Birth, Death, and Diversification of Mobile Promoters in Prokaryotes

Mark W.J. van Passel,^{*,1} Harm Nijveen,[†] and Lindi M. Wahl^{‡,2}

^{*}Laboratory of Systems and Synthetic Biology and [†]Laboratory of Bioinformatics, Wageningen University, Wageningen 6703 HB, the Netherlands, and [‡]Applied Mathematics, Western University, London, Ontario N6A 5B7, Canada

ABSTRACT A previous study of prokaryotic genomes identified large reservoirs of putative mobile promoters (PMPs), that is, homologous promoter sequences associated with nonhomologous coding sequences. Here we extend this data set to identify the full complement of mobile promoters in sequenced prokaryotic genomes. The expanded search identifies nearly 40,000 PMP sequences, 90% of which occur in noncoding regions of the genome. To gain further insight from this data set, we develop a birth–death–diversification model for mobile genetic elements subject to sequence diversification; applying the model to PMPs we are able to quantify the relative importance of duplication, loss, horizontal gene transfer (HGT), and diversification to the maintenance of the PMP reservoir. The model predicts low rates of HGT relative to the duplication and loss of PMP copies, rapid dynamics of PMP families, and a pool of PMPs that exist as a single copy in a genome at any given time, despite their mobility. We report evidence of these “singletons” at high frequencies in prokaryotic genomes. We also demonstrate that including selection, either for or against PMPs, was not necessary to describe the observed data.

PROKARYOTIC genomes are highly dynamic, as the processes of gene acquisition through horizontal gene transfer (HGT), gene duplication, and gene loss combine to diversify the genomes of even closely related strains (Ochman *et al.* 2000; van Passel *et al.* 2007, 2008). Evolutionary modifications to transcriptional circuits, or “transcriptional rewiring,” may be a critical aspect of this rapid gene turnover, as regulatory circuits re-equilibrate to gene acquisition and loss events (Perez and Groisman 2009a,b). For example, genes acquired through HGT may be silenced by the recipient genome (Baños *et al.* 2009), and expression then requires the evolution of a functional promoter: either *de novo* or copied from existing promoters elsewhere in the genome. Similarly, recent evidence from long-term experimental evolution has shown that promoter capture may underlie key steps in adaptive evolution (Blount *et al.* 2012).

In some bacteria, mobile elements that can disperse throughout the genome and reregulate gene expression have been identified, such as Correia elements (Siddique *et al.* 2011), ERICs (De Gregorio *et al.* 2005), and REPINS (Bertels and Rainey 2011). To investigate the prevalence of mobile regulatory elements, Matus-Garcia and co-workers performed an extensive search for mobile promoters in sequenced prokaryote genomes (Matus-Garcia *et al.* 2012; Nijveen *et al.* 2012). The putative mobile promoters (PMPs) identified in these studies were homologous promoter sequences with nonhomologous upstream and downstream coding sequences, representing instances of promoter mobilization within a genome, not duplication events involving both coding sequence and promoter. Putative mobile promoters were identified in 1043 of 1360 genomes tested, with some genomes containing >20 families of PMPs and particular families containing over 90 homologous copies of a PMP in a single genome. Similar to the transmission of insertion sequences (ISs) via horizontal gene transfer (Touchon and Rocha 2007), homologous PMPs were identified in phylogenetically distinct taxa (Nijveen *et al.* 2012), suggesting they could be functional in different genomic contexts.

The abundance of these promoter sequences both within and among genomes suggests a number of intriguing questions about their maintenance, diversity, and evolutionary dynamics, similar to the well-studied issues surrounding transposable elements in eukaryotes (Brookfield 1986b, 2005; Charlesworth

Copyright © 2014 by the Genetics Society of America

doi: 10.1534/genetics.114.162883

Manuscript received October 8, 2013; accepted for publication February 21, 2014; published Early Online February 27, 2014.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162883/-/DC1>.

¹Current address: National Institute for Public Health, Centre for Infectious Disease Control, P.O. Box 1, 3720 BA Bilthoven, the Netherlands

²Corresponding author: Western University, Applied Math, London, ON N6A 5B7, Canada. E-mail: lwahl@uwo.ca

and Langley 1989; Le Rouzic and Deceliere 2005; Blumenstiel 2011). For transposable elements (TEs), a rich history in population genetics has informed our understanding of the distribution of copy number within genomes (Charlesworth and Charlesworth 1983; Langley *et al.* 1983), the degree of relatedness between copies in a family (Ohta 1985; Slatkin 1985; Brookfield 1986a; Hudson and Kaplan 1986), and the role of selection in limiting, or maintaining, copy number (Hickey 1982; Charlesworth and Charlesworth 1983; Golding *et al.* 1986; Edwards and Brookfield 2003; Le Rouzic *et al.* 2007). These approaches typically include transposition and excision, as well as selection, drift, recombination, and mutation, which drives the diversification of TE copies. Various limiting cases have analytically exact solutions (for example, neutral selection, infinite population size, or the absence of excision), while more complex cases have been investigated numerically and by simulation (e.g., Le Rouzic and Deceliere 2005; Le Rouzic *et al.* 2007).

In contrast, for transposable elements in prokaryotes (Wagner 2009), the equilibrium distribution of copy number can be recovered using a simpler branching process or Markov chain approach (Sawyer and Hartl 1986; Moody 1988; Basten and Moody 1991; but also see Dolgin and Charlesworth 2006; Wagner 2006). This approach has proven particularly useful for direct comparisons with data from insertion sequences, suggesting that horizontal gene transfer is essential in maintaining bacterial ISs (Hartl and Sawyer 1988), while purifying selection may not be necessary to limit their expansion (Bichsel *et al.* 2013).

