DEFINING THE CONSTRAINTS ON MICROBIAL EVOLUTION VIA HORIZONTAL GENE TRANSFER: UNCOVERING THE ROLES OF PROTEIN COMPLEXITY, FUNCTION AND DIVERGENCE.

Artur Romanchuk

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology.

Chapel Hill
2014

Approved by:

Corbin D. Jones

Christina L. Burch

Mara Duncan

Charles Mitchell

John Rawls

ABSTRACT

Artur Romanchuk: Defining the Constraints on Microbial Evolution via Horizontal Gene Transfer: Uncovering the Roles of Protein Complexity, Function and Divergence
(Under the direction of Corbin D. Jones and Christina L. Burch)

Although much of the observed prokaryotic diversity on Earth is the product of incremental accumulation of beneficial mutations with small phenotypic effects, horizontal transfer of whole genes (HGT) has shaped adaptive prokaryotic evolution as well. Understanding the evolutionary and cellular mechanisms governing horizontal gene transfer (HGT) is critical for predicting how bacterial genomes and phenotypes evolve. Experiments were used to investigate the covariance between changes in *Escherichia coli* (*E.coli*) fitness caused by horizontal gene transfer (HGT) of exogenously expressed genes and the number and complexity of the protein connectivity of the manipulated genes. Prior work investigating natural HGT among multiple bacterial taxa in combination with network theory, which provides a theoretical framework for characterizing the complexity of protein interaction (protein connectivity), shows that not all genes are transferred between bacteria equally. Surprisingly the underrepresented genes tend to occupy highly connected network positions. The manipulative experiments presented here tested the hypothesis that the protein connectivity in the recipient genome profoundly influences HGT. First, changes in relative fitness were measured in a pooled population created from approximately 4122 *E.coli* cell lineages, each of which expressed a single different *E.coli* gene transferred via an HGT. This work showed that the covariance between the protein connectivity complexity and gene transferability was more complicated than previously suggested. While the complexity of protein connectivity was important, the clustering

of those interactions and the biological function of the gene also had a significant role. In a subsequent experiment, the role of sequence divergence was included into the analysis via individual fitness measurements for 178 *E.coli* cell linages each over expressing genetic homologs from *Vibrio cholera* (89 genes) and *Staphylococcus aureous* (89 genes). Surprisingly, the role of protein connectivity was insignificant when compared to the role of divergence. Finally, a study of the population dynamics of a large megaplasmid was used to illustrate the patterns and processes governing the acquisition and maintenance of large transfers of genetic material via HGT. In the conclusion, patterns of covariance between cell fitness and protein connectivity are discussed.

Dedicated to my wife Pamela Romanchuk for complete and unwavering support throughout the years and my daughter Emma Romanchuk for being the source of much joy and perspective.

ACKNOWLEDGEMENTS

Thanks to my adviser Corbin for keeping me on my toes and not letting me forget the bigger picture.

Thank you to my adviser Christina Burch for helping me maintain rigor and clarity in experimental design and writing

Thanks to my labmates, department colleagues, & associates and cats.

Thanks to my collaborator David Baltrus for making this work possible and keeping me sane and hopeful.

This research would not have been possible without funding from the Carolina Center for Genome Sciences.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS AND SYMBOLS

Bp	Base pairs
C*i*	Clustering coefficient for gene *i* (protein clustering)
d	Pairwise amino acid sequence divergence
GFP	Green fluorescent protein
HGT	Horizontal gene transfer
m	Malthusian growth parameter
ORF	Open reading frame
pi	Probability of maintaining all protein-protein interactions.
PPI	Protein-protein interaction (protein connectivity)
r	selection coefficient
Vmax	Maximum population growth velocity (post transition from lag to exponential growth)

CHAPTER ONE: INTRODUCTION

The gradualist view of adaptive evolution suggests that the observed complexity of life on Earth is the product of natural selection driving the incremental accumulation of beneficial mutations with small phenotypic effects (Ohno 1999, Ohno et al. 1968). Bacteria can be exceptions to this view as bacteria can also evolve through transfer of genetic material across species boundaries via horizontal gene transfer (HGT). HGT genes can be different from those already present in the recipient genome and thus may quickly increase genomic, proteomic, and phenotypic complexity (Matzke et al. 2014, Brembu rt al. 2013). Although HGT occurs by several known mechanisms (Skipper et al. 2013, Sun et al. 2013, Navarro et al. 2013, Gardiner et al. 2013), plasmids likely contribute strongly to the saltational nature of HGT-driven evolution as plasmids are common in bacterial populations (Clewell et al. 2014, San Millan et al. 2014, Henry et al. 2013) and plasmids can potentially transfer large amounts (>1Mb) of genetic information (Sakai et al. 2014, Althabegoiti et al. 2014, Baltrus et al. 2011).

The abrupt genetic change caused by HGT may incur greater pleiotropic costs than the nucleotide substitutions expected under the gradualist view. Put simply, adding a new gene(s) to an established network of genes may have profoundly deleterious effects. These deleterious effects can be caused by perturbing the stoichiometry of a gene network, creation of heterologous dimers, being insensitive to regulation, etc (Baltrus 2013, Starikova et al. 2012, Levin and Cornejo 2009). This idea is consistent with the seminal work of H. A. Orr who predicted a significant reduction in the rate of adaptation as the number of interaction increases (Orr 2000); Orr has dubbed this a "cost of complexity."

Direct perturbations of protein network interactions in model systems suggest that the more interactions a protein has, the more likely that a perturbation will have a high cost. For instinstance, when a subset of ribosomal genes (that typically have many interactions) were cloned and expressed in *E. coli* on a plasmid there was visible growth retardation suggesting that expression of these genes decreased the intrinsic growth rate (Sorek et al. 2007), although some suggest otherwise (Wagner et al. 2013, Gophna and Ofran 2011). Likewise in the yeast, *S. cerevisiae,* the likelihood that removal of a protein will prove lethal correlates with the number of interactions the protein has (Lu et al. 2010, Telavera et al. 2013, Manke et al. 2005). Finally, transcript depletion of 26 highly interacting *E. coli* genes resulted in a lethal phenotype (Bergmiller et al. 2012).

Much like the direct manipulation experiments in model systems, phylogenetic investigations of natural HGT events have shown that genes with few protein interactions (low protein connectivity) undergo HGT more often than genes with many protein interactions (high protein connectivity) (Wiedenbeck and Cohan 2011, Cohen et al. 2011 Wellner et al. 2007). For example, in a study of HGT among six bacterial groups, Ravi Jain (1999) showed that highly connected ribosomal genes are transferred rarely. Similarly, in an extensive phylogenetic analysis (Cohen et al. 2011) of the whole genomes of 50 bacterial species confirmed that genes with many protein-protein interactions are underepresented among the HGT. These data suggest that transferred of highly connected genes have deleterious fitness effects.

The Complexity Hypothesis was proposed by Jain et al. (1999) to explain the observed relationship between HGT success and low numbers of protein interactions. The Complexity Hypothesis envisions a simplified HGT model where a transferred gene replaces a native homolog. If a transferred protein interacts successfully with all of the proteins nornally

interacted with by the native homolog, then there is no fitness cost to this HGT event. However it is likely that the transferred protein is somewhat divergent from the native protein in percent amino acid identity and that this divergence causes some or all protein-protein interactions to fail. Fitness cost is thus the result of failed interactions. The Complexity hypothesis predicts that gene transferability should decline with increasing protein connectivity and the rate of that decline should accelerate with increasing divergence. That is, the most divergent genes with many protein interactions are the most costly and least divergent genes with few interactions are the least costly (Rivera et al. 1998, Doolittle et al. 1999, Jain et al. 2002, Lercher and Pal 2008).

The Complexity Hypothesis also suggests that gene with specific types of functions are less likely to HGT than others. Specifically, the Complexity Hypothesis states that, *"the complexity of translational and transcriptional apparatuses, which are large complex systems with many gene interactions, is a significant factor that restricts their successful horizontal transfer rates relative to the high horizontal transfer rates observed for less complex enzymatic assemblies of a few gene products"* (Jain et al. 1999). These translational and transcriptional—a.k.a. "Informational" genes—are predicted to transfer particularly poorly. As predicted, these genes (such as the ribosomal genes described above) appear to have high fitness costs when HGT and are underrepresented among the HGT events surveyed in earlier studies (Cohen et al. 2011, Wellner et al. 2007).

The Complexity Hypothesis has received some criticism. Historical data is always colored by the exigencies of the examples analyzed and it is often difficult to provide a clear hypothesis test. For instance, tests of the Complexity Hypothesis using comparative data often lump the effects of genetic diversity between genomes participating in HGT with the biological

function of the transferred protein with protein interaction complexity within the bacterial cell (Cohen et al. 2011). Additionally these studies of the Complexity Hypothesis offer little insight into how the arrangement and timing of protein interactions influences HGT success. Most notably there is a lack of understanding whether high protein connectivity, the way in which protein interactions are arranged (protein clustering), or whether interactions are permanent or temporal is responsible for the observed evolutionary trends (Gelperin et al. 2005).

