A neutral theory of genome evolution and the frequency distribution of genes

Bart Haegeman, Joshua Weitz

BMC Genomics

RESEARCH ARTICLE Open Access

A neutral theory of genome evolution and the frequency distribution of genes

Bart Haegeman[1*†] and Joshua S Weitz[2,3*†]

Abstract

Background: The gene composition of bacteria of the same species can differ significantly between isolates. Variability in gene composition can be summarized in terms of gene frequency distributions, in which individual genes are ranked according to the frequency of genomes in which they appear. Empirical gene frequency distributions possess a U-shape, such that there are many rare genes, some genes of intermediate occurrence, and many common genes. It would seem that U-shaped gene frequency distributions can be used to infer the essentiality and/or importance of a gene to a species. Here, we ask: can U-shaped gene frequency distributions, instead, arise generically via neutral processes of genome evolution?

Results: We introduce a neutral model of genome evolution which combines birth-death processes at the organismal level with gene uptake and loss at the genomic level. This model predicts that gene frequency distributions possess a characteristic U-shape even in the absence of selective forces driving genome and population structure. We compare the model predictions to empirical gene frequency distributions from 6 multiply sequenced species of bacterial pathogens. We fit the model with constant population size to data, matching U-shape distributions albeit without matching all quantitative features of the distribution. We find stronger model fits in the case where we consider exponentially growing populations. We also show that two alternative models which contain a "rigid" and "flexible" core component of genomes provide strong fits to gene frequency distributions.

Conclusions: The analysis of neutral models of genome evolution suggests that U-shaped gene frequency distributions provide less information than previously suggested regarding gene essentiality. We discuss the need for additional theory and genomic level information to disentangle the roles of evolutionary mechanisms operating within and amongst individuals in driving the dynamics of gene distributions.

Keywords: Bacteria, Neutral model, Pan-genome, Population genomics, Selection

Background

The gene content of genomes of closely related bacteria can differ significantly. For example, pair-wise comparisons of genome sequences from isolates of the same species often do not share a substantial fraction of their gene content [1-10]. When a large number of genomes within a species or closely related group of bacteria are sequenced, the gene content variability can be summarized as a gene frequency distribution: given G sequenced genomes, some genes are found in a fraction $1 \leq k \leq G$ of all genomes. Empirically, such gene frequency distributions possess a characteristic U-shape, such that there are many genes which only appear in one genome, fewer genes which appear in $2 \leq k \leq G - 1$ genomes, and many genes which appear in all genomes. Genes within each of these three categories have been labeled accessory, character and core genes, respectively [11]. It is tempting to conflate gene frequency with relative essentiality, but is it valid? For example, is it necessarily true that a gene that appears in all genomes in a sample should be classified as a "core" gene? Could such a gene have become common through neutral processes that diminish variability, and occasionally, lead to fixation of types? Likewise, should a gene which appears in only one genome in a sample be considered as "accessory" to the function of that organism? Could such a gene

* Correspondence: bart.haegeman@inria.fr; jsweitz@gatech.edu
† Contributed equally
[1]INRIA Research Team MODEMIC, UMR MISTEA, 34060 Montpellier, France
[2]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA
Full list of author information is available at the end of the article

have become rare via a neutral path toward extinction, or have been recently introduced to a lineage without significant effect on individual fitness?

Here, we argue that a suitable null model is necessary with which to deduce how much weight be given to gene frequency data as a means to generate hypotheses regarding essentiality. For example, a recently proposed neutral theory of biogeography and biodiversity has proved useful in clarifying when and how species abundance distributions in ecology can provide information about selective processes in complex communities [12-15]. Hence, in this manuscript we ask: is it possible to recapitulate findings of U-shaped gene frequency distributions in the absence of selective forces driving genomic and population composition? We answer this question in the affirmative by proposing a simple and analytically tractable neutral model of genome evolution that explicitly accounts for gene composition of genomes. In this model, genomes undergo birth-death processes in a neutral sense and also acquire and lose genes which we term "gene transfer". The model differs from most previous efforts to analyze genome evolution [16,17] by self-consistently treating the dynamics at two scales: population level drift and genomic level change (for an exception, see [18] whose model we address in the Results and Discussion). Analysis of the current model leads to the following major results.

First, we find that gene frequency distributions derived from this model possess a characteristic U-shape for a robust range of model parameters. Hence, we propose that prevalence of a gene does not necessarily imply its essentiality, and that gene frequency distributions may be more limited than previously acknowledged in generating inferences regarding essentiality. Second, we estimate the best fit parameters for a given empirical gene-frequency distribution of sequenced genomes and in so doing find a reasonable correspondence between our neutral model and data from six distinct bacterial species with sequenced genomes from multiple isolates. However, our model assuming constant population sizes predicts gene frequency distributions with systematically fewer rare genes than the empirical distributions. Hence, we show that assuming other types of population dynamics (such as exponentially growing populations) can change the model predictions in line with empirical data, providing a basis for investigating the role of population dynamics in shaping gene frequency distributions. Further, we extend the model to include a rigid and flexible core in the genome, and show that other assumptions about genome structure are consistent with gene frequency data. Finally, we show that a recently proposed gene diversity index - genomic fluidity [19] - is a natural parameter emerging in the neutral models of genome evolution described here. Whereas previously this parameter was entirely statistical in nature, we discuss here how genomic fluidity can be seen as a proxy for the relative importance of gene uptake in shaping the gene composition of genomes between species. In so doing, we discuss how other observations could be combined with gene frequency distributions to improve inferences regarding evolutionary mechanisms shaping genome composition.

Results and discussion

A neutral model of genome evolution combines birth-death events with gene transfer events

We propose the following neutral model of genome evolution, see Figure 1. Consider a population consisting of N organisms in which each organism has a genome consisting of M unique genes. The dynamics consist of a sequence of reproduction and gene transfer events. In a reproduction event, one of the N organisms (chosen at random) dies, and is replaced by offspring of one of the other organisms (chosen at random). The offspring genome is identical to the parent genome. Note that there are still N organisms after the event and hence, this step is equivalent to a birth-death event of the Moran model [20,21]. In a gene transfer event, one of the N organisms acquires a gene from the environment. We assume that this gene is new, i.e., a gene that has not been present in the population before, and hence, this step is comparable to a mutation event in the infinitely many alleles model of population genetics [20,21]. In the model, gene transfer events do not affect birth and death rates of the individual (e.g. [22] present evidence for the neutrality or near-neutrality of transferred genes). We also assume that the acquisition of the new gene induces the loss of another gene in the genome, so that the organism's genome still consists of M genes after the event. We utilize a constant value of M to facilitate mathematical analysis and note that bioinformatic based estimates of total gene counts vary approximately 10% between genomes of the same species, as considered here.

Individual birth-death events cause genetic drift in the population as the number of organisms having a particular gene fluctuates over time. Genetic drift has the tendency to reduce the genetic diversity in the population. Indeed, when the last organism carrying a particular gene dies, this gene disappears from the population, and has no opportunity of re-entering the population. Gene birth-death events, on the other hand, maintain the genetic diversity in the population. New genes enter at low frequency due to "gene transfer" events. These transfer events may be due to uptake of genes from the environment, insertion of genes via viruses, or conjugation with other individuals. In our model, individual birth-death events and gene transfer events have associated rate parameters: we denote by ρ the rate of reproduction per individual, and by σ the rate of gene

