Evolution of bacterial genomes under horizontal gene transfer

Baumdicker, Franz
Pfaffelhuber, Peter
Albert-Ludwigs-Universität Freiburg, Abteilung für Mathematische Stochastik
Eckerstraße 1
79104 Freiburg, Germany
E-mail: baumdicker@stochastik.uni-freiburg.de
E-mail: p.p@stochastik.uni-freiburg.de

Introduction

Unraveling the evolutionary forces shaping bacterial diversity can today be tackled using a growing amount of genomic data. In recent years, the number of completely sequenced prokaryotic genomes has increased to around 1700 (NCBI). In particular, first datasets are available for samples of complete genomes from closely related strains, which are of the same bacterial species (Medini et al., 2005; Tettelin et al., 2005, 2008). Such datasets mark a revolution of microbial evolutionary biology, which turned from a theory-rich/data-poor subject into a data-rich/theory-poor field in the last decade.
While the genome of eukaryotes is highly stable, bacterial genomes from cells of the same species highly vary in gene content. For example, the pathogenic strain *E. coli* O157:H7 carries 1387 genes which are absent in the commensal *E. coli* K-12 (Perna et al., 2001). This huge variation in gene content led to the concepts of the *distributed genome* of bacteria and their *pangenome* (Tettelin et al., 2005; Ehrlich et al., 2005). In datasets, genes present in all genomes of a taxon are called *core genes* while genes present in only some but not all individuals comprise the *accessory genome*.
Gene content diversity originates in horizontal exchange of genomic material and pseudogenozation followed by gene loss. In particular, the amount of genetic exchange within a bacterial species determines the level of clonality (Smith et al., 1993). There are three different mechanisms of horizontal genetic exchange: (a) Transformation is the uptake of genetic material from the environment. (b) When a bacterium is infected by a lysogenic virus (phage) it provides additional genetic material that can be built in the bacterial genome. This process is known as transduction. (c) Conjugation requires a direct link (pilus) between two bacterial cells and leads to exchange of genetic material. These three mechanisms are usually referred to as horizontal gene flow. Recently, small virus-like elements called Gene Transfer Agents (GTAs) have been hotly debated to be the most important source for horizontal genetic exchange in some species (McDaniel et al., 2010).
We present a population genetic model for gene content evolution which accounts for several mechanisms. Gene uptake from the environment is modeled by events of *gene gain* along the genealogical tree, which describes the relationships between the individuals of the population. Pseudogenization may lead to deletion of genes and is incorporated by *gene loss*. These two mechanisms were studied by Huson and Steel (2004) using a fixed phylogenetic tree. Taking the random genealogy given by the coalescent (Kingman, 1982; Hudson, 1983), we studied the resulting genomic diversity already in Baumdicker et al. (2010) (see also Baumdicker et al., 2011). In the present paper, we extend the model in order to incorporate events of intraspecies horizontal gene transfer. Within this model, we derive expectations for the gene frequency spectrum and other quantities of interest.

The model

We consider the following model for bacterial evolution: Each bacterial cell carries a set of *genes* and every gene belongs either to the *core genome* or the *accessory genome*. The infinite set $I := [0, 1]$ is the set of conceivable accessory genes and $\mathcal{G}_c$ with $\mathcal{G}_c \cap I = \emptyset$ is the core genome. A population of

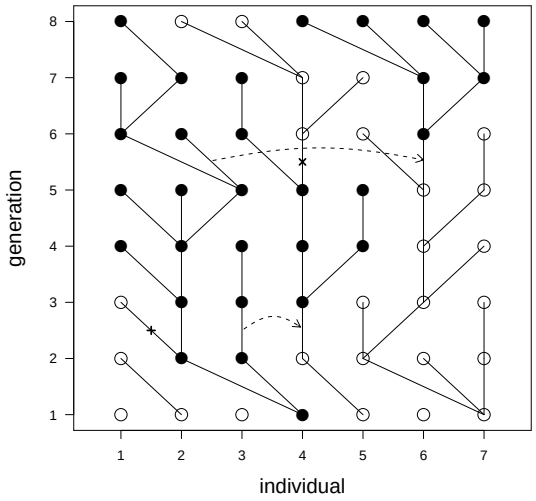

Figure 1: Genes are with probability $\theta/(2N)$ per individual per generation. In this illustration only one gene is gained in generation 1 in individual 4. Individuals carrying this gene are shown in black. Offspring inherits the gene from its ancestor, unless a loss event occurs (cross), with probability $\rho/(2N)$. With probability $\gamma/(2N^3)X(1-X)$, a gene is transferred to a random individual, such that now both the donor and the acceptor carry the gene.