I adapted network theory and developed an experimental evolution approach to investigate how HGT success is affected by protein connectivity in addition to testing some aspects of the Complexity Hypothesis. The network approach is useful because it is a clear mathematical abstraction that can provide insight into fundamental organizational principles of organismal genetics and metabolism. The network approach allows systematic quantification for each peptide of it's number of protein-ptotein interactions, termed connectivity, and of the arrangement of those interactions, termed clustering. Considering HGT the most important question is whether connectivity or clustering reliably influences fitness, and if so, by how much. Answers to these questions determine whether genetic interaction networks constrain or promote evolution towards increasing complexity.

My experimental approach measured the fitness consequences of three aspects of HGT. First, I quantified the relationship between transferability and connectivity among 4122 non-divergent HGTs into *E.coli* using a novel high-throughput assay (Chapter 2). Second, I quantify the combined connectivity and divergence effect on fitness outcome of 89 HGT events by transferring genes from *V. cholera* and *S. aureous* into *E. coli* genetic background and measuring resulting changes in fitness using a traditional pairwise competition assay (Chapter 3). Third, I estimate how many genes can be transferred between genomes in a single HGT

event by studying transfer and persistence dynamics of a massive megaplasmid using a suite of experimental approaches (Chapter 4). These data suggest that the successful HGT of a gene likely depends on the complicated relationships among the protein interaction networks, divergence, and metabolic functions within genome.

REFERENCES

Baltrus, D.A., 2013. Exploring the costs of horizontal gene transfer. Trends Ecol. Evol. (Amst.) 28, 489–495. doi:10.1016/j.tree.2013.04.002

Baltrus, D.A., Nishimura, M.T., Romanchuk, A., Chang, J.H., Mukhtar, M.S., Cherkis, K., Roach, J., Grant, S.R., Jones, C.D., Dangl, J.L., 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 7, e1002132. doi:10.1371/journal.ppat.1002132

Bergmiller, T., Ackermann, M., Silander, O.K., 2012. Patterns of evolutionary conservation of essential genes correlate with their compensability. PLoS Genet. 8, e1002803. doi:10.1371/journal.pgen.1002803

Brembu, T., Winge, P., Tooming-Klunderud, A., Nederbragt, A.J., Jakobsen, K.S., Bones, A.M., 2013. The chloroplast genome of the diatom Seminavis robusta: New features introduced through multiple mechanisms of horizontal gene transfer. Mar Genomics. doi:10.1016/j.margen.2013.12.002

Clewell, D.B., Weaver, K.E., Dunny, G.M., Coque, T.M., Francia, M.V., Hayes, F., 2014. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology, in: Gilmore, M.S., Clewell, D.B., Ike, Y., Shankar, N. (Eds.), Enterococci: From Commensals to Leading Causes of Drug Resistant Infection. Massachusetts Eye and Ear Infirmary, Boston.

Cohen, O., Gophna, U., Pupko, T., 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28, 1481–1489. doi:10.1093/molbev/msq333

Doolittle, W.F., 1999. Lateral genomics. Trends Cell Biol. 9, M5–8.

Gardiner, D.M., Kazan, K., Manners, J.M., 2013. Cross-kingdom gene transfer facilitates the evolution of virulence in fungal pathogens. Plant Sci. 210, 151–158. doi:10.1016/j.plantsci.2013.06.002

Gelperin, D.M., White, M.A., Wilkinson, M.L., Kon, Y., Kung, L.A., Wise, K.J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H., Gerstein, M., Dumont, M.E., Phizicky, E.M., Snyder, M., Grayhack, E.J., 2005. Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev. 19, 2816–2826. doi:10.1101/gad.1362105

Gophna, U., Ofran, Y., 2011. Lateral acquisition of genes is affected by the friendliness of their products. Proc. Natl. Acad. Sci. U.S.A. 108, 343–348. doi:10.1073/pnas.1009775108

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U.S.A. 96, 3801–3806.

Jain, R., Rivera, M.C., Moore, J.E., Lake, J.A., 2002a. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61, 489–495.

Jain, R., Rivera, M.C., Moore, J.E., Lake, J.A., 2002b. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61, 489–495.

Lercher, M.J., Pál, C., 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25, 559–567. doi:10.1093/molbev/msm283

Lu, C., Hu, X., Wang, G., Leach, L.J., Yang, S., Kearsey, M.J., Luo, Z.W., 2010. Why do essential proteins tend to be clustered in the yeast interactome network? Mol Biosyst 6, 871–877. doi:10.1039/b921069e

Manke, T., Demetrius, L., Vingron, M., 2005. Lethality and entropy of protein interaction networks. Genome Inform 16, 159–163.

Matzke, N.J., Shih, P.M., Kerfeld, C.A., 2014. Bayesian analysis of congruence of core genes in prochlorococcus and synechococcus and implications on horizontal gene transfer. PLoS ONE 9, e85103. doi:10.1371/journal.pone.0085103

Ohno, S., 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin. Cell Dev. Biol. 10, 517–522. doi:10.1006/scdb.1999.0332

Ohno, S., Wolf, U., Atkin, N.B., 1968. Evolution from fish to mammals by gene duplication. Hereditas 59, 169–187.

Orr, H.A., 2000. Adaptation and the cost of complexity. Evolution 54, 13–20.

Sakai, Y., Ogawa, N., Shimomura, Y., Fujii, T., 2014. A 2,4-dichlorophenoxyacetic acid degradation plasmid pM7012 discloses distribution of an unclassified megaplasmid group across bacterial species. Microbiology (Reading, Engl.) 160, 525–536. doi:10.1099/mic.0.074369-0

Skipper, K.A., Andersen, P.R., Sharma, N., Mikkelsen, J.G., 2013. DNA transposon-based gene vehicles - scenes from an evolutionary drive. J. Biomed. Sci. 20, 92. doi:10.1186/1423-0127-20-92

Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., Rubin, E.M., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318, 1449–1452. doi:10.1126/science.1147112

Starikova, I., Harms, K., Haugen, P., Lunde, T.T.M., Primicerio, R., Samuelsen, Ø., Nielsen, K.M., Johnsen, P.J., 2012a. A trade-off between the fitness cost of functional integrases and long-term stability of integrons. PLoS Pathog. 8, e1003043. doi:10.1371/journal.ppat.1003043

Starikova, I., Harms, K., Haugen, P., Lunde, T.T.M., Primicerio, R., Samuelsen, Ø., Nielsen, K.M., Johnsen, P.J., 2012b. A trade-off between the fitness cost of functional integrases and long-term stability of integrons. PLoS Pathog. 8, e1003043. doi:10.1371/journal.ppat.1003043

Sun, D., Wang, B., Zhu, L., 2013. [Advances in molecular mechanisms of bacterial resistance caused by stress-induced transfer of resistance genes--a review]. Wei Sheng Wu Xue Bao 53, 641–647.

Talavera, D., Robertson, D.L., Lovell, S.C., 2013. The role of protein interactions in mediating essentiality and synthetic lethality. PLoS ONE 8, e62866. doi:10.1371/journal.pone.0062866

Wagner, A., Zarecki, R., Reshef, L., Gochev, C., Sorek, R., Gophna, U., Ruppin, E., 2013. Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. Proc. Natl. Acad. Sci. U.S.A. 110, 19166–19171. doi:10.1073/pnas.1312361110

Wellner, A., Lurie, M.N., Gophna, U., 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. 8, R156. doi:10.1186/gb-2007-8-8-r156

Wiedenbeck, J., Cohan, F.M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol. Rev. 35, 957–976. doi:10.1111/j.1574-6976.2011.00292.x

CHAPTER TWO: COST OF GENE OVER EXPRESSION IN *ESCHERICHIA COLI* DEPENDS ON GENE FUNCTION MORE THAN NETWORK CONNECTIVITY.

Authors: Artur Romanchuk, Christina L. Burch, Corbin D. Jones

ABSTRACT

Understanding the evolutionary and cellular mechanisms governing horizontal gene transfer (HGT) is critical for predicting how bacterial genomes and phenotypes evolve. Phylogenetic studies of microbial genome evolution suggest that the structural organisation of genetic pathways, such as interactions among proteins, biases horizontal gene transfer toward genes with few interactions. Here we investigate the contribution to this pattern to the fitness costs associated with the increased expression of transferred genes. We integrate existing data from *E. coli* describing protein-protein interactions (protein connectivity), gene function, and other gene characteristics with a novel experimental measure of the fitness costs associated with over-expression of 4102 *E.coli* genes. Expression of genes important to informational processes within the cell was more costly than that of genes in other functional categories. In contrast to the phylogenetic data, this higher cost did not result from the tendency of Informational genes to have a high number of gene interactions. Among other functional classes of genes, the strongest effect was that protein connectivity unexpectedly reduced fitness cost. Our data suggest that gene transferability, at least of low divergence homologues, may be determined more by the functional (and perhaps ecological) context of the gene than by the genomic or cellular milieu of the host genome.

INTRODUCTION

Horizontal gene transfer (HGT) is a major contributor to bacterial genome evolution (Skippington and Ragan 2011). Typically HGT contributes ~14% of the protein coding genome of a bacterial species, with some species exceeding 20% (Gorgarten et al. 2005, Nakamura et al. 2004)**.** As a result HGT generates phenotypic diversity (Hanage 2013, Guttman 1997, Baltrus 2011), drives the expansion of protein families (Treangen and Rocha 2011), and facilitates the evolution of new phenotypes (Moran and Jarvik 2010), new metabolic pathways (Soyer and Creevey 2010), and new species (Schaack et al. 2010).