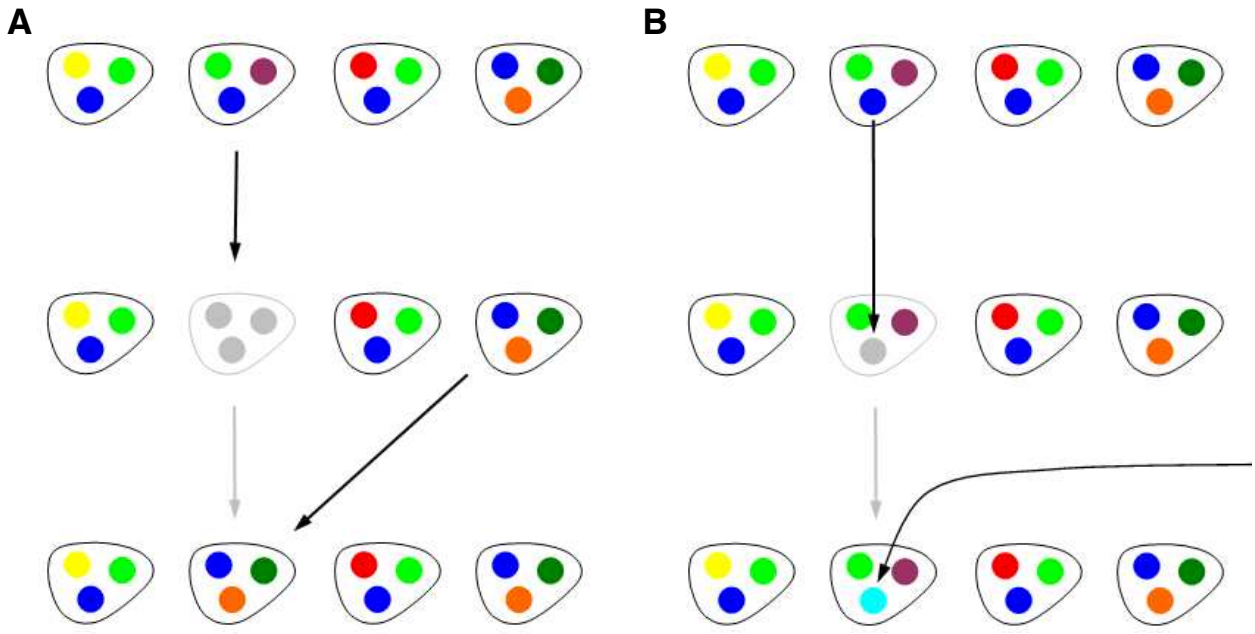

Figure 1 Graphical illustration of a neutral model of genome evolution. A population of $N = 4$ organisms is shown, with genomes consisting of $M = 3$ genes. Colors denote different gene identities. We chose small values for N and M for illustrative purposes; realistic values are, e.g., $N \sim 10^8$ and $M \sim 2000$. (A) In a birth-death event an organism dies and is replaced by offspring of another organism. The offspring genome is identical to the parent genome. (B) In a gene transfer event a gene of one of the organisms is replaced by a gene from the environment. We assume that this new gene has not been present in the population before.

transfer per individual. Intuitively we expect that variability in gene composition of genomes should increase when gene transfer rate σ increases relative to the individual reproduction rate ρ, and vice-versa. Further, we expect that gene frequency distributions will tend toward U-shaped distributions because of the tension between gene transfer (which would favor increasing rarity of genes) and individual birth-death (which would favor increasing commonness of genes, due to neutral drift). We evaluate this prediction of U-shaped distributions in the following section.

Neutral model of genome evolution predicts U-shaped gene frequency distributions

The distribution of genes over genomes can be characterized in detail for the neutral model of genome evolution described above. For example, the gene frequency distribution can be computed explicitly, see Additional file 1: Appendix S1. To describe the solution, we consider a sample of G genomes taken from the population, and we denote the average number of genes appearing in k of the G genomes by g_k. The gene frequencies predicted by the neutral model of genome evolution are

$$g_k = \frac{M\theta}{k}\frac{G!}{(G-k)!}\frac{(\theta)_{G-k}}{(\theta)_G}, \qquad (1)$$

with

$$\theta = \frac{N\sigma}{M\rho} \quad \text{and} \quad (\theta)_k = \theta(\theta+1)\dots(\theta+k-1),$$

where θ is an effective gene transfer rate. The distribution in Eq.(1) appears in solutions to allele distributions in the infinitely many alleles model of population genetics [20,21].

As the number of genes M in a genome and the number of genomes G in the sample are given by the data, the gene frequencies g_k are parametrized by the dimensionless parameter θ which combines the effects of both gene transfer and birth-death processes. Hence, different combinations of N, ρ and σ lead to identical gene frequency distributions. In particular, the predicted distribution is insensitive to accelerating simultaneously gene transfer and reproduction, because the distribution depends on the ratio $\frac{\sigma}{\rho}$, and not on ρ and σ individually. Moreover, an increase of the ratio $\frac{\sigma}{\rho}$, the relative rate of gene transfer, can be compensated by a smaller population size N, increasing the intensity of genetic drift. Note that although in practice the sample size G is much smaller than the population size N, Eq. (1) is also valid for $G = N$, i.e., it can be used to compute the

(empirically inaccessible) gene frequency distribution of the entire population.

We plot Eq. (1) for cases where $\theta = 0.03$, $\theta = 0.3$ and $\theta = 3$ in Figure 2. As anticipated, the weight of the gene frequency distribution shifts from the common genes for small values of θ (left panel) to the rare genes for larger values of θ (right panel). The gene frequency distributions have a U-shape for a robust range of parameters. The U-shape is generic so long as $\theta < 1$; for $\theta > 1$ the distribution is monotonically decreasing. As shown in Additional file 1: Figure S1, the gene frequency distribution changes when sampling more genomes, but the characteristic U-shape remains. These observations for the neutral model of genome evolution show that U-shaped frequency distributions do not require invoking selection at the genome level. Further, the observations suggest that findings of prevalent genes need not be an indicator of essentiality in the absence of other information about gene function.

Comparing empirical gene frequency distributions of multiply sequenced bacterial species to model predictions

We collect and analyze empirical gene frequency distributions from 6 species of bacterial pathogens: *B. anthracis, E. coli, Staph. aureus, Strep. pneumoniae, Strep. pyogenes* and *N. meningitidis*. Gene frequency distributions were compiled by applying an automated genomic pipeline to remove the impact of curation bias and to normalize comparisons between species [23]. Hence, what we compile are frequency distributions of clusters of homologous genes (for details, see [19,23]), which we denote, for simplicity, as "genes" in this manuscript. We find that the empirical gene frequency distributions have a characteristic U-shape in that there are many genes which only appear in a single genome, many genes which appear in all genomes, and fewer genes that appear in an intermediate number of genomes (see Figure 3). This characteristic U-shape is robust to reasonable changes in the values of identity and coverage utilized for comparing genes in our genomic pipeline. Notice that the gene frequency distribution is on a log scale, hence the U-shape is in fact highly pronounced, in that there may be 50 times as many genes that appear in all genomes than appear in half the genomes.

The neutral model of genome evolution can be fit to data using Eq. (1). To do so, we determine the parameter θ that minimizes the distance between the predicted and empirical gene frequency distribution (see Materials and Methods). We find that the neutral model is in reasonable correspondence with data (see Figure 3). First, both data and predictions have a U-shape. Next, model predictions agree well with observations for the total number of core genes. These predictions are made with a single free parameter, θ. On the other hand, the model underpredicts the number of rare genes and overpredicts the number of genes in the intermediate part of the distribution. In particular, the model predicts a $g_k \sim 1/k$ dependence for the peak at small k, whereas the data are closer to a steeper $g_k \sim 1/k^2$ dependence. In the next section we show that this deviation can be partially remedied by dropping the assumption of a constant population size. Finally, it is important to note that although our model assumes constant M, we find that there is approximately a 10% difference in total gene content within the genomes of each of the six species.

Predictions for the observed pan and core genome size in a sample of genomes can also be obtained. In the past, the pan genome size has been defined as the number of genes in all genomes of the population. Similarly, the core genome size has been defined as the number of genes found in every genome in the population. However, we and colleagues have previously shown that estimating pan and core genome sizes are unreliable because they

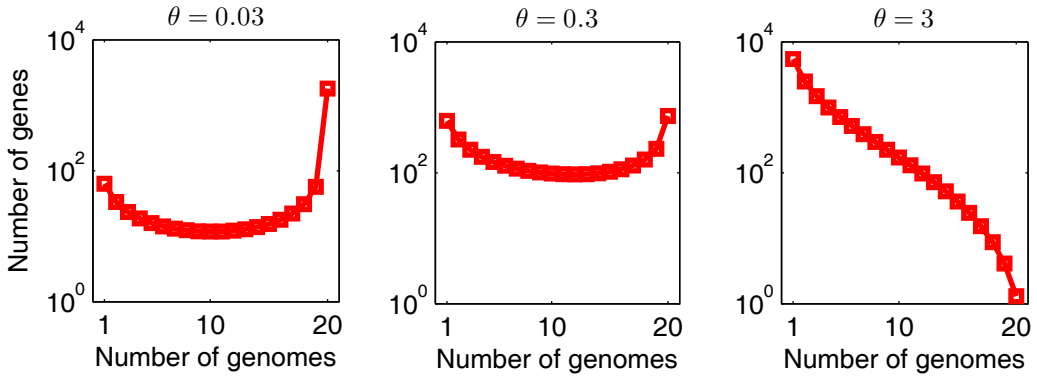

Figure 2 Gene frequency distributions for neutral model of genome evolution (model A). Genome size $M = 2000$ and sample size $G = 20$. Gene transfer parameter θ: in left panel, $\theta = 0.03$; in middle panel, $\theta = 0.3$; in right panel: $\theta = 3$