Although the literature surrounding TEs in both prokaryotes and eukaryotes is clearly diverse and well developed, our understanding of this new set of mobile genetic elements, PMPs, is limited. In the sections to follow, we hope to improve this understanding in two ways. First, we substantially extend the available data regarding the distribution of PMPs within and among sequenced prokaryotic genomes. The data set we present more than triples the known number of PMP elements, revealing ~40,000 PMP sequences. Second, we develop a model of the dynamics of PMP copies and families, motivated by previous work modeling TEs in prokaryotes (Sawyer and Hartl 1986; Moody 1988; Basten and Moody 1991; Bichsel *et al.* 2013). The model includes duplication and excision, as well as the possibility of horizontal gene transfer and genetic diversification. Applying the model to our extended data set, we perform careful model identification and reduction; this process allows us to draw robust conclusions regarding the relative importance of horizontal gene transfer, duplication, and diversification in the maintenance of the PMP reservoir. Since the model predicts that many genomes will contain only a single copy of a given PMP family, we also designed and implemented a search strategy with which to identify these solitary PMPs.

Methods

Extending the reservoir

In previous work, PMPs were identified by searching each fully sequenced prokaryote genome for homologous sequen-

ces in promoter regions adjacent to genes, 150–50 nucleotides upstream of the translation start sites (TLSs). Sequences in promoter regions that shared 80% identity over at least 50 nucleotides, but had nonhomologous genes both upstream and downstream, were identified as PMP families (Matus-Garcia *et al.* 2012). However, since mobile elements can also move to other locations in the genome, here we extend the analysis by searching the entire genome for sequences that are homologous to previously identified PMPs in that genome. This uncovers the full complement of PMPs per genome; although the number of PMP families in each genome does not increase, the total number of PMP elements per family can. We refer to the results of this search as the “full” data set; this data set was used throughout the model selection, data fitting, and sensitivity analyses described below.

Searching for singletons

Matus-Garcia *et al.* (2012) used intragenomic comparisons to independently identify promoter families in each prokaryote genome. However, in several instances families that apparently occurred in close homology across several genomes were identified. These families were therefore clustered to identify a subset of nonredundant PMPs.

A limitation of the initial analysis was that to identify PMPs, at least two homologous sequences in a single genome were necessary. In this study, we further extend the data set by using the set of nonredundant PMP sequences, searching each genome for these previously identified families from other genomes. This extended search is able to identify single copies of a PMP sequence, if the sequence is homologous to a PMP family in another genome. The resulting “singleton data set” is thus limited to PMPs that exist in multiple copies in at least one other prokaryote genome. We use this data set to verify the prediction that PMP sequences should frequently occur as a single copy in a given genome, as described in the final subsection of the *Results*, *Mapping singleton PMPs*. This data set is not used in any other section of the *Methods* or *Results*.

In both the full genome and singleton searches, the query was performed as described previously, using the same cutoff threshold (Matus-Garcia *et al.* 2012). In both cases we also included a randomized query, *i.e.*, permuted versions of the query sequences, as a negative control to quantify the false detection rate.

The birth–death–diversification model

Copy-level model: We develop a model of PMP dynamics at two levels, describing both the number of promoter copies in each intragenomic PMP family, n , and the number of PMP families per genome, m . The approach is quite general, and similar to multitype branching processes that have been applied to other mobile genetic elements (Sawyer and Hartl 1986; Moody 1988; Basten and Moody 1991; Bichsel *et al.* 2013): in developing the model we avoid assumptions that are specific to promoters. Although we develop the model in

terms of a Markov chain, we later compare only the predicted equilibrium distribution of the state space to the data. Thus the model is essentially deterministic, describing the average rates at which, for example, new copies of a promoter sequence arrive via HGT in a genome.

As illustrated in Figure 1A, u and w give the rate at which each PMP sequence is duplicated in the same genome or lost from the genome. In a standard birth–death process, each copy has an independent chance of duplication or excision, and these processes occur at overall rates nu and nw per generation. We also investigated the more general model $n^{\alpha_u}u$ and $n^{\alpha_w}w$, with α_u and α_w either 0 or 1 (Bichsel *et al.* 2013). Note that the excision rate w is the rate at which a copy in the family is lost, either by excision or by sequence divergence, which results in the loss of mobility. For duplication, although the mechanism is unknown, PMPs might, for example, be mobilized by transposases acting in *trans*; at least some identified PMPs have flanking inverted repeats (Matus-Garcia *et al.* 2012), suggesting possible mobility signatures, as seen in other short nonautonomous mobile elements (Stavrinides *et al.* 2012).

To capture the dynamics of related PMP families within genomes, we added the possibility of sequence diversification to the standard birth–death model. As shown in Figure 1, if a copy in a family of n copies diversifies, the family has $n - 1$ copies, and a new “family” of a single copy, a singleton, is created. We let v denote the per-copy probability of diversification, so that a family of n copies diversifies at rate $n^{\alpha_v}v$, where again we investigate $\alpha_v = 0$ or 1.