Comparative studies of gene gain in bacteria suggest that the frequency with which individual genes successfully undergo HGT varies considerably (Pennisi et al. 2004, Gogarten and Townsend 2005, Lawrence 1999, Diaz-Ricci and Herna´ndez 2000, Pal et al. 2005). Two patterns are striking: 1) genes with few protein-protein interactions (PPIs—a measure of the complexity of protein connectivity) transfer more often than genes with many PPIs (Rivera et al. 1998, Doolittle 1999, Jain et al. 1999, Jaim et al. 2002, Sicheritz-Ponten and Andersson 2001, Gogarten et al. 2002, Wellner et al. 2007, Lercher and Pal 2008, Gelperin et al. 2005, Cohen, et al. 2011) and 2) genes involved in "Informational" functions like genome replication, transcription, and translation are less often transferred than non-informational genes (Ge et al. 2005). For example, Lercher and Pal (2008) investigated the interaction between genes that had naturally transferred into *E. coli* across evolutionary time and degrees of sequence divergence. They showed that genes that transferred recently have lower protein connectivity compared to non-transferred genes. From these data they concluded that HGT of genes involved in complex protein–protein interactions are rarely successfully transferred.

The Complexity Hypothesis, which posits an association between a gene's transferability and the complexity of its protein connectivity, is the leading explanation for the observed variation in HGT frequency among genes (Jain 1999). This hypothesis explains the observed pattern, the low rate of HGT by high PPI genes, by arguing that members of large interconnected protein complexes can not perform any beneficial function without their respective partner proteins, which are often not transferred during the same HGT event. Moreover, the chance that the transferred gene will engage in at least one deleterious interaction is expected to increase with its total number of interactions (Nakamura et al. 2004, Cohen et al. 2008, Hao and Golding 2008, Jain et al 1999). Informational genes, which are enriched for interactions, should therefore not be transferred often (Figure 2.1A)(Jain et al. 1999, Sorek et al. 2007). Mechanistically, Jain et al. argued that increased divergence of the transferred homologs would result in an increase chance for a deleterious interaction and the chance of these deleterious events increased with the number of interactions, although others have suggested differences in gene expression alone could lead to the observed pattern (Jain 1999, Wellner 2007).

Because measuring the complexity of interactions is key to testing the Complexity Hypothesis, recent work has used network theory to quantify the complexity of gene interactions. Jain et al (1999) did not explicitly define the complexity of gene interactions. Instead they chose two examples—the genes for thioredoxin and ribosomal protein S5—to illustrate the difference between low complexity (low protein connectivity) and high complexity (high protein connectivity) gene interactions respectively. To be more quantitative, subsequent work has used network theory to measure the complexity of interactions (Barabasi and Oltvai 2004). The network approach quantifies gene *connectivity*, the number of other proteins with

which it interacts, and *clustering* (Ci), the arrangement of those protein interactions. Clustering ranges from protein-protein interactions that form complete interconnected circles (Ci = 1) to those that are branched (0< Ci <1) or linear (Ci = 0). While clustering and connectivity are commonly used in theoretical models of gene interactions, most data driven analyses focus on connectivity, often as measured by protein-protein interaction (PPI).

Direct experimental tests of the Complexity Hypothesis are scarce and have yielded mixed results. Whereas small scale studies considering a handful of genes have shown that the fitness impact of foreign or overexpressed native genes on native complexes can be minimal (Wellner and Gophna 2008, Omer et al. 2010, Bergmiller T. 2012, see also Dykhuizen et al. 1987 and Breen et al. 2012), two larger scale studies recapitulated the negative relationship between transferability and connectivity seen in the comparative data. In a study of several dozen genes in yeast, Papp et al. (2003) found that native gene overexpression had stronger negative impacts on fitness for genes in protein complexes compared to unconnected genes. In a large survey of ~250,000 bacterial genes that could or could not be transformed into *E. coli* in laboratory, Sorek et al. (2007) suggested that toxicity to the host increased with gene dosage for genes with high connectivity but not necessarily for genes with few protein-protein interactions. They hypothesized that regardless of the species of origin or divergence the associated increased expression inhibited HGT, a result that was supported by a later computational analysis of factors affecting HGT, which concluded that gene expression level was the primary factor affecting HGT (Park and Zhang 2012).

Jain's Complexity Hypothesis remains the *de facto* explanation of biases in HGT. However, the scarcity of experimental tests makes it difficult to disentangle the role of protein connectivity from other factors that influence gene transferability. Here we describe a genome-

wide experimental test of the Complexity Hypothesis that investigates the impacts of HGT-induced increases in native gene expression in the absence of sequence or functional divergence in the transferring genes. In essence we are modeling one of the early stages of HGT: the initial introduction of a new copy of a gene into the host genome. We then determine the potential fitness cost of this new copy and see how that cost relates to gene function and protein connectivity. To take a broad measure of the complexity of protein connectivity, we consider both PPI as well as regulatory interactions and quantify both types of interactions using the connectivity and clustering metrics. Consistent with the Complexity Hypothesis we show that Informational genes have a high cost when HGT; contrary to the Complexity Hypothesis we show that this cost is largely independent of the complexity of any measure of protein connectivity. For other types of gene functions, we show that protein connectivity partially predicts the fitness effect of HGT for each transferred gene, but not in the direction predicted by the Complexity Hypothesis.

MATERIAL AND METHODS

Bacterial Strains and Plasmids

We used ASKA *E.coli* duplicate collection (Kitagawa et al. 2005), which consists of 4122 strains of *E. coli* that have each been transformed with a plasmid expressing a different *E. coli* gene under an IPTG inducible promoter. The parent *E. coli* strain AG1 W3110 (F- *recA, endA, gyrA, thi, hsdR*(r_k^- m_k^+)*, supE, relA*) is a recombination-suppressing derivative of *E. coli* MG1655 (Kitagawa et al. 2005). The expression plasmid used to generate the collection was pCA24N, a 5kb non-conjugative plasmid that drives expression of the cloned ORF from an IPTG-inducible promoter that is strictly repressed by the lacI(q) gene product. pCA24N also

encodes chloramphenicol resistance. The 4122 cloned genes all have the exact sequence of the gene encoded on the *E. coli* AG1 W3110 chromosome.

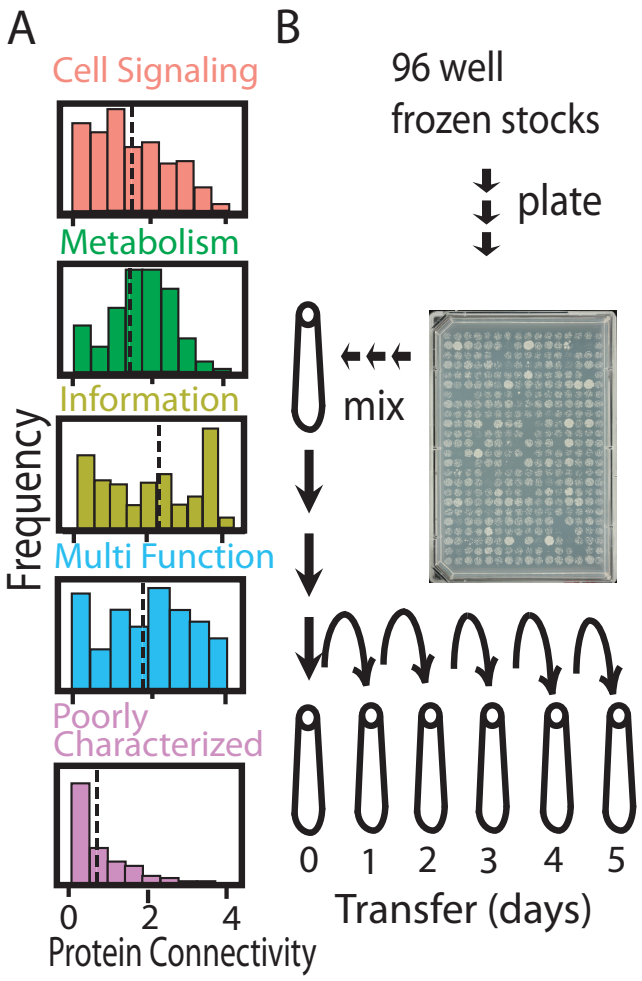

Figure 2.1. Distributions of the protein connectivity of different functional classes of *Escherichia coli* genes used in this experiment and the experimental design used to assay their relative cost. A) Histograms showing medians and distributions of protein connectivity (x-axis log scale number of protein-protein interactions) for genes from each cellular function category (pink – Cell Signaling, green – Metabolism, light green – Informational, blue – Multi Function, purple – Poorly Characterized). Informational genes have the highest median protein connectivity. B) Experimental design showing sample preparation, mixing, propagation and experimental timeline. All ASKA (4122) HGT variant lines were plated by the vendor on multiple agar plates using colony replicators. All colonies were pooled together into a common mix and each replicate population was started with an inoculum from the common mix. Each replicate population was propagated every 24 hours for five days via serial dilution. Population samples from days one, three and five were sequenced to estimate relative frequencies of each HGT variant strains at each time point.