constant size consists of N individuals (bacterial cells). We model the accessory genome of individual i by a finite counting measure $\mathcal{G}_i(t)$ on I. We will identify finite counting measures with the set of atoms, i.e. we write $u \in \mathcal{G}_i(t)$ if $\langle \mathcal{G}_i(t), 1_u \rangle \geq 1$.
The population evolves according to Wright–Fisher dynamics. That is, generations are discrete and individual (bacterial cell) j in generation $t+1$ *chooses* a parent from generation t purely at random and independent of all other individuals at time $t+1$. We denote that parent by $A_j(t)$. In order to obtain the genome $\mathcal{G}_j(t+1)$, we follow the mechanisms:

1. *Gene loss*: Denote the $1-\rho/(2N)$-thinning of $\mathcal{G}_{A_j(t)}(t)$ by $\mathcal{G}'_j(t+1)$. That is, $u \in \mathcal{G}'_j(t+1)$ iff $u \in \mathcal{G}_{A_j(t)}(t)$ and an independent coin with success probability $1-\rho/(2N)$ shows a success.

2. *Gene gain*: Choose an independent random counting measure $\mathcal{H}'_j(t+1)$ according to a Poisson process on I with intensity $\theta/(2N)$.

3. *Horizontal gene transfer*: For every $i = 1, \ldots, N$ (the *donor*) and $v \in \mathcal{G}_i(t)$ (the *transferred gene*), let $v \in \mathcal{H}''_j(t+1)$ with probability $\gamma/(2N^3)$. In this event, individual j is called the *acceptor* of gene v.

Finally, set

$$\mathcal{G}_j(t+1) = (\mathcal{G}'_j(t+1) + \mathcal{H}'_j(t+1) + \mathcal{H}''_j(t+1)) \wedge 1$$

for the genome of individual j in generation $t+1$. The '$\wedge 1$'-term indicates that we do not model paralogous genes, i.e. horizontal gene transfer events have no effect if the acceptor individual j already carries the transferred gene. We refer to $(\mathcal{G}_1(t), \ldots, \mathcal{G}_N(t))_{t=0,1,2,\ldots}$ undergoing the above dynamics as the Wright–Fisher model for bacterial genomes with horizontal gene flow. It can be shown that this Markov chain is Harris recurrent and hence, has a unique equilibrium.

We are mainly interested in large populations and a rescaling of time by a factor of N. The corresponding limit is usually refered to as large population limit in the population genetic literature. The following argument is crucial in the proof of our main result, Theorem 1. Let $X^N(t)$ be the frequency of gene u in generation $[tN]$ in the population of size N. Then, in the large population limit, $N \to \infty$, the process $(X^N(t))_{t\geq 0}$ converges weakly to the solution of the SDE

$$dX = \big(-\tfrac{\rho}{2}X + \tfrac{\gamma}{2}X(1-X)\big)dt + \sqrt{X(1-X)}dW \tag{1}$$

for some Brownian motion W. To see this, note that the evolution of frequencies of gene u is an autonomous process. The diffusion term is associated with random reproduction events and has the well-known form from (1) (Ewens, 2004), known as Wright–Fisher noise. Gene loss reduces X^N with probability approximately proportional to X and $\rho/(2N)$. After rescaling of time, this turns into the rate $-(\rho X/2)dt$. Last, horizontal gene transfer increases X^N with probability approximately proportional to $\gamma/(2N^3)$ and to the number of pairs where the horizontal gene transfer events has an effect, $N^2X^N(1-X^N)$. After rescaling of time, this turns into the rate $\frac{\gamma}{2}X(1-X)dt$.

