Part I Theoretical Considerations on the Evolution of Bacterial Pathogens

CHAPTER 1

Genomes in Motion: Gene Transfer as a Catalyst for Genome Change

Jeffrey G. Lawrence and Heather Hendrickson

> The change from species to species is not a change involving more and more additional atomistic changes, but a complete change of the primary pattern or reaction system into a new one, which afterwards may again produce intraspecific variation by micromutation.
>
> – Richard Goldschmidt, 1940

1.1. INTRODUCTION

Despite our interest and motivation, bacteria are not particularly easy organisms to study; their niches are complex and poorly understood and the vast majority of these species are difficult to culture or to manipulate in the laboratory. Of all bacteria, it is pathogens whose physical, social, and economic impact on our day-to-day lives garners the most attention, from both scientists and non-scientists alike. As a result, pathogens are among the best-studied bacteria, and lessons we learn from them are often generalized to other, non-pathogenic bacteria. Not surprisingly, the first lessons learned in the so-called genomic era came from pathogens, which were the first organisms with fully sequenced genomes. The promise of genomics was that the limitations of conventional microbiology could be overcome by studies of genome sequences and careful analysis of the genes contained therein. Here we examine how genomics has shaped our understanding of microbial genome evolution and ask how extensible these lessons may be. Among the notions that attracted widespread attention was the finding that certain clusters of genes are specifically responsible for virulence and that these loci are often obviously of foreign origin, having been introduced by the then under-appreciated process of lateral gene transfer or LGT.

Coming more than a decade after these findings, this volume is focused on the hugely influential role of LGT in the evolution of genomes, particularly those of pathogenic bacteria. The debate over the relative importance of lateral gene transfer is as old as discussion of the phenomenon itself (Doolittle, 1999a, 1999b; Kurland, 2000). Yet most will agree that gene transfer has not only been instrumental in the origin and diversification of pathogenic bacteria and fungi (Groisman and Ochman, 1994, 1997), but that the very nature of pathogenic organisms has been shaped by this process. In a world lacking LGT, pathogens would be very different creatures from the ones we see today. But to many biologists, LGT is difficult to integrate into a conceptual framework of organismal evolution. Darwinian evolution is traditionally viewed as gradual change, with endless cycles of minute refinements shaping and adapting an organism to its environment. Here, individuals combine and reassort subtle variants within populations to make fluid and natural transitions between phenotypic states. Those changes which increase fitness within a particular environment would be retained. In contrast, change induced by lateral transfer can be both startlingly quick and strikingly large; LGT can catalyze the sort of dramatic phenotypic change that recalls discussion of "hopeful monsters" in the middle part of the last century (Goldschmidt, 1940) wherein speciation was viewed as a fundamental organismal change. With the aid of LGT, lineages are no longer restricted to exploring logically accessible niches; rather, they could acquire any set of genes – from any other organism, living at any place or time – to thrust them into completely novel, previously unavailable ecological contexts. The LGT effectively removes the barriers – both genetic and metaphorical – between bacterial taxa, allowing heritable information as well as evolutionary and ecological potential to flow between them. It is this seemingly unbounded potential for taxonomic mixis that can be puzzling. If LGT is the norm for genome evolution, why isn't the world filled with hopeful chimeras? Why do organisms fall into groups with shared properties? Shouldn't a group of chimeras, descended from a long line of chimeras, be impossible to classify in this way? Is LGT a convenient excuse for unexplainable genomic phenomena? Such wariness is justified, but progress over the past decade has shown that while gene transfer can be a powerful source of genetic innovation, it does not necessarily lead to phylogenetic chaos (Gogarten *et al.*, 2002; Lawrence and Hendrickson, 2003). For example, there are constraints on which genes are transferred with ease and which are more recalcitrant, as encapsulated by the complexity hypothesis (Jain *et al.*, 1999), or mobile genome hypothesis leading to a broad pan-genome (Tettelin *et al.*, 2005). In addition, there are apparent