Generation of the Mutant Pool

The 4122 ASKA strains shipped as 4122 colonies that had been spotted onto 57 LB agar plates using a 96-pin replicator tool. These colonies were kept on ice or at 4°C for less than one week while we created local archives of the individual strains by stabbing the colonies with a 96-pin replicator. Colonies from individual plates were then harvested by suspending in 5 ml of LB + 30 ng/mL Chloramphenicol. The resulting cell suspensions were mixed to yield a total volume of 300 mL, the mixture was centrifuged for 10 minutes at 10K rpm, and the supernatant was decanted. The cell pellet was resuspended in 300 mL of 60% LB + 40% glycerol and divided into 1.5 mL aliquots that were archived at -80°C.

Pooled Competition

Bulk competition experiments were conducted using a serial transfer protocol as described in Travisano et al. (1995). Experiments were initiated by transferring a volume of 100 µL of the mutant pool (described above) into 10 mL of liquid growth media (Luria Broth supplemented with 0.1 mM IPTG to induce gene expression and 30 ng/mL Chloramphenicol to maintain plasmid presence, time1: 100 uL from t_0 freezer into 10mL of broth, 100 uL has 3.68 x 10^5 cells and so 3.68x10^7 cells in the initial inoculum. 100 uL from t_1 incubated flask has 3.62 x 10^7 cells). The resulting culture was incubated with shaking for 24 hours at 37ºC, at which time 100 µL was diluted into 10 mL fresh growth media. This cycle of growth followed by dilution into fresh media was repeated 5 times, and resulted in approximately 6.7 generations per cycle. After each 24 hour incubation, multiple 1.5 mL aliquots were mixed with 0.5 mL glycerol and archived at -20ºC for later analysis. Eight replicate bulk competition experiments

were conducted in this manner. Of the 4122 ASKA strains, 20 failed to grow sufficiently and were eliminated from the experiment.

Plasmid DNA Extraction and Sequencing

Total pCA24N plasmid DNA was extracted from 2 mL of individual archived cultures using the Fermentas Plasmid miniprep kit (Thermo Scientific Waltham, MA kit#K0502). Plasmid was extracted from each of the eight (8) replicate competition experiment after the 1st, 3rd, and 5th serial transfers, yielding a total of 24 samples. These samples were barcoded using standard Illumina TrueSeq (www.illumina.com/truseq.ilmn) multiplex barcoding, pooled paired read libraries were generated and 50 bp long paired end reads were sequenced using Illumina HiSeq 2000 sequencing at the UNC High throughput Sequencing Facility. We obtained ~16.6 million reads per sample.

Mapping Sequencing Reads

Sequencing reads were mapped using Burrows-Wheeler Aligner version 0.7.7. (BWA, Li and Durbin 2009) and SAMtools version v1.3 (Li et al 2009). Our analysis identified read pairs that could be unambiguously mapped to a plasmid rather than the *E. coli* chromosome. Read pairs were mapped using full length, both pair, exact match BWA parameters. We conducted three independent alignments: to the *E. coli* W3110 genome, the pCA24N plasmid backbone, and the complete set of ASKA gene-plasmid junctions (ASKA genes flanked by 500bp of plasmid sequence both up- and down- stream). Two categories of pairs unambiguously mapped to a plasmid: 1) Pairs in which one read mapped to an ASKA gene in the *E. coli* chromosome and the other read mapped to the pCA24N plasmid backbone; and 2)

Pairs in which one read mapped to the region spanning 25bp on either side of a gene-plasmid junction. Reads that mapped to *E. coli* genes absent from the ASKA collection were unambiguously identified as chromosomal contamination (this fraction was ~10%).

Statistical analysis

All statistical analyses were conducted in R (version 3.0.0). Survival analyses were conducted using the *survreg* package, specifying an accelerating failure time model with an exponential distribution. All genes were given a survival value of observed (1) or not observed (0), with no censoring. Survival was modeled as a function of the fixed effects described in Results, including a random effect of replicate. Linear modeling was conducted using the *nlme* package. For each gene, only replicates in which that gene was observed after both transfers 1 and 3 were included in the analysis. 1163 genes met these criteria in at least one replicate and, even among these genes the number of included replicates was variable. In each replicate, we calculated *ln*(fitness) of each gene as $ln(f_3/f_1)$ where f_i is the frequency after transfer *i*. *ln*(fitness) was modeled as a function of the fixed effects specified in Results and a random effect of replicate, weighting the contribution of each gene by the number of replicates in which it was observed after both transfers 1 and 3.

RESULTS

Our experimental design (Figure 2.1B) made use of the ASKA library of 4122 strains of *E. coli,* each transformed with a plasmid capable of expressing a different *E. coli* open reading frame (Kitagawa, M et al. 2005). In all strains, the 'transferred' gene on the plasmid is identical in sequence to the 'native' gene on the recipient *E. coli* chromosome. Of the

4122 strains, 20 failed to grow sufficiently and were eliminated from the experiment. We measured the fitness effects associated with expression of the transferred genes by pooling the remaining 4102 strains in approximately equal numbers and serially transferring the pool in media that induced expression of the transferred genes (Methods). Serial transfers of 100 uL into 10 mL fresh media were conducted every 24 hours for 5 days. We monitored changes in the frequencies of the transferred genes in eight replicate serial transfer experiments by deep sequencing total plasmid isolated from each replicate after days 1, 3, and 5.

Close visual inspection of the resulting data revealed several general patterns. Trajectories of individual gene frequencies over time exhibited low variance among replicates, but high variance among genes (Figures 2A and S1). Most transferred genes rapidly declined in frequency, becoming undetectable (frequency < 10^{-6}) by day 5, whereas only a few increased in frequency throughout (upper right quadrant of Figure 2.2A). Of the 4122 genes in the ASKA collection, 4102 were detected in at least one replicate on day 1, 3360 on day 3, and only 1968 on day 5. The change in gene frequency between days 1 and 3 (measured as $ln(f_3/f_1)$, where f_i is the transferred gene frequency measured at day i) was negatively correlated with the change in gene frequency between days 3 and 5 (Figure 2.2A; Pearson's ρ = -0.339, $p < 0.0001$). More specifically, genes that rapidly decrease in frequency initially tend to slow their rate of decrease later and genes that increase in frequency initially tend to decrease in frequency later. Although the former pattern was unexpected, the latter pattern is expected to result from the increase in population mean fitness that occurs as high fitness genes rapidly increase in frequency.

Examination of the highest fitness genes emphasizes the high consistency among the experimental replicates, but also reveals the major source of variance between replicates. Frequency trajectories of the 20 genes that achieved the highest average frequencies on day 5

are plotted in Figure 2.2B and Table S1. Together these genes made up the bulk of each replicate population by day five and the relative contributions of the 20 genes are consistent among 7 of the 8 replicate populations. In a single replicate, the *yacH* gene rose to an exceptionally high frequency by day 5, possibly because it started at an uncharacteristically high frequency on day 1 (Figure 2.2B, replicate 6). It is apparent that the dynamics of this gene had an impact on the dynamics of many other genes in this replicate (Figure S1).

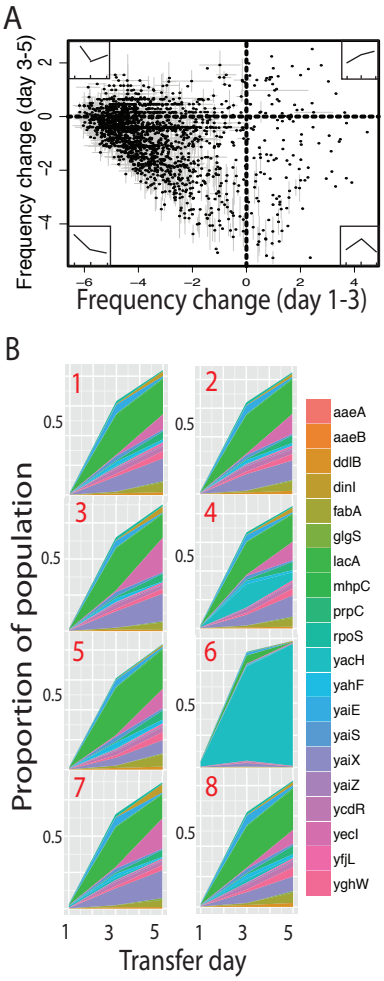

Figure 2.2. Most genes are deleterious and quickly become undetectable, although a consistent minority of tested genes increase in frequency over the competition period. A) A global view of gene frequency change between day 1/3 and day 3/5 with inset illustrative gene trajectory diagrams. Most HGT variant strains decrease in frequency rapidly between day one and three, but slow the rate of decrease between day three and five (top and bottom left). A small fraction of HGT variant strains increased in frequency over time (top right). Finally, a few genes briefly rise in frequency, but then drop in frequency (bottom right). B) Within replicate change in relative proportions of the top 20 HGT variant strains over the three days is consistent among replicates with the exception of replicate six. Each sub-panel represents one of the eight replicate populations. All of the pictured genes belong to the poorly characterized cell function category except *fabA*, which is a metabolic gene.