Finally, *de novo* mobilization, or recruitment of new mobile promoters via HGT, may be important. To capture these effects, we include a parameter η , which describes the rate at which new singleton promoters enter the genome; these arise from families of n copies at HGT rate $n^{\alpha_\eta}\eta$, with α_η either 0 or 1. Thus the full model has potentially four free parameters— u , v , w , and η —and four unknown exponents, the corresponding α values.

The probabilities outlined above define a transition matrix, T , of the associated Markov chain in a straightforward way. If T_{ij} gives the rate at which families with j copies create families with i copies in the next generation, we have

$$\begin{aligned} T_{i,i-1} &= (i-1)^{\alpha_u} u \\ T_{i,i} &= 1 - i^{\alpha_u} u - i^{\alpha_v} v - i^{\alpha_w} w \\ T_{i,i+1} &= (i+1)^{\alpha_w} w + (i+1)^{\alpha_v} v. \end{aligned} \quad (1)$$

In addition to these terms, for $j \geq 2$ we must add $j^{\alpha_v}v$ and $j^{\alpha_\eta}\eta$ to $T_{1,j}$ to capture the addition of singletons via diversification and HGT, respectively. Setting N , the maximum number of copies per family, to be well above the maximum in the observed data, we compute the eigenvector associated with the spectral radius of T ; this yields c_n , the expected fraction of families with n copies at equilibrium. We note that fitting c_n to the data yields ratios of the parameters; that is, we obtain predicted values for the ratios v/u , w/u , and η/u , but not each parameter in isolation.

A Copy-level (copies per family)

B Genome-level (families per genome)

Figure 1 Diagram of the model at two levels. (A) promoter copies per family, and (B) promoter families per genome.

Expanding the model to the genome level: At the genome level, the model is also based on a birth–death process, as shown in Figure 1B. The diversification rate p describes the rate at which a new *family* is created from existing families. In other words, the rate at which a genome with m families gains a family (by diversification of an existing family) is given by $m^{\alpha_p}p$, with $\alpha_p = 0$ or 1. Likewise, especially since most of the families have small numbers of copies, a family might be lost. This happens at rate q per family, for an overall transition rate from m families to $m - 1$ families of $m^{\alpha_q}q$, where $\alpha_q = 0$ or 1. For $\alpha_p = \alpha_q = 1$, this is a standard birth–death process.

To this model we add μ , which gives the rate at which new mobile promoter families enter the genome, either via *de novo* mobilization or HGT of a new family. We assume that this rate does not depend on the number of families already in the genome.

For given values of p , q , and μ , we can again define the transition matrix of the associated Markov process and thus find the equilibrium distribution, that is, the expected proportion of genomes that contain m families, f_m . The distribution f_m can also be computed analytically by balancing transitions between states.

Note that f_m gives the expected fraction of genomes with m families, where the count m includes families of a single copy. In contrast, the full data set includes families of two or more copies. Since c_1 gives the fraction of families that are singletons, a genome with m families contains i singletons with probability

$$P_{m,i} = \binom{m}{i} c_1^i (1 - c_1)^{m-i}.$$

We thus find $\hat{f}_m$, the expected fraction of genomes with m *nonsingleton* families by looking at genomes with $k \geq m$ families in total and computing the probability that these genomes have exactly $k - m$ singletons:

$$\hat{f}_m = \sum_{k=m}^M P_{k,k-m} f_k, \quad (2)$$

where M is the maximum number of families per genome.

Integrating the two levels: One of the powerful features of the birth–death–diversification model is that the rate constants at the genome level are functions of the parameters at the copy level. This allows us to fit the model at both levels simultaneously to the entire data set. At the genome level, each family is lost at rate q , while at the copy level, families are lost only if they are singletons, and the singleton is excised. This allows us to write

$$q = c_1 w. \quad (3)$$

At the genome level, families diversify to form new families at rate p . At the copy level, a new family is formed when one copy in an existing family diversifies, as long as there are two copies or more in the original family. The expected rate at which each family adds new families to the genome is thus

$$p = v \sum_{n=2}^N c_n n^{\alpha_v}. \quad (4)$$

Finally, we compute the rate at which new families arrive in each genome by HGT. On average, each genome has $\bar{f} = \sum m f_m$ families, and each family of n copies generates new singletons at rate $n^{\alpha_n} \eta$ per generation. Weighting by the expected distribution of n , we find the rate at which each genome generates new singleton families by HGT,

$$\mu = \eta \bar{f} \sum_{n=1}^N c_n n^{\alpha_n}, \quad (5)$$

which must also equal the rate at which new singleton families arrive, per genome per generation.

Selection: The insertion of a novel promoter sequence is unlikely to be selectively neutral. However, it is unclear to what extent the available PMP data set, from fully sequenced prokaryote genomes, captures transient population polymorphisms, for example, as selection purges deleterious PMPs from the population. If these dynamics are captured in the data, the fitness of genomes carrying multiple families might be reduced relative to “uninfected” genomes.

We investigated this by allowing for the possibility that genomes carrying more PMP families might be more (or less) fit. In the model without selection, each entry in the genome-level transition matrix, M_{ij} , gives the rate at which genomes with j families create genomes with i families in the next generation. To include selection, we replace each of these entries by $w_j M_{ij}$, where w_j is the fitness of a genome carrying j families. For these fitness coefficients, we tested three cases: (1) exponential reductions in fitness (w_j is proportional to e^{-sj}), (2) linear reductions (w_j is proportional to $1 - sj$, or 0 if $1 - sj$ is negative), or (3) a uniform reduction if PMPs are present ($w_j = 1$ when $j = 0$, otherwise $w_j = 1 - s$). Despite our expectation that fitness would be reduced by

PMPs, in all three cases s was a free parameter that could be positive or negative.