constraints on the phylogenetic breadth of donors and recipients (Beiko *et al.*, 2005), a conclusion validated by molecular mechanisms which may constrain gene transfer (Hendrickson and Lawrence, 2006; Lawrence and Hendrickson, 2003, 2004). Both sets of constraints have served to maintain order among bacterial genomes in the face of gene transfer. That is, apparent phylogenetic order (Brown *et al.*, 2001; Fitz-Gibbon and House, 1999; Snel *et al.*, 1999; Tekaia *et al.*, 1999) does not lead to the conclusion that the impacts and lateral gene transfer have been overstated; rather, there are rules governing gene transfer that have served to preserve the phylogenetic hierarchy we observe.

1.2. GENOME EVOLUTION IN PATHOGENS

The first views of genome evolution arose from traditional models of gene evolution; that is, genome evolution could be viewed as collective gene evolution. Population genetic studies of gene evolution have focused on the fate of mutations, which represent gradual changes in genes. The similarity in gene order among closely related bacteria seemingly reinforced the view that genomes are relatively stable and change content slowly. Yet the availability of complete genome sequences shows that gene content can vary widely among even very closely related strains (Konstantinidis and Tiedje, 2005). The framework of shared genes necessarily minimizes the apparent roles of gene gain, gene loss, genome rearrangement, and allelic replacement of genes via recombination with conspecific strains. More importantly, these are inter-related processes: examination of available genome sequences suggests that LGT has had more impact on genome evolution than simply expanding gene inventory. The tempo and targets of other processes – including rearrangement, gene loss, and gene replacement – are affected as well. In effect, gene acquisitions catalyze genome evolution by a wide array of mechanisms. Herein we discuss how genomes have become more fluid in their content and organization as a result of lateral gene transfer and associated processes.

1.3. DIRECT IMPACT OF GENE ACQUISITION

The most obvious impact of LGT is the introduction of novel genes into a genome. Because the acquisition of novel traits via gene acquisition differs markedly from phenotypic evolution via the modification of genes within the chromosome, the phenotypic, physiological, and evolutionary effects of

such transfers cannot be overstated. First, acquired genes will have had their functions honed by selection in their donor genome and can produce fully functional products immediately after arrival, thereby allowing effective competition in new niches. In contrast, the modification of existing genes can explore functions that are available within small numbers of mutational steps from the existing sequences. This is because, in population genetic terms, the genes do not provide robust functions during transition from one adaptive peak to another, making exploration of the adaptive landscape problematic. In addition, only those proteins available within the genome can be modified in this way, whereas LGT can introduce new protein families. Beyond the introduction of novel genes, gene transfer can introduce genes for complex functions, because more than one gene may be mobilized at the same time. These large regions of acquired DNA are termed pathogenicity islands or PAI and contained dozens of genes (Blum *et al.*, 1994; Hacker *et al.*, 1997; Hacker and Kaper, 2000). It is not feasible for this larger number of genes to have evolved by mutational processes simultaneously, even if paralogs were available in the genome as starting substrates. Importantly, the acquisition of large fragments of DNA is not confined to pathogens. Similar large fragments in other genomes have been termed genomic islands, or GEI, reflecting the ability of introgressed DNA to alter any organism's physiological toolbox (Dobrindt *et al.*, 2004). Functions required large numbers of genes – such as the synthesis of complex coenzymes (Lawrence and Roth, 1996a), or the suite of genes required for methanogenesis (Chistoserdova *et al.*, 1998) – can be introduced at one time. We will not dwell on the numerous functions encoded by PAI or GEI, as there are chapters in the volume discussing them in detail, but pathogenicity islands can contain encode enormously intricate functions, such as the synthesis and deployment of type III secretion systems (Ochman *et al.*, 1996). Dramatic changes may occur very quickly; for example, a pathogenic fungus is believed to have arisen in within the past few decades following the acquisition of pathogenicity islands (Friesen *et al.*, 2006). These adaptations may use functions that were refined elsewhere for different physiological functions. For example, the acquisition of siderophores allows pathogens to obtain critical ions present at low concentration (e.g., Fe^{2+} or HPO_4^{2-}); yet these functions evolved in non-pathogenic bacteria living in non-host-associated, ion-poor environments. The ability of LGT to introduce all of the genes required to perform complex functions provides the cell with a route for ecological adaptation that is, for all intents and purposes, entirely unavailable via traditional gene modification by mutation and selection. This forces microbiologists to consider avenues of phenotypic evolution once thought to be unconventional.