We used a survival analysis to investigate the mechanistic basis of fitness differences among genes in our experiment. The survival analysis allowed us to include all of the genes in our analysis even though many were undetectable (i.e., were lost from the population) on days 3 and/or 5. Thus the survival analysis identifies across the whole experiment the parameters that affected the rate and probability of gene loss. The full survival model considered several predictor variables: gene function, gene length, position in operon, expression level, and connectivity and clustering for both protein-protein and regulatory interactions (Table S2 and S3).

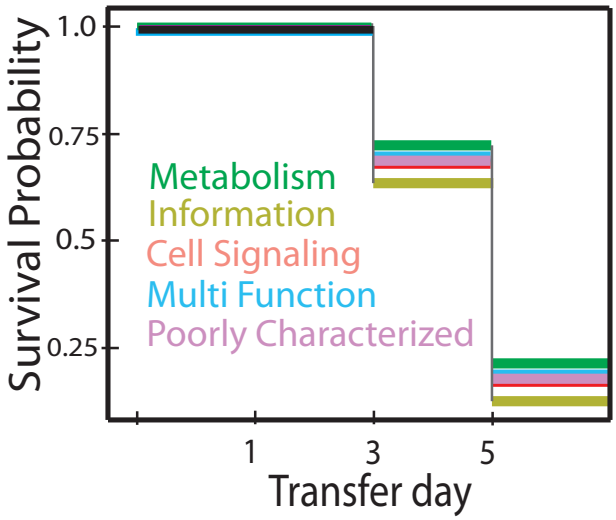

Figure 2.3. Information genes have the poorest survival. Survival analysis shows high fitness cost of Information genes compared to Metabolic genes. Colors are as in figure 2.1(pink – Cell Signaling, green – Metabolism, light green – Information, blue – Multi Function, purple – Poorly Characterized).

We divided gene function into five broad categories based on their annotated function: cell signaling, informational, metabolic, multi-functional, and poorly characterized (Table S3). Table 2.1 shows the analysis of deviance table corresponding to this full model. Among the predictor variables we considered, gene function had the largest impact on survival (Deviance = $-2 \Delta \ln L = 86.89$, df = 4, $p = 6 \times 10^{-18}$), with Metabolic and Informational genes showing the

highest and lowest survival, respectively (Figure 3.2). Protein connectivity had a significantly positive effect, and regulatory connectivity had a significantly negative effect. In addition, gene length and position in operon (*i.e.* relative order within the operon) had statistically significant negative effects on gene survival.

Table 2.1. Analysis of Deviance Table for the Full Survival Model.

Parameter	**Estimate**	**df**	**Deviance**	**p-value**
Gene function		4	86.89001	6.02E-18
Information	-0.121			
Metabolism	0.102			
Multiple	0.0374			
Poorly Characterized	0.0102			
Gene length	-3.3E-05	1	3.905165	0.048138
Position in operon	-0.013	1	10.38915	0.001268
Expression level	-5.6E-06	1	1.534769	0.215398
Protein connectivity	0.00283	1	13.54742	0.000233
Protein clustering	-0.0186	1	0.599019	0.438953
Regulation connectivity	-0.001	1	3.843957	0.049926
Regulation clustering	0.0523	1	3.7948	0.051412

The significant contributions of gene function and protein and regulatory connectivity motivated a closer investigation of the predictors most directly involved in the Complexity Hypothesis, including the interactions between them. Using a survival analysis on a reduced model, we estimated the effects of gene function, connectivity, clustering and all two and three-way interactions between them separately for protein (Table S4) and regulatory interactions (Table S5). Both analyses reveal significant effects of parameter interactions. With respect to the Complexity Hypothesis, several important patterns are apparent in the data (Figure 2.4). Informational genes were more costly as before (*i.e*. had lower survival probabilities) than genes in other functional categories, but the higher cost did not result from the tendency of

informational genes to be highly connected. That is, Informational genes had high costs regardless of their protein or regulatory connectivity or clustering. Among the other functional categories, the strongest network effect was an unexpected positive effect of protein connectivity on survival. Only protein clustering and regulatory connectivity had the expected negative effects on survival.

Table 2.2. Full Linear Model: Effects on Relative Fitness.

Parameter	Estimate	df	denDF	F-value	p-value	
(Intercept)	-3.1626	1	4448	959.33	<.0001	***
Protein connectivity	-0.0175	1	4448	25.84	<.0001	***
Protein clustering	0.1907	1	4448	1.14	0.2857	
Gene function		4	4448	16.94	<.0001	***
Information	0.7965					
Metabolism	-0.0219					
Multiple	-0.3002					
Poorly Characterized	0.3007					
Gene length	0.0001	1	4448	3.38	0.0662	
Position in operon	-0.0071	1	4448	0.09	0.7608	
Regulation connectivity	0.0071	1	4448	6.07	0.0138	*
Regulation clustering	0.0911	1	4448	32.09	<.0001	***
Expression level	0	1	4448	0.39	0.531	

As suggested in Figure 2.2, the dynamics of individual genes are more complicated than the decline in frequency to extinction modeled in the survival analysis. We investigated these dynamics by examining the effects of the transferred gene characteristics on relative fitness among only the 1164 genes (28% of the total) which are detected at day 3 in at least one replicate. We measured fitness relative to the population mean as $ln(f_3/f_1)$, where f_i is the transferred gene frequency measured at day i. We used a linear model to investigate the effects of gene and network characteristics on relative fitness (Table 2.2). In this analysis, gene function and the network statistics protein connectivity, regulatory connectivity, and regulatory clustering had statistically significant effects on relative fitness. More focused examinations of

gene function, connectivity, clustering and their two- and three-way interactions (Tables S6) revealed different patterns than were apparent from the survival analysis. Among the surviving genes, informational genes had higher average fitness than genes in other functional categories (Figure 2.5A). Connectivity had the expected negative effect on fitness (Figure 2.5B). Finally, clustering had a strong but variable effect that depended on gene function and on the type of gene interaction, protein or regulatory (Figure 2.5B).

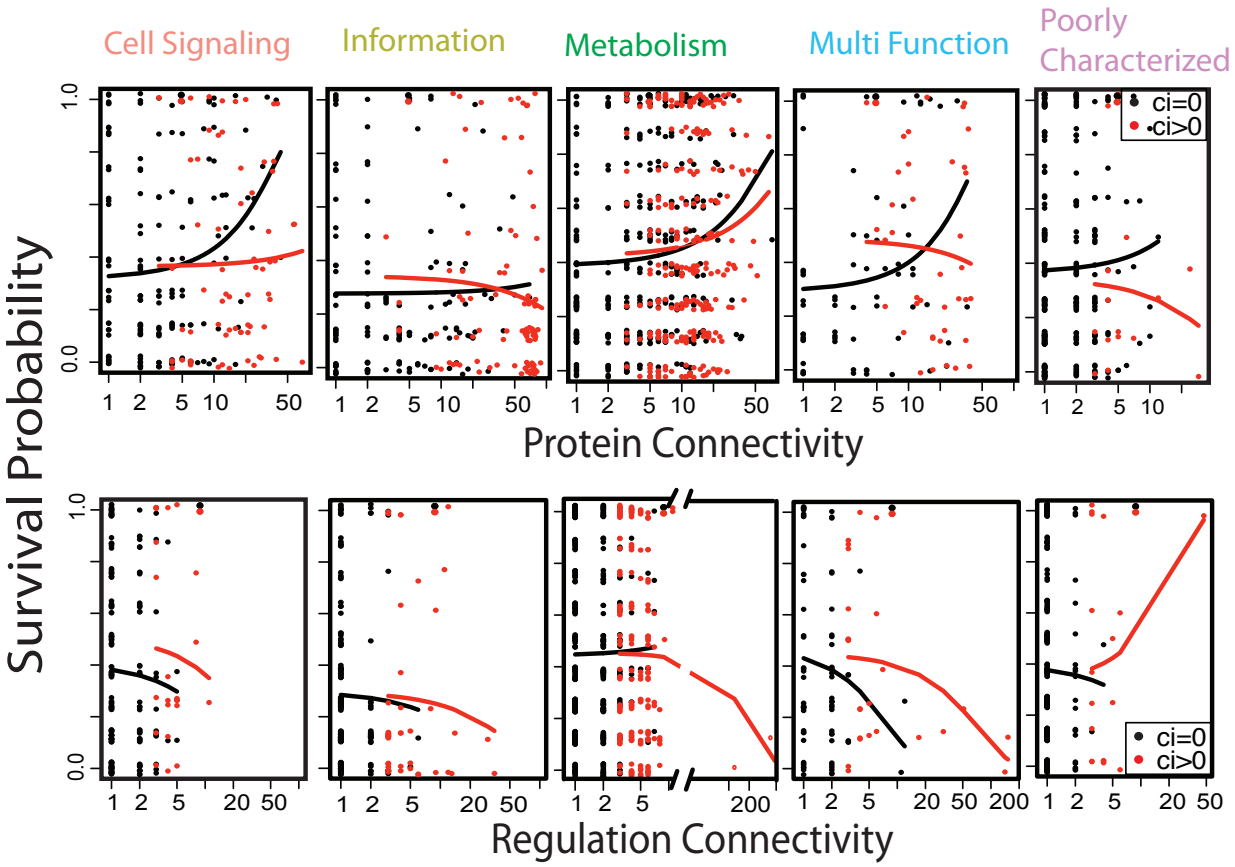

Figure 2.4. Unlike genes from all other functional categories, Informational gene survival is independent of protein and regulatory network connectivity and clustering. Surprisingly high protein connectivity is either not deleterious or beneficial for many non-Informational genes. Scatter plots show survival probability (percent survival as measured by the number of replicates in which gene frequency was > 0 out of eight total replicates) vs. protein and regulation connectivity (log scale plots of linear model regression to show all data points). For all but Informational genes the probability of survival correlates with increased number of protein interactions; probability of survival correlates decreased by clustering (top panel). Increasing regulation connectivity negatively affects fitness across all gene functions (bottom panel).