As described further in the *Results*, we found no evidence that fitness varies with the number of PMP families in the genome and thus did not further investigate the possibility that fitness varies with number of copies per family.

Data fitting

We performed rigorous model identification and model reduction to determine which of the processes described in the models above were statistically justified. We first determined the best-fit orders of the various processes and then eliminated processes one by one to arrive at the simplest model that was consistent with the observed data.

To perform model selection, we first considered the copy-level model and obtained best fits for all possible combinations of the four exponents α_u , α_v , α_w , and α_n either 0 or 1. We then fixed the copy-level exponents and compared data fits of the model at both levels simultaneously for combinations of α_p and α_q either 0 or 1. For a given set of exponents (the α -values), we used a Nelder–Mead simplex routine (*fminsearch* in Matlab) to search for parameters that maximize the log-likelihood.

Having established the best-fit values of these exponents with all model parameters freely varying, we next evaluated which free parameters were justified, evaluating nested models using the Akaike information criterion (AIC). Note that when a nested model excluded any of p , q , or μ , we evaluated the effect of either setting the individual parameter to zero or constraining the parameter according to the values predicted in Equations 3, 4, and 5. For any nested model with AIC values lower than those of the full model, we continued evaluating nested submodels. In comparing a proposed model to the current “best” model yielding $AIC_{\min}$, the value $R = \exp[(AIC_{\min} - AIC)/2]$ gives the relative probability that the given model minimizes the information loss in model fitting, i.e., is a better model with which to explain the data.

Sensitivity

To evaluate the sensitivity of these parameter estimates, we assumed that the observed counts in the data corresponded to the “true” distributions and evaluated the extent to which the parameter estimates would change under different random samplings of that distribution. We first computed C_n as the observed proportion of families with n copies and likewise F_m as the observed proportion of genomes with m families. We then drew 1360 random samples from F_m (1360 genomes with random numbers of families per genome, distributed according to F_m) and 4074 random samples from C_n (4074 families with copies per family distributed according to C_n). This set of samples was then used as data in our model-fitting algorithm and new parameter estimates were obtained. We repeated this 100 times to estimate the coefficient of variation of each of our parameter estimates

and also to check for covariance among estimates, an indication of nonindependence among parameters.

Results

An expanded reservoir of putative mobile promoters

In our previous study, only the PMPs located directly upstream of genes were mapped. By querying the same set of 1360 prokaryote genomes with a representative of each PMP family, searching the whole genome in each case, we substantially expanded the total number of PMPs: from 13,111 to 39,441 sequences (supporting information, Table S1). In marked contrast, the randomized queries of the same genomes resulted in only 43 hits, all of which were associated with two PMP sequences that contained poly(T) runs of >50 nucleotides. The observed increase in the number of PMPs identified in the whole genome search is strong evidence for the putative mobility of these sequences.

The bars in Figure 4 and Figure 5 show the distributions of PMP families per genome, and copies per family, in the full data set. While in the original analysis, 2771 of the 4074 families of PMPs (68%) consisted of families containing only two sequences, expanding the analysis to the entire genome reduced the number of families consisting of two elements to 1129 (28%). Strikingly, the mean number of copies per family increased from 3.2 to 9.7 in the full data set, while the mean number of families per genome remained unchanged at 3.0. Thus on average, each prokaryote genome in our data set carries three families with about 10 copies of the promoter in each family.

Although the number of PMPs identified was tripled by searching outside of the original narrow window upstream of the translational start sites, the distribution of PMPs across the genome was highly nonrandom. In particular, of the 39,441 PMPs identified in the full genome search, 35,295 were located intergenically; only 10.5% (4146) overlapped with coding sequences. Since coding sequences make up ~90% of prokaryote genomes (Mira *et al.* 2001), this suggests a strong bias against these regions for PMP insertion sites.

Model selection and data fitting

Using this full data set, significantly better fits to the data were obtained when the exponents were set to $\alpha_u = \alpha_v = \alpha_w = 1$, while $\alpha_\eta = 0$; other combinations of exponents were $<10^{-6}$ times as likely as this model to best explain the data. These exponent values imply that duplication, excision, and diversification rates depend, as we might expect, on the number of copies in the family. In contrast, since $\alpha_\eta = 0$, each PMP family generates a HGT event with roughly the same probability, irrespective of the number of copies in the family. Taking these values for the copy-level exponents, the best fit of the full model was obtained with $\alpha_p = \alpha_q = 1$. Once again, alternative exponent combinations yielded models that were $<10^{-10}$ times as likely as this model to

best explain the data. This implies that the overall gain and loss of PMP families depends on the number of families already present in the genome.

Using these exponents, we next evaluated the three models that included selection. While none of these models provided a substantially improved fit to the data, the best of the three yielded a relative probability R of 0.092. This value was obtained for exponentially decreasing fitness, with $s = 5.35 \times 10^{-5}$ per PMP family.

If PMPs are selectively deleterious, the overall number of PMPs carried by a genome might be limited by selective pressures. To address this, we also investigated whether genomes that carry larger numbers of PMP families supported fewer PMP copies per family. Results of this analysis are shown in Figure 2. It is clear from these data that the number of copies per family is not reduced in genomes with large numbers of families, but instead appears to increase to a plateau of ~10 copies per family. We conclude that the fitness effect of PMP families is not essential to fit the data, and we address the implications of these results in the *Discussion*.