Sample statistics

Consider a sample $\mathcal{G}_1, \ldots, \mathcal{G}_n$ of size n taken from the population. We consider several statistics under the above dynamics:
The *average number of genes (in the accessory genome)* is given by

$$A := A^{(n)} := \frac{1}{n}\sum_{i=1}^{n} |\mathcal{G}_i| \tag{2}$$

where $|\mathcal{G}_i| := \langle \mathcal{G}_i, 1\rangle$ is the total number of accessory genes in individual i.
The *average number of pairwise differences* is given by

$$D := D^{(n)} := \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} |\mathcal{G}_i \setminus \mathcal{G}_j| \tag{3}$$

where $\mathcal{G}_i \setminus \mathcal{G}_j := (\mathcal{G}_i - \mathcal{G}_j)^+$ are the genes present in i but not in j.
The *size of the accessory genome* is given by

$$G := G^{(n)} := \Big|\bigcup_{i=1}^{n} \mathcal{G}_i\Big| \tag{4}$$

where $\bigcup_{i=1}^n \mathcal{G}_i = \Big(\sum_{i=1}^n \mathcal{G}_i\Big) \wedge 1$ is the set of genes present in any individual from the sample.
The *gene frequency spectrum (of the accessory genome)* is given by $G_1 := G_1^{(n)}, \ldots, G_n := G_n^{(n)}$, where

$$G_k^{(n)} := G_k := |\{u \in I : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } i\}|. \tag{5}$$

Results

Using diffusion theory, we obtain first moments of all of the above statistics in equilibrium. We start with expectations of $G_1^{(n)}, \ldots, G_n^{(n)}$, since all other quantities can be expressed in terms of the gene frequency spectrum. The proof of Theorem 1 can be found at the end of the manuscript.

Theorem 1 (Gene frequency spectrum)**.** *Consider a sample of size n taken from the Wright–Fisher model for bacterial genomes with horizontal gene flow with $\rho > 0, \theta > 0, \gamma \geq 0$ in equilibrium. Then, as $N \to \infty$,*

$$\mathbb{E}[G_k^{(n)}] = \frac{\theta}{k}\frac{n\cdots(n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)}\Big(1 + \sum_{m=1}^{\infty}\frac{(k)_m\gamma^m}{(n+\rho)_m m!}\Big) \tag{6}$$

with $(a)_b := a(a+1)\cdots(a+b-1)$.

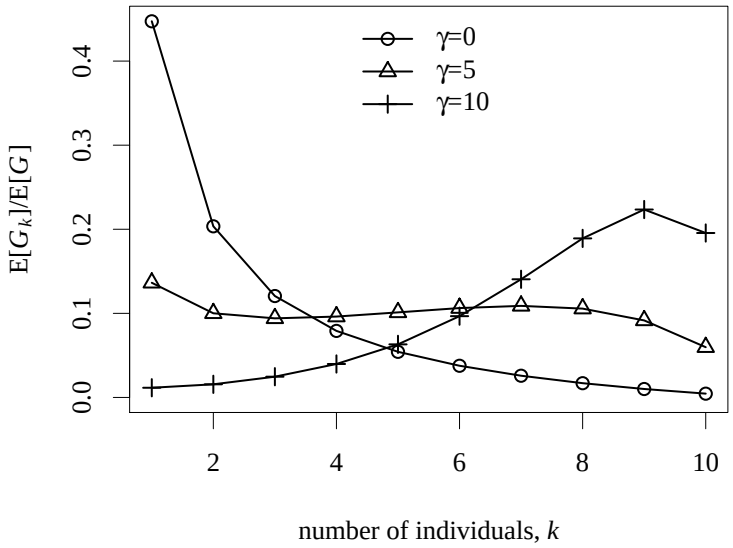

Figure 2: The expected gene frequency spectrum is highly dependent of γ, the rate of horizontal gene flow. For high values of γ, most genes are in high frequency, leading to a closed pangenome. We use $\rho = 2$ in the figure.

