin is an RNA virus mutagen. Nat. Med. **6**:1375–1379.

De Boer, A. S., M. W. Borgdorff, P. E. W. de Haas, N. J. D. Nagelkerke, J. D. A. van Embden, and D. van Soolingen. 1999. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J. Infect. Dis. **180**:1238–1244.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

Duval-Valentin, G., C. Normand, V. Khemici, B. Marty, and M. Chandler. 2001. Transient promoter formation: a new feedback mechanism for regulation of IS*911* transposition. EMBO J. **20**:5802–5811.

Eilers, P. H. C., D. V. Soolingen, N. T. N. Lan, R. M. Warren, and M. W. Borgdorff. 2004. Transposition rates of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns. J. Clin. Microbiol. **42**:2461–2464.

Fayet, O., P. Ramond, P. Polard, M. F. Prere, and M. Chandler. 1990. Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? Mol. Microbiol. **4**:1771–1777.

Ghanekar, K., A. McBride, O. Dellagostin, S. Thorne, R. Mooney, and J. McFadden. 1999. Stimulation of transposition of the *Mycobacterium tuberculosis* insertion sequence IS*6110* by exposure to a microaerobic environment. Mol. Microbiol. **33**:982–993.

Gray, Y. H. M., M. M. Tanaka, and J. A. Sved. 1996. P-element-induced recombination in *Drosophila melanogaster*: hybrid element insertion. Genetics **144**:1601–1610.

Haren, L., C. Normand, P. Polard, R. Alazard, and M. Chandler. 2000. IS*911* transposition is regulated by protein-protein interactions via a leucine zipper motif. J. Mol. Biol. **296**: 757–768.

Hermans, P. W., D. van Soolingen, J. W. Dale, A. R. Schuitema, R. A. McAdam, D. Catty, and J. D. van Embden. 1990. Insertion element IS*986* from *Mycobacterium tuberculosis*: a useful tool for diagnosis and epidemiology of tuberculosis. J. Clin. Microbiol. **28**:2051–2058.

Kidwell, M. G., J. F. Kidwell, and J. A. Sved. 1977. Hybrid dysgenesis: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics **86**:813–833.

Kleckner, N., R. M. Chalmers, D. Kwon, J. Sakai, and S. Bolland. 1996. Tn10 and IS*10* transposition and chromosome rearrangements: mechanism and regulation in vivo and in vitro. Curr. Top. Microbiol. Immunol. **204**:49–82.

Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52**:223–235.

McAdam, R. A., P. W. Hermans, D. van Soolingen, Z. F. Zainuddin, D. Catty, J. D. van Embden, and J. W. Dale. 1990. Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS*3* family. Mol. Microbiol. **4**:1607–1613.

Niemann, S., E. Richter, and S. Rusch-Gerdes. 1999. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J. Clin. Microbiol. **37**:409–412.

Orgel, L. E. and F. H. C. Crick. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

Rosenberg, N. A., A. G. Tsolaki, and M. M. Tanaka. 2003. Estimating change rates of genetic markers using serial samples: applications to the transposon IS*6110* in *Mycobacterium tuberculosis*. Theor. Popul. Biol. **63**:347–363.

Safi, H., P. F. Barnes, D. L. Lakey, H. Shams, B. Samten, R. Vankayalapati, and S. T. Howard. 2004. IS*6110* functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. Mol. Microbiol. **52**:999–1012.

Sawyer, S., D. E. Dykhuizen, R. F. DuBose, L. Green, T. Mutagadura-Mhlanga, D. F. Wolczyk, and D. L. Hartl. 1987. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics **115**:51–63.

Sawyer, S. and D. Hartl. 1986. Distribution of transposable elements in prokaryotes. Theor. Popul. Biol. **30**:1–16.

Sekine, Y., N. Eisaki, and E. Ohtsubo. 1994. Translational control in production of transposase and in transposition of insertion sequence IS*3*. J. Mol. Biol. **235**:1406–1420.

Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik. 1994. The epidemiology of tuberculosis in San Francisco: A population-based study using conventional and molecular methods. N. Engl. J. Med. **330**:1703–1709.

Stanley, J., and N. Saunders. 1996. DNA insertion sequences and the molecular epidemiology of Salmonella and Mycobacterium. J. Med. Microbiol. **45**:236–251.

Tanaka, M. M., and N. A. Rosenberg. 2001. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Stat. Med. **20**:2409–2420.

Tanaka, M. M., P. M. Small, H. Salamon, and M. W. Feldman. 2000. The dynamics of repeated elements: applications to the epidemiology of tuberculosis. Proc. Natl. Acad. Sci. U. S. A. **97**:3532–3537.

Van der Spuy, G. D., R. M. Warren, M. Richardson, N. Beyers, M. A. Behr, and P. D. van Helden. 2003. Use of genetic distance as a measure of ongoing transmission of *Mycobacterium tuberculosis*. J. Clin. Microbiol. **41**:5640–5644.

Van Embden, J. D. A., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick, et al. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting—recommendations for a standardized methodology. J. Clin. Microbiol. **31**:406–409.

Wall, S., K. Ghanekar, J. McFadden, and J. W. Dale. 1999. Context-sensitive transposition of IS*6110* in Mycobacteria. Microbiology **145**(Pt 11):3169–3176.

Yeh, R. W., A. Ponce De Leon, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small. 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J. Infect. Dis. **177**:1107–1111.

Jonathan Eisen, Associate Editor

Accepted July 26, 2004