1.4. THE ROLE OF GATEWAY GENES

While the direct effects of gene gain can be extensive, the real potential for genome change lies in the long-term, indirect effects. The addition of new genes to a bacterial genome results in numerous related, downstream events that may change the evolutionary trajectory of bacterial chromosomes as much as the gene acquisition event itself. In the broadest sense, gene gain can cause a dramatic change in the ecological niche exploited by the cell. Consider that the foreign genes we observe in bacterial chromosomes represent only a small fraction of those that were introduced into that cell's cytoplasm and recombined into a stable replicon. After this occurs, all gene acquisition events are filtered, so to speak, by natural selection. Genes which provide no useful function, or those that are actively problematic, will not be retained. Others may persist for short periods of time when their benefits are transient; for example, we see such evolutionary lability with antibiotic resistance genes in most lineages, which can be lost as readily as they are gained (Kirkup and Riley, 2004).

In very few cases, incoming genes will provide a function that is beneficial to the cell and can be retained for long periods of time (Lawrence and Ochman, 1997, 1998). In this context, the new gene inventory will predispose a cell for a substantially different lifestyle than that of its ancestor. As a result, the suite of incoming genes that would be potentially beneficial to the cell will also differ. That is, each gene acquisition event alters the potential fitness contributions of other incoming genes, effectively making genomes 'moving targets' for LGT. Genes which would otherwise have provided no benefit to the cell may become highly advantageous once that cell has acquired other, niche-altering genes. When acquired genes change the adaptive landscape to such a large degree, we can consider them to be 'gateway' genes, differentially conditioning the cell to the acquisition of other genes. Many consider initially acquired pathogenicity islands to be gateway genes, transforming an otherwise commensal organism into a pathogen (although see below for further discussion of this point). The adoption of the pathogenic lifestyle then favorably disposes the cell to the acquisition of other pathogenicity islands. For example, genomic analysis of the plant pathogen *Erwinia carotovora* suggests that they acquired genes required for life in and around plants; these genes include those involved in type III protein secretion, phytotoxin production, plant-cell adhesion, and nitrogen fixation (Toth *et al.*, 2006). These functions complement other acquired genes – such as those allowing for degradation of plant cell-wall polysaccharides, or the production of siderophores – in the making, a formidable pathogen, but one attacking plants, not animals. The

route towards plant pathogenesis was imparted by acquisition of a gateway gene cluster opening up this ecological niche to the nascent *Erwinia* lineage.

1.4.1. Providing Sites for Additional Gene Insertion

The integration of large regions of DNA can facilitate additional gene acquisition events in a passive way by providing non-disruptive sites wherein incoming genes can insert. Bacterial chromosomes are information rich, and little intergenic space is available for the insertion of new DNA without disrupting the expression of existing genes. Some bacteriophages circumvent this constraint by using integrases that target tRNA genes, and recombination there is non-disruptive, because phage-borne sequences reconstruct the tRNA, thereby masking the effects of phage insertion. Lacking this gene-specific option, random insertion of DNA will have minimal impact when the host genes that are disrupted upon integration in to the chromosome are of little or no value. Recently acquired DNA contains a high fraction of genes that do not persist for long periods of time (Lawrence and Ochman, 1998), making them conveniently dispensable targets. Over time, less useful genes will be culled by deletion, but shortly following acquisition they likely provide the bulk of the available sites for further insertions. As a result, higher rates of foreign gene integration can facilitate further gene acquisition, a relationship resembling a positive feedback loop.