Corollary 2 (More summary statistics of gene content)**.**
Under the same assumptions as in Theorem 1,

$$\mathbb{E}[A^{(n)}] = \frac{\theta}{\rho}\left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(1+\rho)_m}\right), \tag{7}$$

$$\mathbb{E}[D^{(n)}] = \frac{\theta}{1+\rho}\left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(2+\rho)_m}\right), \tag{8}$$

$$\mathbb{E}[G^{(n)}] = \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \theta \sum_{m=1}^{\infty} \frac{\gamma^m}{m}\Big(\frac{1}{\rho_m} - \frac{1}{(n+\rho)_m}\Big). \tag{9}$$

Discussion

We introduced the Wright–Fisher model for bacterial genomes with horizontal gene flow, measured by the parameter γ. Since the corresponding model without horizontal gene flow was considered in Baumdicker et al. (2010), we note that Theorem 1 and Corollary 2 imply that all sample statistics are continuous at $\gamma = 0$.
Recently, the concepts of *open* and *closed* pangenomes were introduced (Medini et al., 2005). If, after sequencing a finite number of genomes, all genes present in the population are found, one speaks of a closed pangenome. If new genes are found even after sequencing many cells, the pangenome is *open*. It is not hard to see that high values of γ imply that most genes are in high-frequency. In other words, sequencing a new individual hardly leads to new genes which were not seen before. This impact of openness and closedness of the pangenome can as well be seen from Figure 2.

Consider the diffusion (1), which describes the approximate frequency of individuals carrying one specific gene $u \in I$. Usually, the $\frac{\gamma}{2}X(1-X)$-term appears in population genetic models only due to a selective force (see e.g. Kimura, 1964; Ewens, 2004; Durrett, 2008). In the present setting, it appears

because horizontal gene flow increases the frequency of the gene by a rate which is proportional to the number of possible donor/acceptor-pairs of individuals.
Due to the close connection of horizontal gene transfer with selective models, a comparison to recent work is appropriate. In particular, the theory for the frequency spectrum in selective models with irreversible mutations is carried out in Fisher (1930); Wright (1938); Kimura (1964, 1969). Additionally, Sawyer and Hartl (1992) developed a Poisson Random Field model for selective sites. (Extensions were e.g. given in Williamson et al., 2005.) They assume that a large set of unlinked loci is under selection. As a result, they obtain predictions for the number of alleles present in a subset k out of n of individuals. However, since their loci are unlinked, the random processes are completely independent for the different processes. This is in contrast to our approach where the reproduction within the Wright–Fisher model affects all genes in the same way, and the horizontal gene transfer only affects single genes. Moreover, our model is reversible in the sense that present genes may as well be lost (but not reintroduced). Since it has been shown in Baumdicker et al. (2010) that discontinuities at $\rho = 0$ (no gene loss) arise, it is not straight-forward to use these classical results in the present setting.

Simulations

We are interested in patterns of presence/absence of genes in a sample of size n in an equilibrium situation. If presence/absence of a gene would be independent of the state of the other genes, the gene frequency spectrum could be simulated by independent copies of the diffusion X, given in (1). However, all genes are inherited along the same lineages, so the frequency of two different genes depend on each other; see Figure 4. For $\gamma = 0$ the composition of the accessory genome of n individuals can be simulated backwards in time using the coalescent. As, for $\gamma = 0$, genomes only depend on events along the ancestral lineages of the sample this is very efficient (Hudson, 2002). In the case $\gamma > 0$, we simulate the accessory genome forward in time. Therefore, we have to consider all individuals of the population, as each of them might influence the accessory genome of the sample of size n by events of horizontal gene transfer. Unfortunately the size of the population has to be much larger than n to obtain values close to the large population limit results. Thus the forward simulations for $\gamma > 0$ are much slower than backward simulations for $\gamma = 0$.
Theorem 1 gives the expected sizes of the gene frequency spectrum. For $\gamma = 0$ it is known that the gene frequency spectrum highly depends on the underlying genealogy. E.g. if the genealogy separates the n individuals into two groups, one of size k and one of size $n - k$, then $G_k^{(n)}$ and $G_{n-k}^{(n)}$ increase. The same is true for simulated data with $\gamma > 0$, see Figure 3.