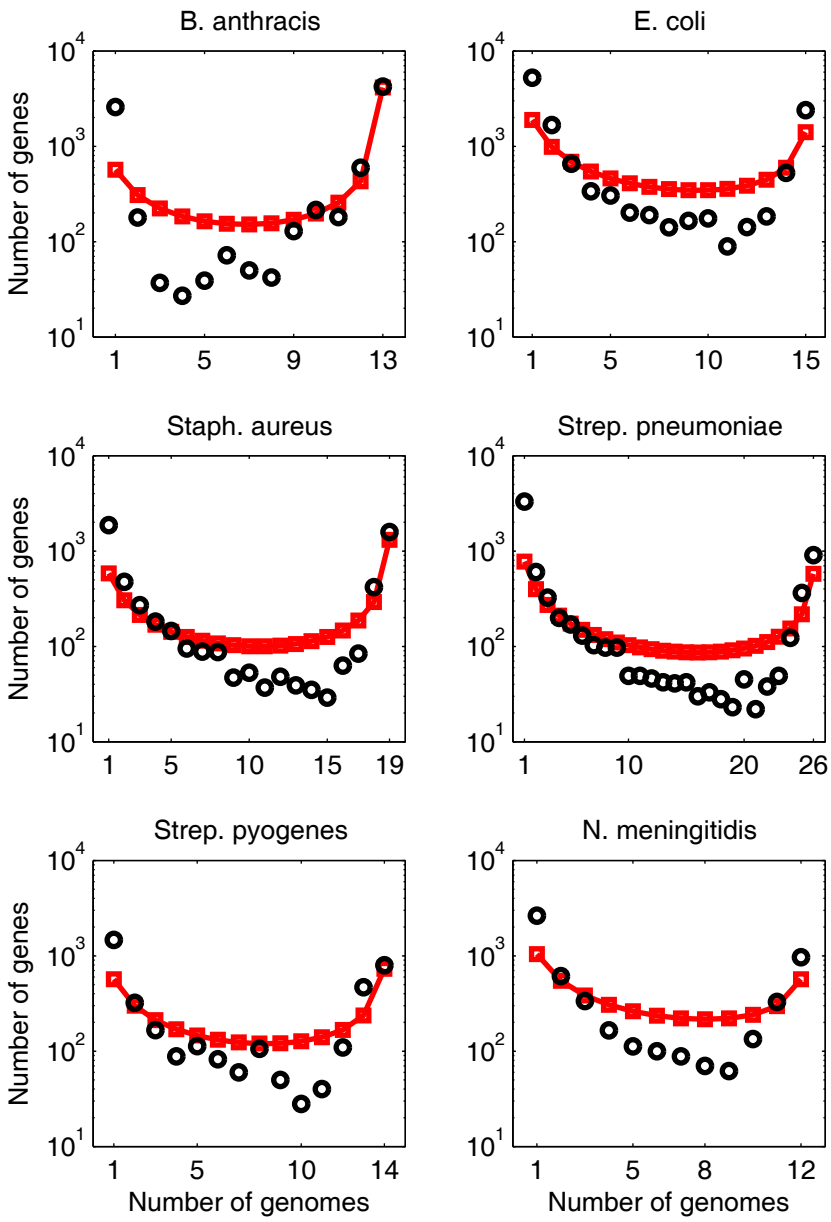

Figure 3 Data comparison for neutral model of genome evolution (model A). Comparison of gene frequency distributions with predictions of the simplest model: the population size is assumed to be constant and all genes are governed by the same gene transfer process. The model has one parameter, the gene transfer parameter θ. Black circles: data; red line with squares: model

depend on observations of rare genes and genomes, respectively, which are difficult to find in samples precisely because they are rare [19]. Here, we define the observed pan and core genome size as the number of unique genes found in all sample genomes and the number of genes common to all sample genomes, respectively. As we derive in Additional file 1: Appendix S1, the model predicts that the pan genome size increases logarithmically with sample size, and that the core genome size decreases as a power law (with exponent $-\theta$). The prediction that gene diversity grows without bound is unsurprising, because we assumed an infinite gene pool (and we caution that gene diversity cannot increase to infinity in reality). Nevertheless, we expect the logarithmic (power-law) dependence of pan (core) genome size also to hold for a finite gene pool (as long as the gene pool is much larger than the set of genes observed in the sample). These findings are corroborated by the data, which exhibit the same qualitative behavior, see Additional file 1: Figure S2. Notice that we used the value for

σ obtained from Figure 3 in the pan and core genome fits of Additional file 1: Figure S2, so that these fits have no additional free parameters.

Estimating gene transfer parameters from model fits

We can utilize parameters estimated from gene frequency distribution fits to also estimate underlying mechanistic parameters driving genome evolution, albeit with caveats that we discuss below. Note that the estimated values for θ are in the range $0.1 < \theta < 0.5$ (see Table 1). For example, with $M \approx 2000$, then the product $N\sigma/\rho$ should be on the order of 10^3. Conventional estimates of effective population size are approximately 10^7 - 10^9 [24,25], suggesting that gene uptake is on the order of $\sigma/\rho \approx 10^{-4}$ - 10^{-6}, approximately once per tens of thousands or million divisions. Assuming bacteria that divide once per hour, then $\sigma \approx 10^{-4}$ - 10^{-6} hr^{-1}. In this model, σ represents gene transfer. Here, we consider one mechanism for such gene uptake - natural uptake of DNA from the environment - and evaluate whether or not empirical parameters associated with transformation are consistent with inferred estimates of σ. Natural transformation rates vary widely depending on strain type, sequence homology, and physiological conditions. Empirically estimated DNA uptake rates are generally reported as transformation frequencies, ϵ, defined as the proportion of colony forming units (CFUs) that have taken up a segment of DNA of interest at the end of some experimental time period, T. We developed a simple model that estimates σ directly from ϵ and T (see Materials and Methods). We find that application of this method yields estimates of gene uptake rate for pathogens in laboratory environments that bracket the value predicted from our model. For example, DNA damaged *Helicobacter pylori* cells exhibit $\epsilon \sim 10^{-4}$ - 10^{-8} in a $T = 2.5$ hr experiment [26] Hence, $\sigma \sim 10^{-4}$ - 10^{-9} hr^{-1}. Likewise, naturally competent *Neisseria gonorrhoeae* cells exhibit transformation frequencies $\epsilon \sim 10^{-3}$ of total cells after $T = 4$ hr, though values range from $\epsilon \sim 10^{-2}$ - 10^{-7} [27]. These values yield uptake rates of $\sigma \sim 10^{-4}$ hr^{-1} with a range of $\sigma \sim 10^{-3}$ - 10^{-8} hr^{-1}. In both cases, these estimates are consistent with our estimate of gene transfer rates in the multiply sequenced pathogens considered here. However, there remains substantial disagreement as to whether the lower, effective number derived from certain population genetic models or the much larger, census number is a better estimate of effective population size [25]. Hence, without species-specific information, we caution that direct estimates of either gene transfer rate or effective population size should be treated with skepticism, even if estimates of their combined effect is more robust. Moreover, as we show in the next section, the value of θ is sensitive to assumptions about population history. Indeed, there is no reason to expect that the population structure and selective effects of genes are as simple as assumed here, providing additional caution to overly rigid interpretations of estimates of either N or σ.