We next evaluated all possible four-, three- and two-parameter models nested in the full five-parameter model without selection. Although in total a large number of models were tested during model identification, the majority were easily eliminated from further consideration because their relative probability, R , was extremely low. Figure 3 illustrates R for models of two to five parameters, grouped by number of free parameters. While none of the two-parameter models provided good data fits, four higher-order models (solid symbols) yielded nearly identical AIC values. These values, along with the parameters that were included for each case, are shown in detail in Table 1; the inset to Figure 3 shows the corresponding relative probabilities. Although we cannot statistically identify a single “best” model, we are able to draw a number of robust conclusions that are shared by these models.

From these results, we conclude that in fitting the genome-level data, all three processes (diversification, loss, and HGT) are clearly required, and the data are best matched when these processes occur at similar rates. In particular, the ratios q/p and μ/p are 1.24 and ~1.1, respectively, in all of the best-fitting models; PMP families are lost at 1.25 times the rate they are gained via diversification and are gained by HGT (or *de novo* mobilization) at an intermediate rate.

At the copy level, we conclude that the duplication and loss of PMP copies occur at similar rates ($w/u \approx 1$) and that these rates are much larger than the diversification or HGT rates. These latter rates are so small that in the best-fitting four-parameter models, either diversification or horizontal gene transfer can be excluded from the model, and a good fit to the data may still be obtained.

However, the best three-parameter model was 0.995 times as likely as the best four-parameter model to minimize the information loss (Figure 3) and is thus not statistically different. In this model, the genome-level parameters p , q , and μ are set by the copy-level parameters according to Equations 3, 4, and 5, yielding $P = 6.26v$, $q = 0.336w$,

Figure 2 Copies per family vs. number of families in each genome. Each shaded dot plots the total number of PMP copies divided by the total number of PMP families in a single genome. Solid circles with error bars give the mean ± 1 SD. The number of copies per family is not reduced in genomes that carry many families.

and $\mu = 2.99\eta$. In this case, both diversification and HGT are required to simultaneously fit the copy- and genome-level data. We also note that in this model, the rates predicted for these processes (ν and η) are similar to the best-fitting parameter values in the unconstrained, five-parameter model.

The best fit of the three-parameter model (red line) to the data (bars) is illustrated in Figures 4 (genome level) and 5 (copy level). For reference, the genome-level data are identical to the data set shown in Matus-Garcia *et al.* (2012, Figure 4), while the number of promoters per family has been expanded as described in *Methods*. The distribution of promoters per family has a very long tail, which we include as a bin for families with >35 promoters. Note that the smallest number of promoters per family in the data set is 2; the inset on the right shows the predicted distribution of promoters per family including singletons, c_n .

Sensitivity

We performed sensitivity analysis using the best-fitting three-parameter model described above. The parameter estimates were robust to sampling noise; the sensitivity analysis yielded coefficients of variation (standard deviation/mean) of 3.9% for ν/u , 0.07% for w/u , and 5.9% for η/u . This high reproducibility is likely due to two factors: the large data set considered (~ 1000 genomes, 4000 families, and 40,000 copies), combined with the small number of parameters estimated in the reduced model. No significant covariance between the parameter estimates was observed ($r^2 < 5\%$ in all cases).

Mapping singleton PMPs

The model predicts a large fraction of singleton PMP homologs, and we were able to identify these by expanding

Figure 3 Results of model selection. The relative probability, $R = \exp[(AIC_{\min} - AIC)/2]$ is plotted against the number of parameters in the model, where $AIC_{\min}$ is the minimum AIC value achieved for any tested model. Shaded symbols correspond to models with low values of R , which were eliminated from consideration; results are not shown for several models with $R < 10^{-60}$. The square corresponds to the three-parameter model whose fit is illustrated in Figure 4 and Figure 5. The inset shows the same data on a linear y-axis.

our search as described in *Methods*. As a query we used the nonredundant data set of 3216 sequences from our previous study and mapped all instances of singleton hits meeting the criteria of at least 80% nucleotide identity over 50 bp (Table S2). For 987 (31%) of these nonredundant queries, we found a singleton hit in at least one genome. In total, the expanded search identified 6348 singletons in 1284 distinct genomes. In contrast, the control data set using permuted versions of the same query sequences resulted in 11 hits in total, across all genomes.

Discussion

The expanded search for mobile promoter sequences increased the number of PMP copies in our data set from $\sim 13,000$ when searching directly upstream of translation start sites to 40,000 when whole genomes were queried. Only 10% of these copies overlapped coding sequences, despite the fact that coding sequences make up $\sim 90\%$ of prokaryote genomes. Thus, on average the PMPs identified in the whole genome search show a very strong bias toward noncoding regions of the genome, similar to RNA-polymerase binding sites (Froula and Francino 2007). Although the focus of our expanded search was to enumerate the instances of PMPs in sequenced genomes, the very large number of PMP sequences uncovered in the whole genome search raises a number of intriguing questions. A clear avenue for future work is a detailed exploration of these newly discovered PMP copies, investigating their distribution across the genome and identifying patterns in their insertion sites, such

Table 1 Parameters and AIC values for the best-fitting models

No. of parameters	Copy level			Genome level		AIC
	v/u	w/u	η/u	q/p	μ/p	
5	0.0493	0.965	0.0690	1.24	1.07	28751.56
4	–	1.01	0.393	1.24	1.07	28749.15
4	0.0598	0.954	–	1.24	1.07	28749.15
3	0.0424	0.981	0.0965	(1.24)	(1.09)	28749.16

Parameters in parentheses were set by Equations 3, 4, and 5.

as the marked pairings identified for REPINS (Bertels and Rainey 2011).