Proof of Theorem 1 and Corollary 2

We consider the diffusion (1) with infinitesimal mean and variance

$$b(x) = -\tfrac{1}{2}\rho x + \tfrac{1}{2}\gamma x(1-x), \qquad a(x) = x(1-x).$$

The Green function for the diffusion, measuring the time the diffusion, i.e. a gene, spends in frequency x until eventual loss, if the current frequency is δ, is given by

$$G(\delta, x) = 2\frac{\phi(\delta)}{a(x)\psi(x)},$$

where

$$\psi(y) := \exp\left(-2\int_0^y \frac{b(z)}{a(z)}dz\right) = (1-y)^{1-\rho}e^{-\gamma y},$$
$$\phi(x) := \int_0^x \psi(y)dy.$$

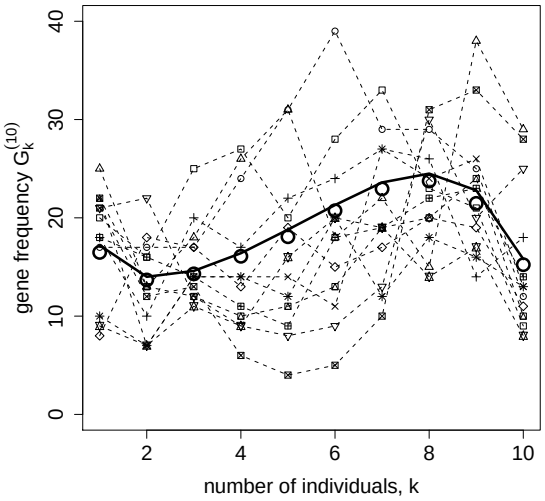

Figure 3: The expected gene frequency spectrum for $n = 10$, $\theta = 10$, $\gamma = 6$ and $\rho = 2$ is shown as a solid black line. For twelve different simulations with $N = 500$ individuals the gene frequency spectrum for $n = 10$ randomly chosen individuals is shown (dashed lines). The mean of 1000 simulations (black circles) is close to the results of Theorem 1.

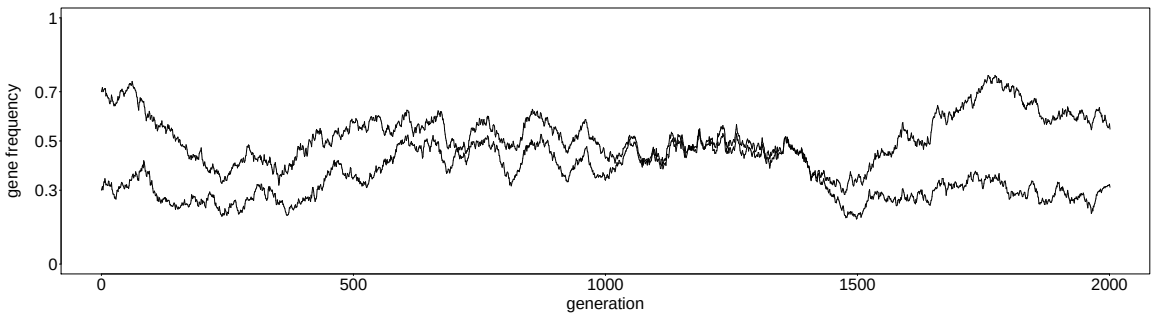

Figure 4: The frequencies for two different genes in a population of size 2000 are shown. Here $\rho = 0.2$ and $\gamma = 1$. At time zero initially 600 individuals carry gene 1, 800 individuals carry both, gene 1 and gene 2 and 600 individuals carry none of the two genes. The frequencies are not independent as both genes depend on the same underlying ancestral lineages. Since gene loss and horizontal gene transfer events occur independently for each gene, they can weaken the dependency of the two frequency paths.