Population structure strongly impacts gene frequency distributions

The current neutral model of genome evolution assumes a fixed population size N. This is a common, but likely unrealistic, assumption as bacterial populations can undergo large and fast size fluctuations. In this model, the introduction of novel genes is decoupled from the history of population size or structure, so that we can select an arbitrary population size or structure and then superimpose the introduction of novel genes on top of the resulting history of individual births and deaths. To illustrate this point, we consider how an exponentially growing population affects the gene frequency distributions g_k. Specifically we denote the population size history as

$$N(t) = N_0 e^{\alpha(t-t_0)}, \tag{2}$$

Table 1 Overview of model fits

			fitting error				fluidity				
	G	M	Δ_A	Δ_B	Δ_C	Δ_D	ϕ^{obs}	φ_A^{pred}	φ_B^{pred}	φ_C^{pred}	φ_D^{pred}
B. anthracis	13	5523	80	21	78	13	0.08	0.09	0.08	0.09	0.08
E. coli	15	4576	98	58	47	2.6	0.25	0.30	0.25	0.29	0.25
Staph. aureus	19	2651	29	16	21	4.3	0.16	0.19	0.16	0.19	0.16
Strep. pneumonia	26	2095	42	21	30	4.3	0.23	0.32	0.24	0.30	0.23
Strep. pyogenes	14	1786	26	10	25	7.5	0.20	0.24	0.20	0.24	0.21
N. meningitidis	12	2080	53	26	31	2.4	0.28	0.33	0.28	0.32	0.28

Model A assumes a constant population size, and the same gene transfer process for all genes. Model B assumes an exponentially growing population size. Model C assumes that a part of the genome is shared by all genomes (a rigid core); the other part is subjected to the same gene transfer process as in model A. Model D assumes two parts in the genomes, governed by different gene transfer rates. We determined for the four models the parameters that minimize the distance Δ between the empirical and the theoretical gene frequency distribution (see Materials and Methods for the definition of Δ). For each of the 6 bacterial species analyzed, we report the number of analyzed genomes G, the genome size M (average number of genes per genome), the distance Δ for the model fits, the genomic fluidity ϕ^{obs} estimated on the data, and the fluidity ϕ^{pred} for the model fits. Recall that model A has one parameter, models B and C have two parameters, and model D has three parameters.

with α the population growth rate, t_0 the present time and $N_0 = N(t_0)$ the present population size. We use a coalescence approach [20,21] to compute the average gene frequencies g_k, see Additional file 1: Appendix S3. The solution for the average gene frequency distribution g_k depends on two dimensionless parameters θ_0 and β,

$$\theta_0 = \frac{N_0\sigma}{M\rho} \quad \text{and} \quad \beta = \frac{N_0\alpha}{2\rho}. \tag{3}$$

The parameter θ_0 is the same as θ for the constant population size model, except that the population size N is replaced by the present population size N_0; again we call θ_0 the gene transfer parameter. The parameter β is a rescaled version of the population growth rate α; we call it the population growth parameter. The constant population size model, which we denote by model A, is a one-parameter model; the variable population size model, which we denote by model B, is a two-parameter model. Hence, we expect a richer set of gene frequency distributions predicted by model B compared to model A.

Additional file 1: Figure S3 shows gene frequency distributions computed for different combinations of the parameters θ_0 and β. For small $\beta \leq 1$ the distributions closely resemble the distributions of the model A with constant population size ($\alpha = \beta = 0$, see Figure 2). When increasing the population growth parameter β, the U-shape becomes more pronounced. For example, the peak at small k has a power-law dependence $g_k \sim k^{-\gamma}$ with $\gamma = 1$ for small β and γ increasing for increasing β. The predicted distributions are often, apart from the peak for the core genes present in all genomes, almost symmetric (see panels with $\theta_0 = 0.03$ or $\theta_0 = 0.3$). Notice that very similar distributions can be obtained for different parameter combinations (e.g., compare panel $\theta_0 = 0.03$, $\beta = 10$ and $\theta_0 = 0.3$, $\beta = 100$), which will affect the parameter estimation (see below).

We can fit empirical distributions to this neutral model of genome evolution (model B). To do so, we determine the parameters θ_0 and β that minimize the distance between the predicted and empirical gene frequency distribution. This computation yields estimates for the gene transfer parameter θ_0 and the population growth parameter β. As shown in Figure 4, the gene frequency distributions for model B fit the data better than those for model A (compare with Figure 3). The predictions of model B are uniformly accurate for the number of rare genes, the number of genes in the intermediate part of the distribution and the number of common genes. However, smaller but systematic deviations remain between observed and predicted gene frequency distributions. In particular, the empirical distributions (except for *B. anthracis*) are left-skewed, whereas the theoretical distributions have no skew or a small right skew. The improved fit is also apparent from the distance Δ between data and model, reported in Table 1, especially for *B. anthracis*. The estimate of the gene transfer parameter θ_0 is an order of magnitude larger in model B than the estimate of the gene transfer parameter θ in model A. However, the population size in model B is that of the present, whereas the population size in model A is the effective population size over the entire coalescent history. Hence, differences in gene transfer parameters are to be expected because they are driven in part by changes in assumptions about population size. This suggests that caution must be applied before utilizing θ, θ_0, or other dimensionless gene transfer parameters, to estimate an effective gene transfer rate without additional information that constrains the estimates of both population size and growth rate. For similar reasons caution should be applied to the interpretation of the estimates of the growth parameter β.

Models with an explicit core genome improve fit of empirical gene frequency distributions

The previous models assume that the gene transfer process affects all genes identically. Indeed, each gene present in a genome has the same chance to be replaced by a gene transfer event, and this replacement has no effect on the reproduction rate. Here we show how to relax this assumption without prohibitively increasing the model complexity. To illustrate this, we study a new model, in which we distinguish two parts in the genomes: one part is governed by the same gene transfer process as model A; the other part does not undergo gene transfer and hence, constitutes a rigid core genome. We term this model C and note that a similar model has also been proposed in the context of the analysis of *Prochlorococcus* genomes [18]. We assume that this rigid core has the same composition for all genomes. One interpretation of the rigid core is that genes in this core are essential, and deletion of any of a subset of genes in the core would be lethal to the individual. The average gene frequency distribution is given by Eq. (1), with an additional contribution for the rigid core, see Additional file 1: Appendix S5. This distribution depends on two parameters: the fraction λ_1 of the fluid part of the genomes, corresponding to $\lambda_1 M$ genes per genome, and the gene transfer parameter θ_1 for this part. The rigid core then represents a fraction $\lambda_2 = 1 - \lambda_1$, or $\lambda_2 M$ genes.

We can determine the parameters for which model C fits best the empirical gene frequency distributions. The model fits, shown in Figure 5 (yellow line), are better than those for model A (Figure 3), but are worse than those for model B (Figure 4), see also Table 1. Note that the fitting error Δ_B and Δ_C of models B and C can be compared directly, because both models have two

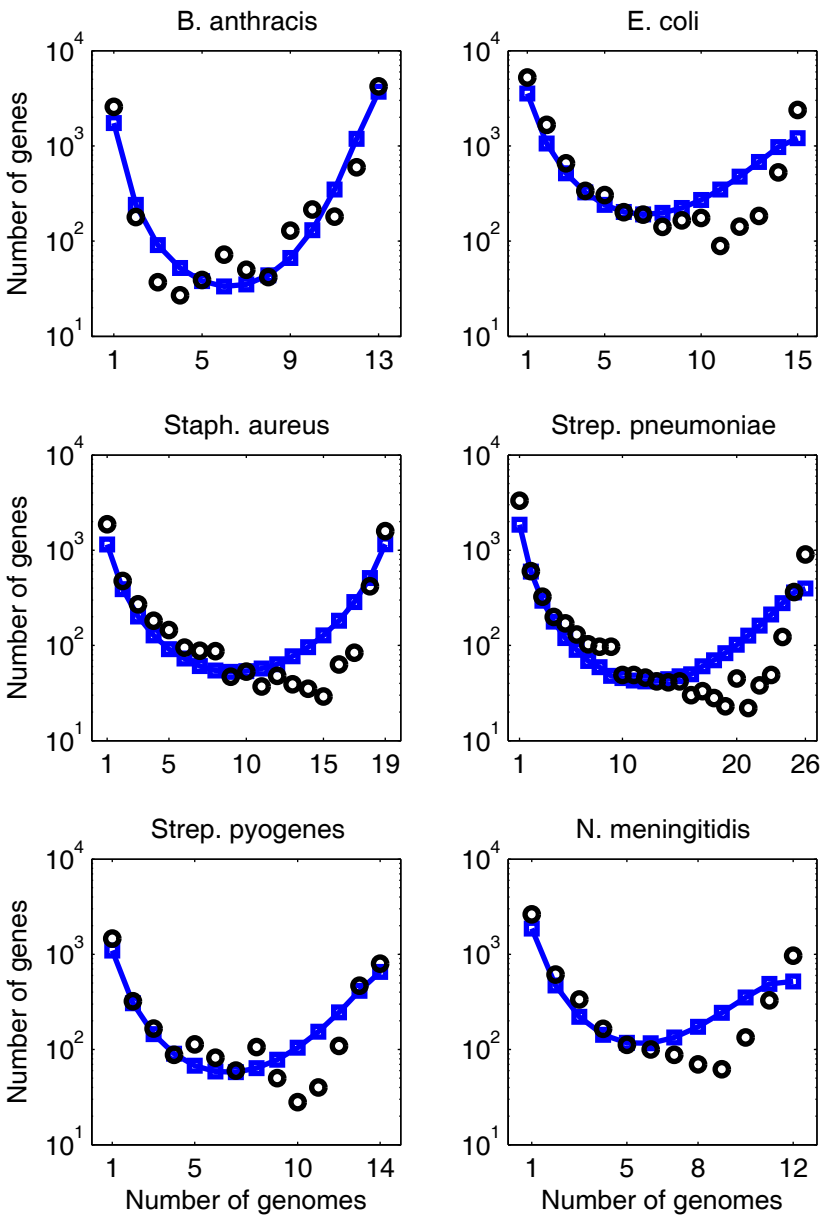

Figure 4 Data comparison for model with exponentially growing populations (model B). Comparison of gene frequency distributions with predictions of the model in which population size is assumed to grow exponentially. The model has two parameters, the gene transfer parameter θ_0 and the population growth parameter β. Black circles: data; blue line with squares: model.