1.4.2. Promotion of Genome Rearrangement

Just as newly acquired genes may allow for additional insertions, these sequences may also promote intragenomic rearrangements such as translocations and inversion. Just as they may act as neutral sites for insertion, they may act as places where inversion and translocations will be minimally disruptive. But beyond acting passively, newly acquired DNA may play a more active role in genomic rearrangements in two ways. First, they may provide catalysts in the form of integrases and transposases whose expression leads to genomic rearrangement. Second, they may provide more active substrates for recombination in the form of repeated DNA sequences. The genomes of closely related species of *Bordetella* illustrate the potential for genomic rearrangement (Parkhill *et al.*, 2003). *B. pertussis* and *B. parapertussis* are the causative agents of whooping cough, while *B. bronchiseptica* is implicated in more benign bronchial infections. Therefore all three organisms are pathogens, although *B. pertussis* has increased its virulence. Whereas the genomes of *B. bronchiseptica* and *B. parapertussis* are largely syntenic,

the *B. pertussis* genome shows a remarkable degree of rearrangement relative to its sibling species (Parkhill *et al.*, 2003). In addition, the *B. pertussis* genome has a large number of transposable elements; these IS elements may have facilitated the formation of inversions and translocations either directly, via transposition, or indirectly by providing sites of action for homologous recombination. While other factors – such as decreases in population size – may have contributed to the inability of *B. pertussis* to counter-select the accumulation of transposons in its genome, it is clear that the presence of IS elements can have a large impact on genome structure.

1.4.3. Effecting Recombination Interference

Following their introduction, laterally transferred genes will begin to evolve according to the directional mutations pressures of their recipient genome (Lobry and Sueoka, 2002; Sueoka, 1988, 1992). But foreign genes and native genes do not evolve independently. Closely related strains of bacteria can exchange genes by homologous recombination, a process termed allelic replacement. Here, sequences that retain a high degree of similarity may recombine if mechanisms of gene exchange – transduction, conjugation, or transformation – introduces the DNA into the cytoplasm of a conspecific strain. Such recombination can be beneficial within bacterial species by allowing the rapid dissemination of advantageous mutations between strains, or in allowing for the repair of deleterious mutations using functional genes from conspecific strains. The impact of recombination is not limited to simple gene conversion; depending on the physical limitations of the mechanisms of gene transfer – for example, the size of transducing particles – strains may also exchange small insertions, rearrangements or deletions. Therefore, inter-strain recombination is an active force in shaping genome composition. Despite acting on shared sequences, this process is not unaffected by acquisition of foreign DNA. As discussed above, gene acquisition will lead to a change in the organism's ecology, so that the newly altered recipient cell no longer occupies the same niche as its ancestor. Yet homologous recombination acts among all organisms whose DNA is sufficiently similar to allow for strand invasion, regardless of their underlying ecology. As populations of bacteria acquire different laterally transferred genes, they will begin to exploit sets of different niches, all the while sharing numerous genes that are not performing any niche-specific functions. Yet recombination at these shared loci will be suppressed if these recombination events affect neighboring, adaptive loci that enable the different lineages to exploit different niches. That is, some recombination events will result in poorly adapted hybrids

and will be counter-selected; the elimination of these less-fit cells from the population leads to lower rates of recombination being observed at loci flanking adaptive genes (Lawrence, 2002). This recombination interference does not affect all of the genes that are shared between newly diverging lineages; recombination will occur normally at shared loci that are not linked to sites of adaptive differences between strains because these recombinants will not suffer any fitness costs. In effect, the acquisition of foreign DNA will act to promote genetic isolation at their flanking loci, those genes which are shared between ecologically distinct strains. Eventually, complete genetic isolation will occur when there is no part of the chromosome that is not linked to an adaptive locus. As recombination rates decrease, point mutations will accumulate which impose pre-mating isolation via mismatch correction systems (Majewski and Cohan, 1999, 1998; Vulic *et al.*, 1999). One can view gene acquisition as an arbiter of bacterial speciation by catalyzing both genetic isolation and the accumulation of mutations in orthologous genes.