Following Durrett (2008), we introduce new genes in frequency δ at rate $\frac{\theta}{2}\frac{1}{\phi(\delta)}$ in a consistent way. That is, the gene raises in frequency to $\varepsilon > \delta$ with probability $\frac{\phi(\delta)}{\phi(\epsilon)}$. Hence the number of genes in frequency x is Poisson with mean

$$\frac{\theta}{2}\frac{1}{\phi(\delta)}G(\delta,x) = \theta\frac{e^{\gamma x}}{x(1-x)^{1-\rho}}.$$

The gene frequency spectrum is now given by

$$\begin{aligned}
\mathbb{E}[G_k^{(n)}] &= \binom{n}{k}\int_0^1 \theta\frac{e^{\gamma x}}{x(1-x)^{1-\rho}}x^k(1-x)^{n-k}dx \\
&= \binom{n}{k}\theta\int_0^1 e^{\gamma x}x^{k-1}(1-x)^{n-k-1+\rho}dx \\
&= \theta\binom{n}{k}(k-1)!\frac{\Gamma(n-k+\rho)}{\Gamma(n+\rho)}{}_1F_1(k;n+\rho;\gamma)
\end{aligned}$$

where ${}_1F_1(k;n+\rho;\gamma) = 1+\sum_{m=1}^{\infty}\frac{(k)_m\gamma^m}{(n+\rho)_m m!}$ is a hypergeometric funtion and $(a)_b := a(a+1)\cdots(a+b-1)$. Given the gene frequency spectrum, it is now easy to compute first moments of A, D and G (see Corollary 2) by using

$$\begin{aligned}
&\mathbb{E}[A] = \mathbb{E}[G_1^{(1)}], \qquad\qquad \mathbb{E}[D] = \mathbb{E}[G_1^{(2)}], \\
&\mathbb{E}[G] = \sum_{k=1}^{n}\tfrac{1}{k}\mathbb{E}[G_1^{(k)}] = \sum_{k=1}^{n}\frac{\theta}{k}\frac{k}{k-1+\rho}\sum_{m=0}^{\infty}\frac{\gamma^m}{(k+\rho)_m} = \theta\sum_{m=0}^{\infty}\gamma^m\sum_{k=0}^{n-1}\frac{1}{(k+\rho)_{m+1}} \\
&\quad = \theta\sum_{k=0}^{n-1}\frac{1}{k+\rho} + \theta\sum_{m=1}^{\infty}\frac{\gamma^m}{m}\sum_{k=0}^{n-1}\Big(\frac{1}{(k+\rho)_m} - \frac{1}{(k+1+\rho)_m}\Big) \\
&\quad = \theta\sum_{k=0}^{n-1}\frac{1}{k+\rho} + \theta\sum_{m=1}^{\infty}\frac{\gamma^m}{m}\Big(\frac{1}{(\rho)_m} - \frac{1}{(n+\rho)_m}\Big).
\end{aligned}$$

References

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2010). The diversity of a distributed genome in bacterial populations. *Ann. Appl. Probab. 20*(5), 1567–1606.

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2011). The infinitely many genes model for the distributed genome of bacteria. *Preprint*.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution. Second edition.* Springer.

Ehrlich, G. D., F. Z. Hu, K. Shen, P. Stoodley, and J. C. Post (2005). Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin. Orthop. Relat. Res. 437*, 20–24.

Ewens, W. J. (2004). *Mathematical Population Genetics. I. Theoretical Introduction. Second edition.* Springer.

Fisher, R. (1930). The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh 50*, 205–220.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*, 183–201.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics 18*, 337–338.

Huson, D. H. and M. Steel (2004). Phylogenetic trees based on gene content. *Bioinformatics 20*(13), 2044–2049.

Kimura, M. (1964). Diffusion Models in Population Genetics. *J. Appl. Probab. 1*(2), 177–192.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics 61*, 893–903.

Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab. 19A*, 27–43.

McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul (2010). High frequency of horizontal gene transfer in the oceans. *Science 330*, 50.

Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev. 15*(6), 589–594.

Perna, N. T., G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Psfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature 409*, 529–533.

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

Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt (1993). How clonal are bacteria? *Proc. Natl. Acad. Sci USA 90*, 4384–4388.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome. *Proc. Natl. Acad. Sci. U.S.A. 102*(39), 13950–13955.

Tettelin, H., D. Riley, C. Cattuto, and D. Medini (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology 11*(5), 472–477.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante (2005, May). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U.S.A. 102*, 7882–7887.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. U.S.A. 24*, 253–259.