independent parameters. Model C predicts that about half of the genome belongs to the rigid core ($\lambda_1 \approx \lambda_2 \approx 0.5$, see Table 2). The other part of the genome is rather fluid, with estimated gene transfer parameter $\theta_1 \approx 1$ (see Table 2). This combination of parameters results (except for *B. anthracis* and *Strep. pyogenes)* in a steep dip for the common (but not core) genes of the frequency distributions. However, such a dip is not present in the data (Figure 5).

To weaken the assumption of a rigid core, we consider another model with an explicit core genome, which we call model D, in which genes in the core genome retain some fluidity. More precisely, as for model C, we divide the genomes into two parts. Both parts are governed by the gene transfer process of model A, but the genes in the first part (fraction λ_1, gene transfer parameter θ_1) are more fluid than the genes in the second part (fraction $\lambda_2 = 1 - \lambda_1$, gene transfer parameter $\theta_2 < \theta_1$). Hence,

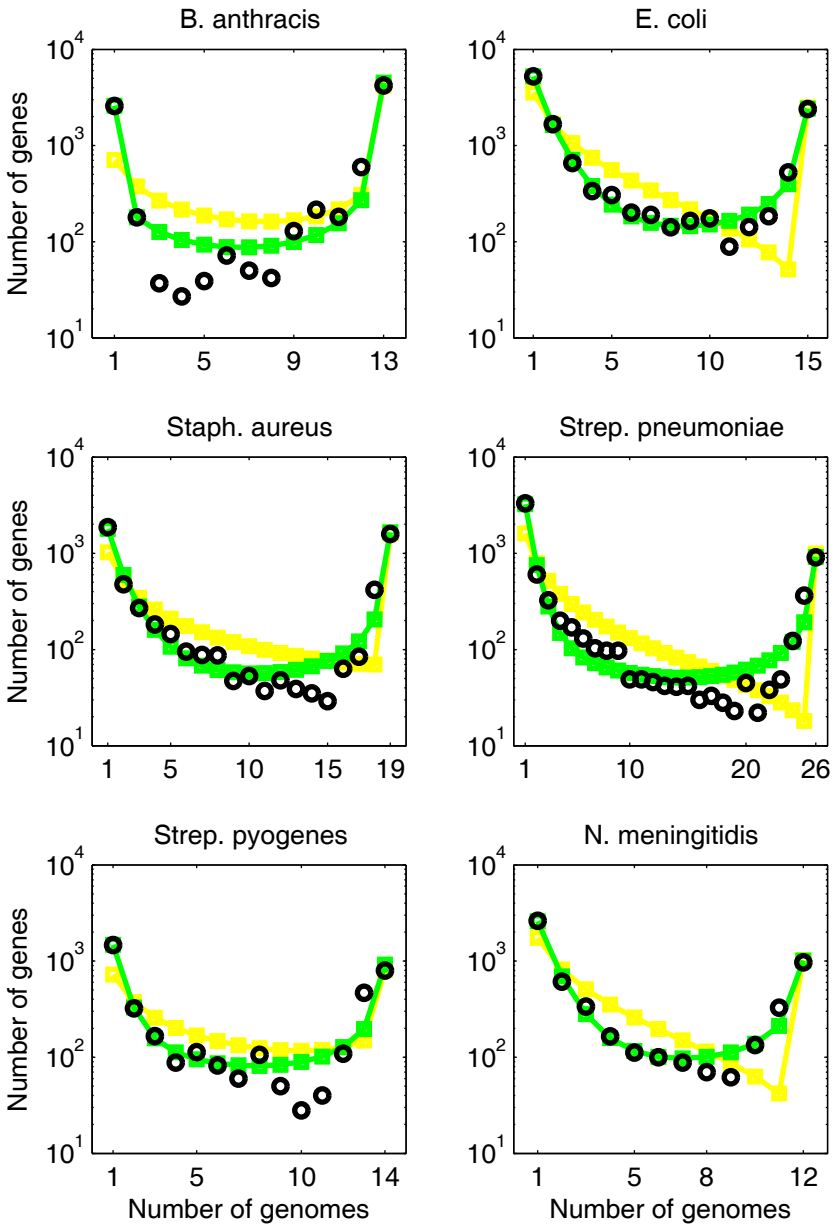

Figure 5 Data comparison for models with rigid and flexible core genomes (models C and D). Comparison of gene frequency distributions with predictions of two models which assume that a part of a genome is more susceptible to gene transfer. The genomes in model C have a rigid core, i.e., some genes cannot be removed from the genomes. The genomes in model D have a flexible core, i.e., theses core genes can be moved around between genomes, but to a lesser extent than the other genes. Model C has two parameters, whereas model D has three parameters. Black circles: data; yellow line with squares: model C; green line with squares: model D.

model D has three independent parameters. The average gene frequency distribution is equal to the sum of two distributions (1), with parameters θ_1 and θ_2, see Additional file 1: Appendix S5. The predicted distributions show an excellent agreement with the empirical data (Figure 5), as can also be seen from the fitting error Δ: for *E. coli* and *N. meningitidis* the error has dropped tenfold compared to the other models, see Table 1. It is also interesting to note that the model consistently (for the 6 bacterial species, although *B. anthracis* clearly stands out) predicts genomes with a small part of high fluidity ($\lambda_1 \approx 0.1$, $\theta_1 \approx 10$) and a large part of low fluidity ($\lambda_2 \approx 0.9$, $\theta_2 \approx 0.1$). Moreover, the model with a flexible core genome (model D) predicts the scaling of sample pan and core genome sizes in close agreement with the data, see Figure 6. Further, because no

Table 2 Parameter values of model fits

	model A	model B		model C		model D	
	θ	θ_0	β	θ_1 (λ_1)	θ_2 (λ_2)	θ_1 (λ_1)	θ_2 (λ_2)
B. anthracis	0.10	6.9	490	0.30 (0.41)	0 (0.59)	∞ (0.03)	0.06 (0.97)
E. coli	0.44	2.1	17	1.77 (0.46)	0 (0.54)	12 (0.15)	0.15 (0.85)
Staph. aureus	0.24	0.87	10	0.92 (0.42)	0 (0.58)	14 (0.07)	0.12 (0.93)
Strep. pneumonia	0.48	1.94	17	1.47 (0.53)	0 (0.47)	41 (0.08)	0.20 (0.92)
Strep. pyogenes	0.33	1.93	23	0.57 (0.68)	0 (0.32)	40 (0.06)	0.20 (0.94)
N. meningitidis	0.50	3.5	30	1.72 (0.52)	0 (0.48)	15 (0.16)	0.19 (0.84)

We determined for each of the four models A, B, C and D the parameters that minimize the distance Δ, see Eq. (6), between the empirical and the theoretical gene frequency distribution. Here we report the gene transfer parameter θ of the model A fit, the gene transfer parameter θ_0 and the population growth parameter β of the model B fit, the genome fractions λ_1 and λ_2, and the gene transfer parameter θ_1 of the model C fit, and the genome fractions λ_1 and λ_2, and the gene transfer parameters θ_1 and θ_2 of the model D fit.

additional free parameters were utilized in making this fit, such scaling represents an additional prediction of each of the models.