1.4.4. Promotion of Gene Loss

Not surprisingly – because organismal genomes are not constantly increasing in size – the gain of genes through horizontal gene transfer leads to gene loss. This occurs for three reasons. First, gene losses can be beneficial, if expression of those genes conflicts with the cell's new lifestyle. For example, losses of the *Shigella cadA* and *ompT* genes were beneficial in that expression of either gene decreases pathogenicity (Day *et al.*, 2001; Nakata *et al.*, 1993). This sort of gene loss is an inevitable consequence of the cell's changing ecological role. Second, gene loss can be neutral. The process of niche reorganization described above will render unimportant those genes which do not contribute to the cell's new lifestyle; without selection for function they will accumulate mutations and will eventually be deleted from the genome. There appears to be a strong "deletion" bias in bacterial genomes that prevents the accumulation of transposons, pseudogenes, and other segments of DNA that are not under positive selection for function (Lawrence, 2001; Mira *et al.*, 2001). As a result, bacterial genomes are information rich, with little intergenic spaces. Therefore, it is not surprising that, once selection is relaxed on some genes following that organism's exploitation of a new lifestyle, genes will be lost by deletion rather than persisting in the genome as pseudogenes. This process does not involve a targeting of specific genes for deletion. Rather, deletions that would have removed these genes from ancestral populations would have been counter-selected, preventing deleted strains from rising to high frequency. But following LGT, some deletions will no longer be detrimental, and genes will be lost. Pseudogenes appear in

high frequency only in those genomes where gene deletion has not kept pace with the rapid rate at which genes became useless, such as in the genomes of the obligate pathogen *Mycobacterium leprae* (Cole *et al.*, 2001). Last, gene loss may be detrimental, but insufficiently so to be prevented; from a population genetics standpoint, gene gain will, on average, necessitate gene loss. Beyond the loss of non-selected DNA, the cell cannot retain additional information in the form of newly acquired genes without sacrificing its ability to retain information elsewhere (Lawrence *et al.*, 1999; Lawrence, 2001). Put simply, a population of organisms can maintain only a finite amount of information under selection at any one time. When mutations in some genes are counter-selected (that is, organisms bearing mutations in these genes are eliminated from the population), then mutations must accumulate in other genes. Those mutations with the most detrimental effects will be eliminated and, by definition, effectively neutral mutations will be allowed to accumulate in their stead. This loss of information being maintained by selection can be manifested by the loss of entire genes. From a population genetic standpoint, although the functions some genes provide may be beneficial to a small degree, they will be insufficiently beneficial to prevent their loss.

1.5. IMPACT OF POPULATION SIZE

The limitations on information content are imposed by the population structure of the bacterium itself (Lawrence *et al.*, 1999; Lawrence, 2001). The exploitation of pathogenic lifestyles is often associated with a decrease in population size, with a commensurate decrease in information content. Bacteria with large population sizes and high rates of recombination can maintain large amounts of information; this large information content may be manifested in large numbers of genes. In contrast, organisms with small population sizes and small recombination rates can maintain far less information. Typically, these organisms have fewer genes as a result. As organisms transit between large and small effective population sizes, genome reduction will occur whereby the majority of genes in their genome may be lost. As expected, extreme genome reduction is seen in pathogens and symbionts with small population sizes, including *Mycoplasma* and *Buchnera* (Andersson, 2000; Andersson and Andersson, 1999b, 1999a; Razin, 1997). Here, extreme ecological specialization has led to very small effective population sizes and low rates of recombination. Because deleterious mutations are less effectively removed from the population, the numbers of genes that can be retained in the face of Muller's ratchet has decreased. Combined with the ever-present process of gene deletion, the genomes have been reduced to very small sizes. In one case, an insect symbiont has been reduced to