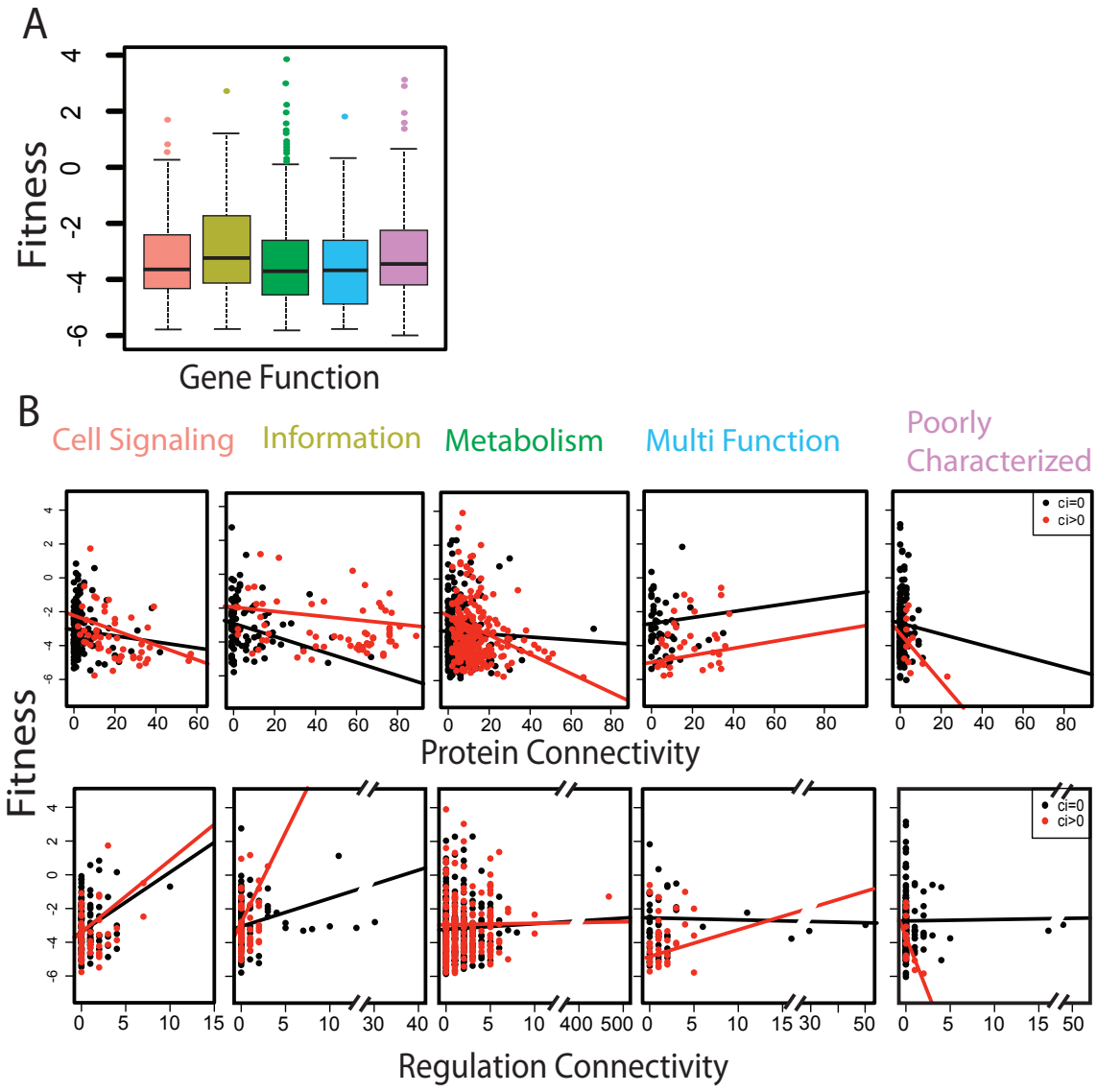

Figure 2.5. Among the surviving genes the relationship between relative fitness and protein connectivity supports the Complexity Hypothesis for all gene function classes. A) Boxplot shows median fitness (intrinsic growth rate $\ln(f_3/f_1)$ where f_t is gene frequency at time point t) for genes from each functional category. Surprisingly, Informational genes have the highest intrinsic growth rate. B) Increase in regulation connectivity and clustering lowers fitness costs of gene over expression. Scatter plots of means and standard error for fitted regressions of HGT variant strains intrinsic growth rate for two clustering classes ($Ci = 0$ and $Ci > 0$) across eight replicates versus number protein-protein interaction among genes (top) and the number of gene regulatory interactions among genes (bottom).

DISCUSSION

Jain et al's (1999) Complexity Hypothesis is commonly invoked to explain why gene function and complexity of gene interactions, typically measured in terms of protein connectivity, correlate with gene transferability during HGT (Cohen et al. 2011). While a number of studies have tested, applied, or extended this idea as genomic data has become more available (Wellner et al. 2007), the evidence showing that complexity of gene interactions is the causal mechanism limiting HGT remains mixed. Is complexity of interactions the cause or a correlate of genes that are hard to HGT?

Here we developed a simplified experimental model of a key early stage of HGT—the initial introduction of a gene into a new host—to dissect this question. We used a bulk competition experiment to directly compare the fitness effects of expressing 4102 non-divergent homologs transferred to an *E. coli* host. As non-divergent homologs are capable of making all normal protein-protein interactions, our data explore the effects of altered regulation rather than altered physical interactions. Qualitatively, our data concur with several earlier observations, *e.g*. that Informational genes are more costly, but our results also show that the relationship between the complexity of gene interactions and HGT gene fitness during this early stage is more nuanced than initially suggested.

Survival analysis shows that exogenous expression of Informational genes decreases fitness the most and that exogenous expression of Metabolic genes decreases fitness the least. This result is consistent with Jain et al. (1999) and several other studies showing that Informational genes, which typically have complex interactions, seldom HGT. Surprisingly, this pattern emerged even though we used non-divergent homologs in our study. Although

defined more broadly now, the Complexity Hypothesis as originally defined by Jain et al. incorporated both the number of gene interactions *and* the amount sequence divergence between the HGT gene and its homologs in the host genome. Divergent homologs were thought to present a greater risk of a deleterious interaction. Our data suggest that the divergence component is not required to recapitulate the bias against Informational genes. Over-expression may be enough to make an Informational gene relatively unfit compared to other gene classes.

Two other unanticipated patterns emerged from our results: (1) the fitness cost of Informational genes did not covary with protein connectivity and (2) in all other functional classes of genes there was a significant *positive* correlation between fitness and protein connectivity. Other genomic variables were also associated with differences in fitness in our survival analysis, but their effect was relatively minor. The both patterns are at odds with the *sensu lato* definition of Complexity Hypothesis—increased complexity of gene interactions should result in decreased fitness. We observed the opposite. These patterns, however, could be reconciled with the Complexity Hypothesis *sensu stricto* if divergence between homologs is a strong determinate of deleteriousness (or if there is a strongly deleterious interaction between divergence and complexity). Our present data cannot fully address this question as our HGT genes are identical to the host's, but our data serves as a strong motivation for further experiments that do investigate the effects of divergence.

The difference between Informational genes and other functional categories in the relationship between fitness and connectivity may partially explain the low rate of Informational gene HGT. Informational genes do not gain the fitness benefit imparted by having more interactions than other functional categories (most striking for Ci = 0, Figure 2.4). Thus for an equivalent number of interactions, Informational genes gain less advantage from interactions

than other categories such as Metabolism. This effect increases as the number of interactions increases resulting in a paucity of Informational genes with a high number of interactions, which is the pattern noted by Jain et al. Our data therefore suggest that complexity of gene interactions does play a role in HGT, but that mechanism of this effect is different from that assumed under the Complexity Hypothesis.

Our survival analysis captures the dynamics of all genes across the whole experiment, but is strongly determined by those genes that go extinct. Our relative fitness analysis gives a view of the fitness trajectories of surviving genes between two time points. Here the picture is less clear than for the survival analysis: surviving Informational genes are slightly less costly and the relationship between connectivity and fitness is not significant for most functional categories, except Metabolism. The former likely reflects that the highly deleterious Informational genes are undetectable by T3. The latter result is largely a lack of power driven by the small number of high PPI genes for most functional categories at the start of the experiment and the fact that by T3 only 28% of genes are still detectable. Despite this weak power, we do see a main effect of clustering and a strong interaction between clustering, connectivity and gene function. This result, consistent with the overall survival analysis, suggests for the genes surviving to T3 that the structure of the gene interactions (*i.e*. clustering) could be as much of a driver of HGT fitness as the number of interactions (*i.e.* connectivity).