Genomic fluidity as a mechanistic summary statistic for gene frequency distributions

In a previous work [19] we advocated for the use of robust diversity indices to describe gene variation between genomes. In doing so we proposed the use of "genomic fluidity" which captures the average dissimilarity of pairs of genomes from within a group based on gene content. Specifically, genomic fluidity is equal to the probability that a randomly chosen gene from one genome is not found in another genome within the same group of organisms. For a sample of G genomes it can be estimated using the following formula:

$$\varphi = \frac{2}{G(G-1)} \sum_{\substack{k,\ell=1 \\ k<\ell}}^{G} \frac{U_k + U_\ell}{M_k + M_\ell}, \tag{4}$$

where U_k and U_ℓ are the number of genes found in either (but not both) genomes k and ℓ respectively in a pairwise comparison, and M_k and M_ℓ are the total number of genes found in genomes k and ℓ respectively. Estimates of genomic fluidity within a sample should agree with the true value of genomic fluidity within the population, in part, because they do not depend on the frequency of rare genomes or genes [19]. Note that genomic fluidity summarizes gene frequency distributions, however multiple gene frequency distributions may be compatible with the same value of genomic fluidity. Hence, here we ask whether genomic fluidity is related to the model parameters, θ, θ_0, and β, that underlie the gene frequency distributions presented here.

For the model with constant population size, the genomic fluidity ϕ and the gene transfer parameter θ are intimately linked. For a population in steady state, we have, see Additional file 1: Appendix S1,

$$\varphi = \frac{\theta}{1+\theta} \quad \text{or} \quad \theta = \frac{\varphi}{1-\varphi}. \tag{5}$$

Hence, genomic fluidity has a one-to-one relationship with θ, the relative rate of gene uptake to genomic replacement. When genomic fluidity approaches 1, then genomes are nearly completely dissimilar, which implies large gene replacements relative to genome reproductions (large θ). When genomic fluidity approaches 0, then processes that promote convergence of genomes are more important than gene-uptake processes (small θ). Previously, we advocated for the use of genomic fluidity on statistical grounds as a means to compare gene diversity between groups of genomes and as an alternative to the estimation of pan and core genome diversity. The constant population size model demonstrates that genomic fluidity may be indicative of processes driving the uptake of genes from the environment vs. genetic drift.

For the model with exponentially growing population size, the relationship between the genomic fluidity ϕ and the parameters θ_0 (for gene transfer) and β (for population growth) is more intricate. Genomic fluidity increases with the gene transfer parameter θ_0 and with the population growth parameter β, but there is no simple formula for $\phi(\theta_0, \beta)$ analogous to Eq. (5). However, the genomic fluidity is useful to clarify the estimation of the parameters θ_0 and β, see Additional file 1: Appendix S4. Indeed, the different model fits return very similar estimates for the genomic fluidity (including the models with rigid and flexible cores), see Table 1. This illustrates the robustness of genomic fluidity, confirming our previous findings [19]. However, this robustness comes with a trade-off: because very different parameter combinations θ_0 and β have the same genomic fluidity, we are unable to infer the gene transfer parameter θ_0 and the gene transfer rate σ from the genomic fluidity alone. This is a very typical finding in dynamic models in that predictions can be robust even when inferences of exact

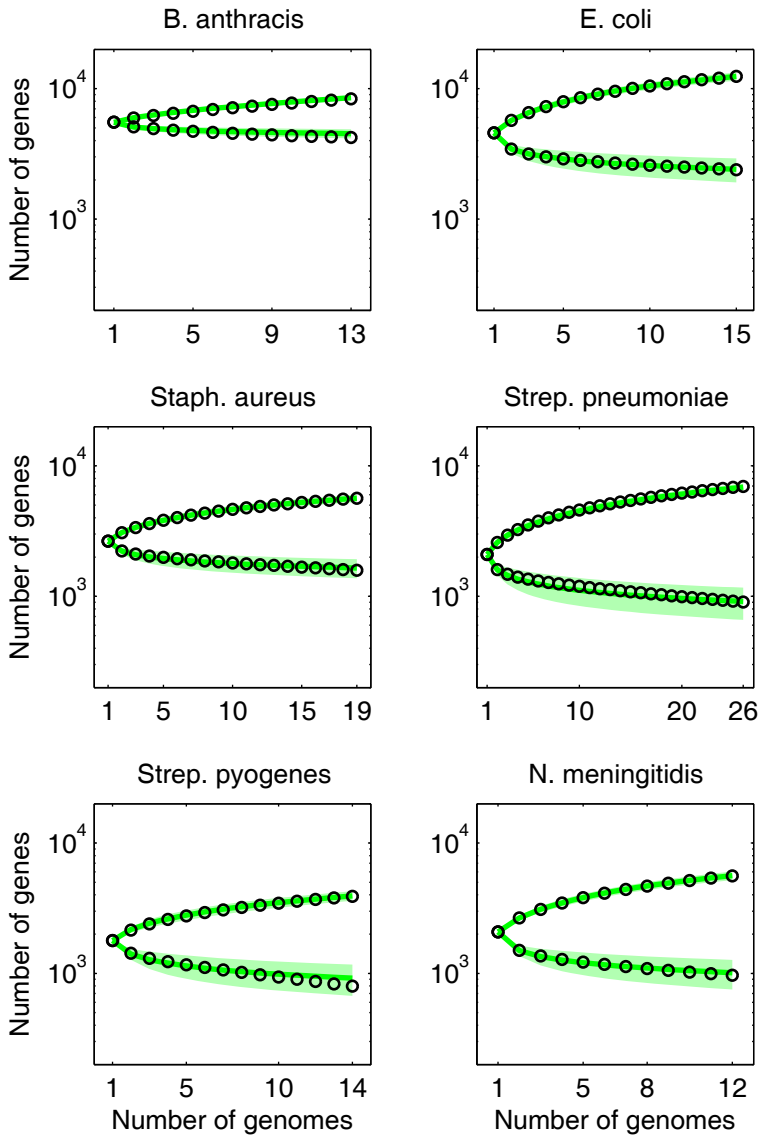

Figure 6 Predictions for observed core and pan genome size for model D. We used the parameters λ_1, θ_1 and θ_2 obtained from fitting the gene frequency distribution (see Figure 5) to evaluate the predicted core and pan genome size (see Additional file 1: Appendix S6). Black circles: data; green line: mean prediction; green shaded region: standard deviation of prediction. The increasing curves are for the pan genome; the decreasing curves are for the core genome.

combinations of mechanistic parameters may not always be possible from model fits [28].

Conclusions

We have presented a neutral model of genome evolution that combines birth-death processes at the population level with gene transfer events at the genome level. We find that this model generically yields U-shaped gene frequency distributions. This result suggests that a gene's prevalence is insufficient to infer its essentiality to a species. We compared our model to empirical gene frequency distributions estimated from sequenced genomes of six bacterial species and found: (i) reasonable fits to data; (ii) improved fits when assuming non-constant population size or including an explicit core genome; (iii) despite the qualitative agreement, that there still remains unexplained aspects of empirical gene frequency distributions, e.g., skewness; (iv) that our neutral model is remarkably compatible with a previous proposal for a robust gene diversity index - genomic fluidity [19]. We have also shown that our modelling framework can easily incorporate more complexity, which not

surprisingly gives improved fits. In this sense our model is also formally related to models of population genetics in which assumptions of population sizes and dynamics are meant to evaluate if and when spatial population structure and even ecological dynamics may alter allele distributions in identifiable ways [20,21]. In the present case the excellent fits obtained, e.g., for our flexible core model (model D), should not be interpreted as an indication for the validity of its assumptions. Rather, our analysis shows that gene frequency distributions do not contain sufficient information for the inference of evolutionary mechanisms underlying the observed distributions. Moreover, the finding that neutral models can generically lead to U-shaped gene frequency distributions suggests the need to incorporate and evaluate random processes in the analysis of gene composition and its dynamics.