As seen in the inset in Figure 5, a simple prediction of our model is the existence of a large class of mobile promoter singletons. These singletons are “families” of one copy that have recently diversified, recently arrived by HGT, or, in the case of PMPs arising from IS elements, recently lost their associated coding sequences. The model predicts that at any given time, 33% of all mobile promoter families have only a single copy. By extending our search to include promoters identified in other genomes, we found that 24% of intergenomic PMP families existed as singletons in at least one genome. This number is an underestimate because we can identify singletons only if homologous multicopy families occur in another genome.

Surprisingly, selection was not necessary to describe the distribution of PMP families in this data set. In addition, we saw no evidence that the number of PMPs that could be sustained in a genome was limited; the number of copies per family did not decrease for genomes with many families. These results strongly suggest that the available PMP data set does not adequately capture transient polymorphisms, but rather reflects a snapshot of PMPs that largely exist at high frequencies in the relevant populations. Sequenced genomes from a large number of individuals of the same lineage, such as experimental evolution data, could allow us to identify the time course of polymorphisms and definitively test for selection.

The best-fit parameter values predict that the rates of gain and loss of PMP copies in prokaryote genomes are roughly equal, while the rate of HGT, per family, is an order of magnitude lower. The rate of diversification, in which a PMP family diversifies in sequence to create a nonhomologous family (for example, by deletions), is smaller yet, about half the HGT rate. The low diversification frequency predicted by our model suggests that PMP sequences may be tightly constrained; that is, a large fraction of their 50- to 100-nucleotide sequence may be essential to their function. Nonetheless, models that excluded diversification were not adequate to fit the data. An example of possible diversification within a promoter family is shown in Figure S1.

Because we recover only the ratios of these rates, we are unable at this stage to make predictions involving time-scales, such as the predicted time to reach equilibrium or

short-term dynamics for a PMP family. In the long term, the rates we estimate predict the persistence of PMPs, with the majority of prokaryote genomes carrying at least one family, but most families having only one or two copies; these predictions are self-evident given that the model fits the data well. The persistence of PMPs is maintained by a balance of gain, through duplication and HGT, and loss, through deletion/excision and diversification. Of these four processes, deletion w and duplication u play the largest roles, and at the copy level the deletion rate is slightly less than the duplication rate. In the best-fitting models, however, the sum of the rates contributing to loss ($w + v$) is a few percent larger than the duplication rate. This suggests that horizontal gene transfer, which adds further copies to the genome at a low rate, is important in maintaining PMPs, as has been found for insertion sequences (Hartl and Sawyer 1988).

For insertion sequences, the rate of HGT, per element per generation, has been estimated to be as low as 10^{-8} for transduction (Jiang and Paul, 1998), in the range 10^{-6} – 10^{-3} for transformation (Williams *et al.* 1996) and 10^{-6} – 10^{-4} for conjugation (Dahlberg *et al.* 1998) (see Bichsel *et al.* 2013, Table 1, for an overview). Although PMPs are much shorter sequences than ISs, if we take 10^{-6} as a conservative estimate of a possible HGT rate for PMPs, this would imply that the duplication and loss rates for PMPs would be of order 10^{-5} . These values would be consistent with recent direct estimates of IS replicative transposition rates of 10^{-5} for six transposable elements in *Escherichia coli* (Sousa *et al.* 2013). Sousa *et al.* also found that for at least two of these ISs (IS2 and IS5), the excision rate was almost as large as the duplication rate. Our results are thus consistent with the hypothesis that transposable elements acting in *trans* may play a role in PMP mobility.

Despite their relatively low diversification rate, the dynamics of PMP families are predicted to be quite rapid, with PMP families being gained and lost by genomes almost as rapidly as individual PMP copies. In particular, the rate at which families generate new families, the rate at which new PMP families arrive by HGT, and the rate at which families are lost are all of the same order of magnitude as the per-copy duplication rate. In hindsight, it is perhaps not surprising that the dynamics of PMP families occur on a similar timescale as the dynamics of PMP copies, given

Figure 4 Results of model fitting: families per genome. Bars give the distribution of number of PMP families per genome, while the solid line is the best simultaneous fit of the three-parameter model to these data. Genome-level parameters p , q , and η are defined by Equations 3, 4, and 5 using copy-level parameter values $v/u = 0.0424$, $w/u = 0.981$, and $\mu/u = 0.0965$.

that about half of all PMP families exist as families of only one or two copies.

The birth–death–diversification model we develop is very similar in spirit to the model developed by Bichsel *et al.* (2013) to estimate the fitness effect of the insertion sequence IS5. The processes of PMP duplication, loss, and *de novo* appearance in the genome are, respectively, in direct analogy with replicative transposition, excision, and HGT as included by Bichsel *et al.* Interestingly, the exponential orders of these processes (the α -values in this article) are also consistent with the conclusions of Bichsel *et al.* (2013) for IS5. Despite this agreement, the two models include distinct effects and address very different questions. Our work addresses both copy number and the number of families per genome, and thus diversification is a key process in our model that is not relevant to the IS5 model. IS5 is an autonomous, 1200-nucleotide insertion sequence, whereas our data capture the dynamics of 50- to 100-nucleotide non-autonomous PMPs. The question of fitness is central to the IS5 work, and Bichsel *et al.* conclude that IS5 may be effectively neutral. We likewise allowed genome fitness to vary with family number, but found no evidence for selective effects.