When proposing the Complexity Hypothesis, Jain et al. illustrated their point with the ribosome and thioredoxin genes**.** We revisited these examples in our analysis (Figure 2.6). The ribosome combines 56 protein coding genes and 22 RNAs into a highly connected and clustered protein interaction network (Figure 2.6A). By contrast thioredoxin genes form a simple network with few protein interactions. Under the Complexity Hypothesis, the high connectivity of the

ribosomal genes should make them more deleterious. Looking at the relative fitness of the surviving genes (Figure 2.6B) shows that ribosomal genes span a wide range of fitness values and neither clustering nor protein connectivity appear to correlate with fitness. In contrast, the thioredoxin network has lower connectivity and a lower relative fitness. Least connected thioredoxin genes (*trxA* and *trxB*) had the lowest fitness; the three most highly connected genes (*nrdA, nrdD,* and *nrdE*) had the highest fitness of genes in the thioredoxin network (Figure 2.6B). While these examples cannot capture the dynamics of the entire experiment, they do qualitatively illustrate the independence of Informational gene fitness and the positive relationship of non-Informational gene fitness with increased connectivity in our experiment.

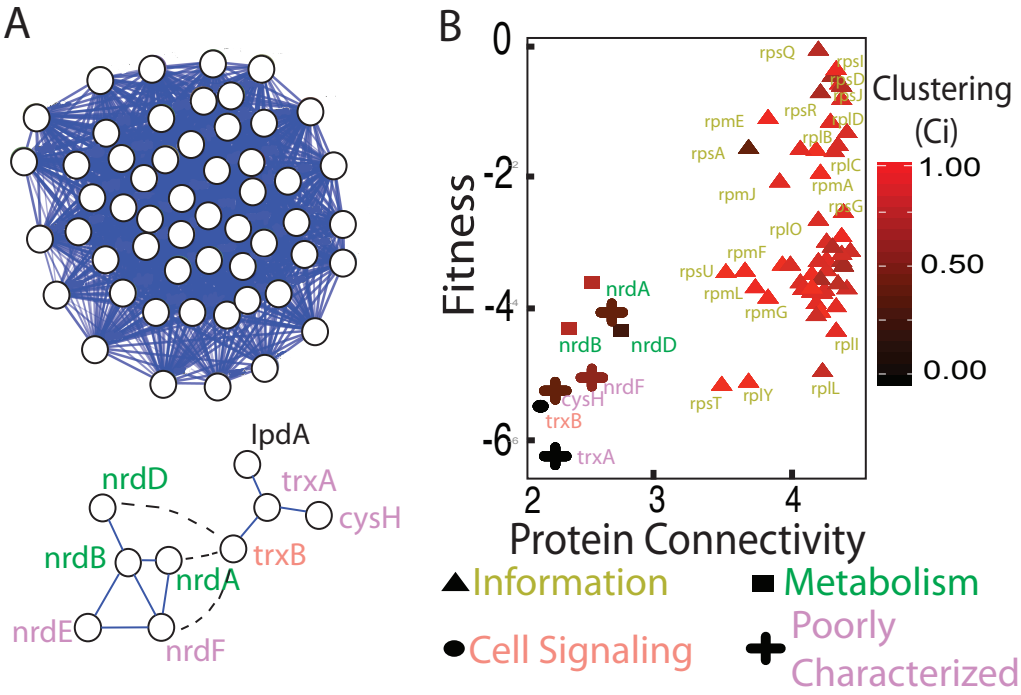

Figure 2.6. Highly connected and clustered ribosomal genes are less costly than less connected and less clustered thioredoxin network genes. A) Graph representations of ribosome and thioredoxin networks illustrate the qualitative difference in connectivity between the ribosome and thioredoxin networks. Solid lines represent protein-protein interactions and dashed lines represent metabolic interactions. B) Scatter plot of the mean relative fitness ($\ln(f_3/f_1)$, where f_t is the gene frequency at time t) across eight replicas versus protein connectivity (number protein-protein interaction between genes). Genes from the ribosome and thioredoxin network are grouped by cellular function category (symbols) and protein clustering coefficient (color gradient).

As noted above, the major difference between the work presented here and that of others is that we measure the fitness cost of non-divergent homologs. As divergence in protein sequence was conceived as the cause of broken protein-protein interactions in Jain et al.'s original formulation of the Complexity Hypothesis, our use of non-divergent homologs may partially explain our unique finding of a non-negative relationship between transferability and connectivity. Given this difference, we turn our attention to the related Balance Hypothesis (Papp et al. 2003, others Zhang et al. 2003, Calvin Bridges 1916) that was originally developed to explain differences in gene duplicability. The Balance Hypothesis does not require divergence to give rise to the negative relationship between transferability and connectivity. Instead the Balance Hypothesis predicts that HGT upsets the stoichiometric balance of protein complexes, and that the probability of stoichiometric upset increases with the connectivity. Although our data do not support this prediction in terms of protein-protein connectivity, we do see a negative relationship between regulatory connectivity and HGT fitness. This means that the more complex a gene's protein-protein interactions, the more robust that gene becomes to over-expression, but the more complex a gene's regulatory interactions, the more sensitive it is to over-expression. This observation may suggest that balanced gene expression maybe be more important for maintaining proper gene regulation than it is for maintaining proper protein-protein interactions.

Although our analysis captures two of the biological factors important to successful HGT—protein connectivity and gene function—these factors explain only a part of the total variance in gene survival in our experiment. The high unexplained variance in survival suggests that ecology, *i.e.* fit to the LB broth environment, had a large impact on gene survival in this experiment. Indeed, ecology (*i.e.* interactions between strains) provides the only explanation

for the unexpected frequency trajectories of genes that declined in frequency between days 1 and 3, but increased in frequency between days 3 and 5 (Figure 2.2A, upper left quadrant). These observations emphasize the need for conducting similar experiments in a broad array of environments (Case and Boucher 2011).

REFERENCES

Agarwal, S., Deane, C.M., Porter, M.A., Jones, N.S., 2010. Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein Interaction Networks. PLoS Comput Biol 6, e1000817. doi:10.1371/journal.pcbi.1000817

Baltrus, D.A., Nishimura, M.T., Romanchuk, A., Chang, J.H., Mukhtar, M.S., Cherkis, K., Roach, J., Grant, S.R., Jones, C.D., Dangl, J.L., 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 7, e1002132. doi:10.1371/journal.ppat.1002132

Barabási, A.-L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113. doi:10.1038/nrg1272

Barve, A., Wagner, A., 2013. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500, 203–206. doi:10.1038/nature12301

Baryshnikova, A., Costanzo, M., Myers, C.L., Andrews, B., Boone, C., 2013. Genetic interaction networks: toward an understanding of heritability. Annu Rev Genomics Hum Genet 14, 111–133. doi:10.1146/annurev-genom-082509-141730

Barzel, B., Barabási, A.-L., 2013. Universality in network dynamics. Nat Phys 9. doi:10.1038/nphys2741

Bergmiller, T., Ackermann, M., Silander, O.K., 2012. Patterns of evolutionary conservation of essential genes correlate with their compensability. PLoS Genet. 8, e1002803. doi:10.1371/journal.pgen.1002803

Bragg, J.G., Wagner, A., 2009. Protein material costs: single atoms can make an evolutionary difference. Trends Genet. 25, 5–8. doi:10.1016/j.tig.2008.10.007

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., Kondrashov, F.A., 2012. Epistasis as the primary factor in molecular evolution. Nature 490, 535–538. doi:10.1038/nature11510

Bridges, C.B., 1916. Non-Disjunction as Proof of the Chromosome Theory of Heredity. Genetics 1, 1–52.

Case, R.J., Boucher, Y., 2011. Molecular musings in microbial ecology and evolution. Biol. Direct 6, 58. doi:10.1186/1745-6150-6-58

Chang, X., Xu, T., Li, Y., Wang, K., 2013. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of /ˆdate/' and /ˆparty/' hubs. Sci. Rep. 3. doi:10.1038/srep01691

Cohen, O., Gophna, U., Pupko, T., 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28, 1481–1489. doi:10.1093/molbev/msq333

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A.H.Y., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.-C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J., Boone, C., 2010. The genetic landscape of a cell. Science 327, 425–431. doi:10.1126/science.1180823

Croucher, N.J., Harris, S.R., Grad, Y.H., Hanage, W.P., 2013a. Bacterial genomes in epidemiology--present and future. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 368, 20120202. doi:10.1098/rstb.2012.0202

Croucher, N.J., Harris, S.R., Grad, Y.H., Hanage, W.P., 2013b. Bacterial genomes in epidemiology--present and future. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 368, 20120202. doi:10.1098/rstb.2012.0202

Dahlberg, C., Chao, L., 2003a. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. Genetics 165, 1641–1649.

Dahlberg, C., Chao, L., 2003b. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. Genetics 165, 1641–1649.

Darling, A.E., Miklós, I., Ragan, M.A., 2008. Dynamics of genome rearrangement in bacterial populations. PLoS Genet. 4, e1000128. doi:10.1371/journal.pgen.1000128

Diaz Ricci, J.C., Hernández, M.E., 2000. Plasmid effects on Escherichia coli metabolism. Crit. Rev. Biotechnol. 20, 79–108. doi:10.1080/07388550008984167

Doolittle, W.F., 1999. Lateral genomics. Trends Cell Biol. 9, M5–8.