Horizontal gene transfer is widely recognized as being an important mechanism driving genome evolution [22,25,29-31]. As such, there are many other models of evolution that address how neutral and selective processes give rise to variation in the state of genes and genomes (e.g., [16,17,20,32-35]). Indeed, the central model of population genetics in which individuals die and are replaced at random by other individuals is utilized here [20,21]. However, in the current model, genetic variation arises via the uptake of a novel gene. A recent paper also proposed a model of genome dynamics in which a rigid core was imposed [18] in order to fit gene frequency distributions estimated from 9 *Prochlorococcus* genomes. As shown here, such a fit may have limited inferential value, since a rigid core is not necessary in order to model gene frequency distributions. However, prior modeling suggests multiple avenues by which our model can be unified with dynamics at different scales. First, we have not considered horizontal gene transfer within genomes of the same species, nor recombination during division, nor of other types of transduction that may help to explain the finer genetic structure of bacterial populations [35-37]. Note that within-species gene transfer makes the acceptor genome more similar to the donor genome, and therefore reduces genetic diversity in the population just like genetic drift. We expect within-species gene transfer to have a smaller impact on genetic diversity than birth-death events, although its quantitative effect on the gene frequency distribution might be different (see [38] for an attempt to account for within-species gene transfer in a model similar to ours). Second, we do not include the fitness effect of mutations, whether neutral, beneficial or deleterious, which would impact the fixation of novel as well as pre-existing genes in genomes [16,17]. Including non-neutral mutational effects would obviously be a departure from our effort here to describe how much of the information on gene variation in genomes can be described using purely neutral models or simple extensions thereof. Note that although the impact of horizontally transferred genes on genome fitness remains controversial, there is evidence that such genes have no, or mild, effects on genome fitness [22]. Finally, a number of models have taken steps toward describing how the sizes of groups of genes, protein domains, proteins, and even categories of proteins (e.g., transcription factors) have changed over long evolutionary scales [32-34,39,40]. These models typically describe the structure within a genome (e.g., the abundance distribution of protein domains within different domain classes [34]), whereas our model describes population structure. It would seem that some unification of these models may be possible.

Development of models to predict and characterize gene composition variation among genomes is motivated by improvements in sequencing technologies which have enabled whole-genome sequencing of multiple isolates of the same bacterial species [41]. However, the gene frequency distribution data upon which we base this model is subject to two caveats. First, we treated the sequenced genomes as if they were sampled uniformly from the population. However, the genomes exhibit phylogenetic structure which should be taken into account. One option would be to use the total divergence of the core genes to correct for the non-uniform sampling (e.g., [42]), although alternative normalizations are possible. Second, determining whether two genes are found in a pair of genomes depends on the use of cutoffs within some comparative alignment scheme. Different cutoffs can be utilized depending on whether one is interested in gene homologs, orthologs, gene families, gene super-families, and so on. If the cutoffs are set too stringently, then nearly every gene will appear to be unique. If the cutoffs are set too loosely, then every gene will appear to be the same as every other. Prior work demonstrated that there exist metrics of gene composition dissimilarity (e.g. genomic fluidity) that are robust to changes in such cutoffs [19]. A unification of the current model with a sequence-based gene model would present opportunities to connect more factors (including mutation and recombination) driving gene variation with empirical patterns. However, we suggest that caution may be necessary in moving forward when attempting to utilize best fit parameters to infer mechanistic rates. In the present case, we showed that our neutral model reveals a well-known phenomenon of having robust predictions within a parameter space that poses an identifiability problem [28]. In essence, there are combinations of evolution parameters that yield similar predictions for gene frequency distributions (see Additional file 1: Figure S5 and Additional file 1: Appendix S4). Hence, more information is required concerning actual population size structure and the nature of gene uptake [43] before we recommend utilizing our best fits to precisely estimate gene transfer

rate, effective population size, growth rate and so on. More generally, fitted parameter values are subject to numerous simplifying assumptions of the models. Although it is interesting to compare the order of magnitude of the parameter fits with experimental data, one should be cautious to interpret the parameters too strictly as directly measurable quantities.

In this manuscript we presented a purely neutral explanation for the non-equal distributions of genes within genomes. The utilization of neutral models in genetics and ecology have yielded similar results in the past: in presenting quantitative arguments for when unequal patterns of appearance imply mechanisms of selection [12,44,45]. For example, a recent proposal for a unified theory of biodiversity and biogeography for forest trees [12] started with a similar dilemma. In that case, ecologists had observed skewed rank-abundance relationships such that some tree species were found at very high abundances and others at very low abundances. Ecologists had by and large assumed that those trees with greater abundance had a fitness advantage over trees with lower abundances. However, Hubbell's model showed that finding a few common trees and many rare trees could also be derived without invoking selection. Hence, in order to determine whether or not tree species had a fitness advantage in different regions one needed to look for correlations between traits and abundance which would not have been expected from a purely neutral model [46,47]. In the case considered here, our neutral model shows that the U-shape of gene frequency distributions provide less information than previously thought about the fitness benefit of genes. Instead, we need to find patterns of genome composition variation that can be explained by neutral models and identify those patterns or deviations from patterns that cannot be explained by neutral models - similar proposals have been advocated in other contexts [48]. Possible examples include the analysis of gene sequences and correlations amongst those gene present or absent amongst a set of genomes. In moving forward we suggest the need to continue to build the toolbox of a quantitative evolutionary genomics specifically adapted to the mechanisms operating within and amongst microbes.

Methods

Empirical estimation of gene frequency distributions

The pipeline has been described in detail elsewhere [19,23]. In brief, it (i) finds genes; (ii) calculates homology between all genes within a group of genomes using a set of cutoffs associated with identity and coverage (here set at 70% identity and coverage); (iii) applies a maximal clustering rule to determines groups of homologous genes; (iv) determines a gene presence-absence matrix of dimension $M_{\text{tot}} \times G$ of the total number of genes M_{tot} in the group of G genomes. We take row-sums of this matrix to find the frequency of each of the M_{tot} genes, and then take the histogram of these row-sums to calculate the gene frequency distribution. See Additional file 2 for the empirical gene frequency distributions of each of the six species analyzed here.

Estimating model parameters given empirical data

The parameter estimation is based on the average gene frequencies g_k. For model A the frequencies are computed using Eq. (1). For model B the frequencies are computed using the algorithm of Additional file 1: Appendix S3. For model C and D the frequencies are computed using the equations of Additional file 1: Appendix S5. To fit an empirical gene frequency distribution, we determine the model parameters that minimize the distance between the observed distribution g_k^{obs} and the predicted distribution g_k^{pred}. We use the following distance function Δ,

$$\Delta\left(g_k^{\text{obs}}, g_k^{\text{pred}}\right) = \frac{1}{G}\sum_{k=1}^{G}\left(\sqrt{g_k^{\text{obs}}} - \sqrt{g_k^{\text{pred}}}\right)^2, \qquad (6)$$

i.e., the mean square difference of the square-root transformed frequency distributions. We use a square-root transform to balance the different contributions to Δ. Without this transform large values of g_k (i.e., the tips of the U-shaped distribution at $k = 1$ and $k = G$) are weighed too heavily; with a logarithm transform small values of g_k (i.e., the intermediate part of gene frequency distribution, $2 \le k \le G - 1$) get proportionally too much weight. The fitted model parameters are reported in Table 2, and the corresponding distance Δ in Table 1. See Additional file 3 for Matlab scripts utilized to estimate the best fit parameters for each model.

Estimating gene uptake rates from transformation frequency

DNA uptake rates are generally presented as transformation frequencies, i.e., the fraction of colony forming units (CFUs) which have taken up a marker sequence relative to the total number of CFUs. Let us denote ϵ as the transformation frequency in an uptake experiments in which growing cells are exposed to DNA for a time T in exponential growth phase. Consider the division rate of the cells to be r, irrespective of whether they have taken up the marker sequence or not. Hence, we can write the dynamics for the population of cells without, $s(t)$, and with, $m(t)$, the marker sequence: $ds/dt = rs - \sigma s$ and $dm/dt = rm + \sigma s$, where σ is the gene uptake rate we would like to estimate. The solutions to these equations are $s(t) = s_0 e^{(r-\sigma)t}$ and $m(t) = s_0 e^{rt}\,(1 - e^{-\sigma t})$, where

s_0 is the initial population of cells. Hence, at the end of the experiment, $\epsilon = m(T)/(s(T) + m(T))$ or $\epsilon = 1 - e^{-\sigma T}$. Hence, σ can be estimated from the measurement of ϵ by solving $\sigma = -\frac{1}{T}\log(1-\varepsilon)$.

Additional material

Additional file 1: Appendix. Composed appendix, including text and figures providing additional information on how to derive gene frequency distributions, pan and core genome scaling, and calculate model fits.

Additional file 2: Gene frequency data. Raw data of the number of genes found in number of genomes for each of the 6 species analyzed here (in RTF format).

Additional file 3: Model fit scripts. A set of Matlab files to estimate the best fit parameters for each of the 4 models when given gene frequency data (in RTF format).