In modeling the full PMP data set, we consider only average rates across all promoters, families, and genomes, which is clearly a large-scale simplification. For example, we expect that the behavior of promoters associated with essential or core genes might differ significantly from this bulk average. In addition, our sample of genomes is clearly nonrandom, with strong biases toward pathogenic microbes and experimental model organisms. Although repeated random samplings of phylogenetically disparate genomes could

Figure 5 Results of model fitting: copies per family. Bars give the distribution of the number of PMP copies per family. The solid line is the best simultaneous fit of the three-parameter model to these data. Values of the parameters are $v/u = 0.0424$, $w/u = 0.981$, and $\mu/u = 0.0965$. The inset shows the predicted proportion of families of each size, c_n , including families of one copy that are not measurable in the data set.

alleviate this bias (Bichsel *et al.* 2013), the data set for PMPs are sufficiently rich that a more detailed approach, incorporating phylogenetics, is another promising avenue for future work. Such an approach has recently been applied to the phylogenies of several transposable elements in two sequenced mosquito genomes (Struchiner *et al.* 2009). Anecdotal, a cursory investigation of the subset of our data from 29 *E. coli* genomes suggests substantially higher rates of HGT within the species, yielding a more compact distribution of families per genome (data not shown).

As well as the natural directions for future work mentioned above, we are in the process of analyzing sequence data from a mutation accumulation experiment, with which we hope to verify the mobility of at least some of the PMP sequences from *E. coli*. We are also analyzing in detail the sequences associated with each PMP family, expanding the previous search (Matus-Garcia *et al.* 2012) for signatures of mobility, such as palindromic cores and flanking inverted repeats. Our hope is that these efforts will further illuminate the structure and function of this intriguing new class of mobile genetic elements.

Acknowledgments

The authors are indebted to M. Matus-Garcia for data collection. L.M.W. is grateful to I. Gordo and A. M. Sousa at the Instituto Gulbenkian de Ciência, Portugal, for insightful comments. M.W.J.V. is funded by the Netherlands Organization for Scientific Research (NWO) via a VENI grant. H.N. is funded by the Netherlands Consortium for Systems Biology, which is part of the Netherlands Genomics Initiative/NWO.

L.M.W. is funded by National Sciences and Engineering Research Council Canada.