Acknowledgements
The authors thank K. Jordan, M. Cosentino Lagomarsino, T. Read, F. Stewart, S. Yi, and the anonymous reviewers for comments and suggestions on the manuscript. The authors thank A. Kislyuk for assistance with bioinformatics analysis. JSW acknowledges the support of Defense Advanced Projects Research Agency under grant HR0011-09-1-0055. JSW holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

Author details
[1]INRIA Research Team MODEMIC, UMR MISTEA, 34060 Montpellier, France. [2]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. [3]School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA.

Authors' contributions
BH and JSW designed the study, developed the model, analyzed models and data, and co-wrote the manuscript. Both authors read and approved the final manuscript.

Received: 3 January 2012 Accepted: 21 May 2012
Published: 21 May 2012

References
1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit , Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae:* Implications for the microbial "pan-genome.** *Proc Natl Acad Sci USA* 2005, **102(39)**:13950-13955.
2. Hotopp JDC, Grifantini R, Kumar N, Tzeng YLL, Fouts D, Frigimelica E, Draghi M, Giuliani MMM, Rappuoli R, Stephens DS, Grandi G, Tettelin H: **Comparative genomics of *Neisseria meningitidis:* core genome, islands of horizontal transfer and pathogen-specific genes.** *Microbiology* 2006, **152(12)**:3733-3749.
3. Hogg J, Hu F, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich G: **Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains.** *Genome Biol* 2007, **8(6)**:R103+.
4. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, Barbadora K, Klimke W, Dernovoy D, Tatusova T, Parkhill J, Bentley SD, Post JC, Ehrlich GD, Hu FZ: **Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome.** *J Bacteriol* 2007, **189(22)**:8186-8195.
5. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J: **The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates.** *J Bacteriol* 2008, **190(20)**:6881-6893.
6. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell D, Wain J, Dolecek C, Achtman M, Dougan G: **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*.** *Nat Gen* 2008, **40(8)**:987-993.
7. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ: **Biogeography of the *Sulfolobus islandicus* pan-genome.** *Proc Natl Acad Sci USA* 2009, **106(21)**:8605-8610.
8. Schoen C, Tettelin H, Parkhill J, Frosch M: **Genome flexibility in *Neisseria meningitidis*.** *Vaccine* 2009, **275(S2)**:B103-B111.
9. Chen PE, *et al*: **Genome characterization of the *Yersinia* genus.** *Genome Biol* 2010, **11(1)**:R1.
10. D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A: ***Legionella pneumophila* pangenome reveals strain-specific virulence factors.** *BMC Genomics* 2010, **17**:181.
11. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome.** *Trends in Genet* 2009, **25(3)**:107-110.
12. Hubbell SP: *The Unified Neutral Theory of Biodiversity and Biogeography* Princeton: Princeton University Press; 2001.
13. Gravel D, Canham CD, Beaudet M, Messier C: **Reconciling niche and neutrality: the continuum hypothesis.** *Ecol Lett* 2006, **9(4)**:399-409.
14. Adler PB, HilleRisLambers J, Levine JM: **A niche for neutrality.** *Ecol Lett* 2007, **10(2)**:95-104.
15. Rosindell J, Hubbell SP, Etienne RS: **The unified neutral theory of biodiversity and biogeography at age ten.** *Trends in Ecol & Evol* 2011, **26(7)**:340-348.
16. Berg OG, Kurland CG: **Evolution of microbial genomes: sequence acquisition and loss.** *Mol Biol Evol* 2002, **19(12)**:2265-2276.
17. Novozhilov AS, Karev GP, Koonin EV: **Mathematical modeling of evolution of horizontally transferred genes.** *Mol Biol Evol* 2005, **22(8)**:1721-1732.
18. Baumdicker F, Hess W, Pfaffelhuber P: **The diversity of a distributed genome in bacterial popoulations.** *Ann Appl Probab* 2010, **20(5)**:1567-1606.
19. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS: **Genomic fluidity: an integrative view of gene diversity within microbial populations.** *BMC Genomics* 2011, **12**:32.
20. Ewens W: *Mathematical Population Genetics.* 2 edition. New York: Springer; 2005.
21. Wakeley J: *Coalescent Theory: An Introduction* Greenwood Village: Roberts and Company; 2009.
22. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution.** *Nat Rev Microbiol* 2005, **3(9)**:679-687.
23. Kislyuk AO, Katz LS, Agrawal S, Hagen MS, Conley AB, Jayaraman P, Nelakuditi V, Humphrey JC, Sammons SA, Govil D, Mair RD, Tatti KM, Tondella ML, Harcourt BH, Mayer LW, Jordan IK: **A computational genomics pipeline for prokaryotic sequencing projects.** *Bioinformatics* 2010, **26(15)**:1819-1826.
24. Fraser C, Hanage W, Spratt B: **Recombination and the nature of bacterial speciation.** *Science* 2007, **315(5811)**:476-480.
25. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP: **The bacterial species challenge: making sense of genetic and ecological diversity.** *Science* 2009, **323(5915)**:741-746.
26. Dorer MS, Fero J, Salama NR: **DNA damage triggers genetic exchange in *Helicobacter pylori*.** *PLoS Pathogens* 2010, **6(7)**:e1001026.
27. Duffin P, Seifer H: **DNA uptake sequence-mediated enhancement of transformation in *Neisseria gonorroeae* is strain dependent.** *J Bacteriol* 2010, **192(17)**:4436-4444.
28. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP: **Universally sloppy parameter sensitivities in systems biology models.** *PLoS Comput Biol* 2007, **3(10)**:e189.
29. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405(6784)**:299-304.
30. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55(1)**:709-742.
31. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19(12)**:2226-2238.

32. Huynen MA, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes.** *Mol Biol Evol* 1998, **15(5)**:583-589.
33. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313(4)**:673-681.
34. Cosentino Lagomarsino M, Sellerio A, Heijning P, Bassetti B: **Universal features in the genome-level evolution of protein domains.** *Genome Biol* 2009, **10(1)**:R12.
35. Vetsigian K, Goldenfeld N: **Global divergence of microbial genome sequences mediated by propagating fronts.** *Proc Natl Acad Sci USA* 2005, **102(120)**:7332-7337.
36. Maynard Smith J, Smith NH, O'Rourke M, Spratt BG: **How clonal are bacteria?** *Proc Natl Acad Sci USA* 1993, **90(10)**:4384-4388.
37. Fraser C, Hanage WP, Spratt BG: **Neutral microepidemic evolution of bacterial pathogens.** *Proc Natl Acad Sci USA* 2005, **102(6)**:1968-1973.
38. Baumdicker F, Pfaffelhuber P: **Evolution of bacterial genomes under horizontal gene transfer.** *Proceedings of the 58th World Statistics Congress (ISI2011)* 2011.
39. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution.** *Nature* 2002, **420(6912)**:218-223.
40. Maslov S, Krishna S, Pang TY, Sneppen K: **Toolbox model of evolution of prokaryotic metabolic networks and their regulation.** *Proc Nat Acad Sci USA* 2009, **106(24)**:9743-9748.
41. Shendure J, Ji HL: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, **26(10)**:1135-1145.
42. Konstantinidis KT, Tiedje JM: **Genomic insights that advance the species definition for prokaryotes.** *Proc Natl Acad Sci USA* 2005, **102(7)**:2567-2572.
43. Beiko RG, Doolittle WF, Charlebois RL: **The impact of reticulate evolution on genome phylogeny.** *Syst Biol* 2008, **57(6)**:844-856.
44. Kimura M: *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press; 1983.
45. Nei M, Suzuki Y, Nozawa M: **The neutral theory of molecular evolution in the genomic era.** *Annu Rev Genomics Hum Genet* 2010, **11**:265-289.
46. Harpole WS, Tilman D: **Non-neutral patterns of species abundance in grassland communities.** *Ecol Lett* 2006, **9(1)**:15-23.
47. Kraft NJB, Valencia R, Ackerly DD: **Functional traits and niche-based tree community assembly in an amazonian forest.** *Science* 2008, **322(5901)**:580-582.
48. Koonin EV: **Are there laws of genome evolution?** *PLoS Comput Biol* 2011, **7(8)**:e1002173.

doi:10.1186/1471-2164-13-196
Cite this article as: Haegeman and Weitz: **A neutral theory of genome evolution and the frequency distribution of genes.** *BMC Genomics* 2012 **13**:196.