Literature Cited

- Baños, R., A. Vivero, S. Aznar, J. García, M. Pons *et al.*, 2009 Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genet.* 5(6): e1000513.
- Basten, C., and M. Moody, 1991 A branching-process model for the evolution of transposable elements incorporating selection. *J. Math. Biol.* 29(8): 743–761.
- Bertels, F., and P. Rainey, 2011 Within-genome evolution of REPINs: a new family of miniature mobile DNA in bacteria. *PLoS Genet.* 7(6): e1002132.
- Bichsel, M., A. Barbour, and A. Wagner, 2013 Estimating the fitness effect of an insertion sequence. *J. Math. Biol.* 66(1): 95–114.
- Blount, Z., J. Barrick, C. Davidson, and R. Lenski, 2012 Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489(7417): 513–518.
- Blumenstiel, J., 2011 Evolutionary dynamics of transposable elements in a small RNA world. *Trends Genet.* 27(1): 23–31.
- Brookfield, J., 1986a A model for DNA sequence evolution within transposable element families. *Genetics* 112: 393–407.
- Brookfield, J., 1986b The population biology of transposable elements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312(1154): 217–226.
- Brookfield, J., 2005 The ecology of the genome: mobile DNA elements and their hosts. *Nat. Rev. Genet.* 6(2): 128–136.
- Charlesworth, B., and D. Charlesworth, 1983 The population dynamics of transposable elements. *Genet. Res.* 42: 1–27.
- Charlesworth, B., and C. Langley, 1989 The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* 23: 251–287.
- Dahlberg, C., M. Bergström, and M. Hermansson, 1998 In situ detection of high levels of horizontal plasmid transfer in marine bacterial communities. *Appl. Environ. Microbiol.* 64(7): 2670–2675.
- De Gregorio, E., G. Silvestro, M. Petrillo, M. Carlomagno, and P. Di Nocera, 2005 Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J. Bacteriol.* 187(23): 7945–7954.
- Dolgin, E., and B. Charlesworth, 2006 The fate of transposable elements in asexual populations. *Genetics* 174: 817–827.
- Edwards, R., and J. Brookfield, 2003 Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Mol. Biol. Evol.* 20(1): 30–37.
- Froula, J., and M. Francino, 2007 Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS ONE* 2(8): e745.
- Golding, G., C. Aquadro, and C. Langley, 1986 Sequence evolution within populations under multiple types of mutation. *Proc. Natl. Acad. Sci. USA* 83(2): 427–431.
- Hartl, D., and S. Sawyer, 1988 Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? *Genetics* 118: 537–541.
- Hickey, D., 1982 Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101: 519–531.
- Hudson, R., and N. Kaplan, 1986 On the divergence of members of a transposable element family. *J. Math. Biol.* 24(2): 207–215.
- Jiang, S., and J. Paul, 1998 Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* 64(8): 2780–2787.
- Langley, C., J. Brookfield, and N. Kaplan, 1983 Transposable elements in mendelian populations. I. A theory. *Genetics* 104: 457–471.
- Le Rouzic, A., and G. Decelie, 2005 Models of the population genetics of transposable elements. *Genet. Res.* 85(3): 171–181.
- Le Rouzic, A., T. Boutin, and P. Capi, 2007 Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. USA* 104(49): 19375–19380.
- Matus-Garcia, M., H. Nijveen, and M. van Passel, 2012 Promoter propagation in prokaryotes. *Nucleic Acids Res.* 40(20): 10032–10040.
- Mira, A., H. Ochman, and N. Moran, 2001 Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17(10): 589–596.
- Moody, M., 1988 A branching process model for the evolution of transposable elements. *J. Math. Biol.* 26(3): 347–357.
- Nijveen, H., M. Matus-Garcia, and M. van Passel, 2012 Promoter reuse in prokaryotes. *Mobile Genet. Elements* 2(6): 279–281.
- Ochman, H., J. Lawrence, and E. Groisman, 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784): 299–304.
- Ohta, T., 1985 A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics* 110: 513–524.
- Perez, J., and E. Groisman, 2009a Evolution of transcriptional regulatory circuits in bacteria. *Cell* 138(2): 233–244.
- Perez, J., and E. Groisman, 2009b Transcription factor function and promoter architecture govern the evolution of bacterial regulons. *Proc. Natl. Acad. Sci. USA* 106(11): 4319–4324.
- Sawyer, S., and D. Hartl, 1986 Distribution of transposable elements in prokaryotes. *Theor. Popul. Biol.* 30(1): 1–16.
- Siddique, A., N. Buisine, and R. Chalmers, 2011 The transposon-like Correia elements encode numerous strong promoters and provide a potential new mechanism for phase variation in the meningococcus. *PLoS Genet.* 7(1): e1001277.
- Slatkin, M., 1985 Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics* 110: 145–158.
- Sousa, A., C. Bourgard, L. Wahl, and I. Gordo, 2013 Rates of transposition in *Escherichia coli*. *Biol. Lett.* 9(6): 20130838.
- Stavrinides, J., M. Kirzinger, F. Beasley, and D. Guttman, 2012 E622, a miniature, virulence-associated mobile element. *J. Bacteriol.* 194(2): 509–517.
- Struchiner, C., E. Massad, Z. Tu, and J. Ribeiro, 2009 The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes. *Evolution* 63(12): 3136–3146.
- Touchon, M., and E. Rocha, 2007 Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24(4): 969–981.
- van Passel, M., C. Smillie, and H. Ochman, 2007 Gene decay in Archaea. *Archaea* 2(2): 137–143.
- van Passel, M., P. Marri, and H. Ochman, 2008 The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLOS Comput. Biol.* 4(4): e1000059.
- Wagner, A., 2006 Cooperation is fleeting in the world of transposable elements. *PLOS Comput. Biol.* 2(12): e162.
- Wagner, A., 2009 Transposable elements as genomic diseases. *Mol. Biosyst.* 5(1): 32–35.
- Williams, H., M. Day, J. Fry, and G. Stewart, 1996 Natural transformation in river epilithon. *Appl. Environ. Microbiol.* 62(8): 2994–2998.

Communicating editor: J. Lawrence

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162883/-/DC1>

Birth, Death, and Diversification of Mobile Promoters in Prokaryotes

Mark W.J. van Passel, Harm Nijveen, and Lindi M. Wahl

```

30022430_prom GT - GGGGAT CAACT GTCCGT AAAAG CCCGATTGGT - TCAACT AAT AAT CCGT GGGGGGAT GAACAAAA - CCCCCACGGATTAAAGTTTCACTTTAT
30019068_prom GT GGGGGAT CAACT GTCCGT AAAAG GTCCGATTGGT GTGGGCT AAT AAT CAGT - GGGGGGAT GAACAAAA - CCCCCACTGATTAAAGTTTCACTTTAT
30020645_prom GT AAGAGAT CAATT GCCCGT AAAAG GTCGATTGGT - TCAACT AAT AAT CAGT - GTGGGAT GAACAAAA - CCCCCACTGATTAAAGTTTCACTTTAT
30022640_prom GT GGAGAAT CAACT GCCCGT AAAAG CCCGATTGGT GAGGGCT GAT AAT CAGT - GGGGGGAT GAACC - - - TCCCCAATGATTAAAGTTTCACTTTAT
30022440_prom ACGGAGAT CCACT GCCCGT AAACGTCGATTGGTAAAGGCT AAT AAT CAGT - - GGGGGT GAACAAAAAT CCCCCACTGATTAAAGTTTCACTTTAT
30018309_prom GTTAGATTT GGGCT GCTCGT AAAAG CCCAATTAACTGTGGGCT AAT AATTAGT - - GGGGAGGAACAAAA - ACCCTACTGATTAAAGTTTCACTTTAT

```

Figure U1 shows an example of a diversification pattern within a promoter family found in *Bacillus cereus*. While the nucleotides at many positions are conserved in all members of the family, nucleotide changes and indels have occurred at several more variable sites. Based on nucleotide identity, the last sequence in the alignment (30018309_prom) is the most likely to diverge away from this family as it accumulates more mutations. Multiple alignment created using Muscle (Edgar, 2004), visualized with Jalview (Waterhouse et al. 2009).

Literature Cited

- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5): 1792-1797.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9): 1189-1191.

Tables S1-S2

Available for download at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.162883/-/DC1>

Table S1 Full genome dataset

Table S2 Singleton dataset