Drummond, D.A., Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. 10, 715–724. doi:10.1038/nrg2662

Dykhuizen, D.E., Dean, A.M., Hartl, D.L., 1987. Metabolic flux and fitness. Genetics 115, 25–31.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., Jensen, L.J., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–815. doi:10.1093/nar/gks1094

Ge, F., Wang, L.-S., Kim, J., 2005. The Cobweb of Life Revealed by Genome-Scale Estimates of Horizontal Gene Transfer. PLoS Biol 3, e316. doi:10.1371/journal.pbio.0030316

Geiler-Samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., Hartl, D.L., Drummond, D.A., 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc. Natl. Acad. Sci. U.S.A. 108, 680–685. doi:10.1073/pnas.1017570108

Gelperin, D.M., White, M.A., Wilkinson, M.L., Kon, Y., Kung, L.A., Wise, K.J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H., Gerstein, M., Dumont, M.E., Phizicky, E.M., Snyder, M., Grayhack, E.J., 2005. Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev. 19, 2816–2826. doi:10.1101/gad.1362105

Gilbert, C., Schaack, S., Feschotte, C., 2010. [Mobile elements jump between parasites and vertebrate hosts]. Med Sci (Paris) 26, 1025–1027. doi:10.1051/medsci/201026121025

Gogarten, J.P., Doolittle, W.F., Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238.

Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. doi:10.1038/nrmicro1204

Gout, J.-F., Kahn, D., Duret, L., Paramecium Post-Genomics Consortium, 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 6, e1000944. doi:10.1371/journal.pgen.1000944

Guttman, D.S., 1997. Recombination and clonality in natural populations of Escherichia coli. Trends Ecol. Evol. (Amst.) 12, 16–22.

Guttman, D.S., Dykhuizen, D.E., 1994. Detecting selective sweeps in naturally occurring Escherichia coli. Genetics 138, 993–1003.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., Vidal, M., 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88–93. doi:10.1038/nature02555

Hanage, W.P., 2013. Fuzzy species revisited. BMC Biol. 11, 41. doi:10.1186/1741-7007-11-41

Hao, W., Golding, G.B., 2008a. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics 9, 235. doi:10.1186/1471-2164-9-235

Hao, W., Golding, G.B., 2008b. High rates of lateral gene transfer are not due to false diagnosis of gene absence. Gene 421, 27–31. doi:10.1016/j.gene.2008.06.015

Harrison, E., Brockhurst, M.A., 2012. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 20, 262–267. doi:10.1016/j.tim.2012.04.003

Intrieri, M.C., Buiatti, M., 2001. The horizontal transfer of Agrobacterium rhizogenes genes and the evolution of the genus Nicotiana. Mol. Phylogenet. Evol. 20, 100–110. doi:10.1006/mpev.2001.0927

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U.S.A. 96, 3801–3806.

Jain, R., Rivera, M.C., Moore, J.E., Lake, J.A., 2002a. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61, 489–495.

Jain, R., Rivera, M.C., Moore, J.E., Lake, J.A., 2002b. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61, 489–495.

Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, E.V., 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol. Biol. 2, 18.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., Mori, H., 2005. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. DNA Res. 12, 291–299. doi:10.1093/dnares/dsi012

Lawrence, J.G., 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr. Opin. Microbiol. 2, 519–523.

Lercher, M.J., Pál, C., 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25, 559–567. doi:10.1093/molbev/msm283

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liu, L.-M., Liu, J.-N., Liu, Z., Yu, Z.-J., Xu, S.-Q., Yang, X.-H., Li, T., Li, S.-S., Guo, L.-D., Liu, J.-Z., 2013. Microbial communities and symbionts in the hard tick Haemaphysalis longicornis (Acari: Ixodidae) from north China. Parasit Vectors 6, 310. doi:10.1186/1756-3305-6-310

Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., Edgar, R.C., Eickhorst, T., Ley, R.E., Hugenholtz, P., Tringe, S.G., Dangl, J.L., 2012. Defining the core Arabidopsis thaliana root microbiome. Nature 488, 86–90. doi:10.1038/nature11237

Moran, N.A., Jarvik, T., 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328, 624–627. doi:10.1126/science.1187113

Nakamura, K., Haruta, S., Ueno, S., Ishii, M., Yokota, A., Igarashi, Y., 2004. Cerasibacillus quisquiliarum gen. nov., sp. nov., isolated from a semi-continuous decomposing system of kitchen refuse. Int. J. Syst. Evol. Microbiol. 54, 1063–1069. doi:10.1099/ijs.0.02883-0

Omer, S., Kovacs, A., Mazor, Y., Gophna, U., 2010. Integration of a foreign gene into a native complex does not impair fitness in an experimental model of lateral gene transfer. Mol. Biol. Evol. 27, 2441–2445. doi:10.1093/molbev/msq145

Pál, C., Papp, B., Lercher, M.J., 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat. Genet. 37, 1372–1375. doi:10.1038/ng1686

Papp, B., Pál, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197. doi:10.1038/nature01771

Park, C., Zhang, J., 2012. High expression hampers horizontal gene transfer. Genome Biol Evol 4, 523–532. doi:10.1093/gbe/evs030

Pennisi, E., 2004. Microbiology. Researchers trade insights about gene swapping. Science 305, 334–335. doi:10.1126/science.305.5682.334

Qian, W., Zhang, J., 2008. Gene dosage and gene duplicability. Genetics 179, 2319–2324. doi:10.1534/genetics.108.090936

Rivera, M.C., Jain, R., Moore, J.E., Lake, J.A., 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. U.S.A. 95, 6239–6244.

Schaack, S., Gilbert, C., Feschotte, C., 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. (Amst.) 25, 537–546. doi:10.1016/j.tree.2010.06.001

Sicheritz-Pontén, T., Andersson, S.G., 2001. A phylogenomic approach to microbial evolution. Nucleic Acids Res. 29, 545–552.

Skippington, E., Ragan, M.A., 2011. Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiol. Rev. 35, 707–735. doi:10.1111/j.1574-6976.2010.00261.x

Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., Rubin, E.M., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318, 1449–1452. doi:10.1126/science.1147112

Soyer, O.S., Creevey, C.J., 2010. Duplicate retention in signalling proteins and constraints from network dynamics. J. Evol. Biol. 23, 2410–2421. doi:10.1111/j.1420-9101.2010.02101.x

Travisano, M., Mongold, J.A., Bennett, A.F., Lenski, R.E., 1995a. Experimental tests of the roles of adaptation, chance, and history in evolution. Science 267, 87–90.

Travisano, M., Mongold, J.A., Bennett, A.F., Lenski, R.E., 1995b. Experimental tests of the roles of adaptation, chance, and history in evolution. Science 267, 87–90.

Travisano, M., Mongold, J.A., Bennett, A.F., Lenski, R.E., 1995c. Experimental tests of the roles of adaptation, chance, and history in evolution. Science 267, 87–90.

Treangen, T.J., Rocha, E.P.C., 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLoS Genet 7, e1001284. doi:10.1371/journal.pgen.1001284

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., Ziv-Ukelson, M., 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 12, R110. doi:10.1186/gb-2011-12-11-r110

Wellner, A., Gophna, U., 2008. Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. Mol. Biol. Evol. 25, 1835–1840. doi:10.1093/molbev/msn131

Wellner, A., Lurie, M.N., Gophna, U., 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. 8, R156. doi:10.1186/gb-2007-8-8-r156

Zhang, Z., Inomata, N., Yamazaki, T., Kishino, H., 2003. Evolutionary history and mode of the amylase multigene family in Drosophila. J. Mol. Evol. 57, 702–709. doi:10.1007/s00239-003-2521-7

Table S1. Genes Dominating the Population at the Last Sample Time Point, their Molecular and Cellular Functions.

Gene	Function	BRITE hierchy
ddlB	D-alanine:D-alanine ligase	D-Alanine metabolism
		Peptidoglycan biosynthesis
glgS		motility and biofilm regulator
lacA	galactoside O-acetyltransferase	Metabolic pathways
mhpC	2-hydroxy-6-ketonona-2,4-dienedioic acid hydrolase	Phenylalanine metabolism
		Microbial metabolism in diverse environments
		Degradation of aromatic compounds
prpC	2-methylcitrate synthase from propanoyl CoA	Carbohydrate metabolism
		Propanoate metabolism
rpoS	RNA polymerase, sigma S (sigma 38) factor	RNA polymerase nonessential primary-like sigma factor
yacH	predicted protein	
yahF	predicted acyl-CoA synthetase with NAD(P)-binding	
yaiE	conserved protein	
yaiS	conserved protein	
yaiX	predicted pseudogene	
yaiZ	predicted inner membrane protein, DUF2754 family	
yfjL	CP4-57 prophage; predicted protein	
yghW	predicted protein	

Table S2. Gene Network Metrics,Ttheir Definitions and Sources.

Metric	Metric definition	Source	Reference	Web link
Protein connectivity	A number of connections a node (gene) has to it's nearest neighbors (requiring confidence > 0.9 out of 1).	STRINGv9.05	Franceschini et al. 2013	http://string-db.org/
Protein clustering	Number of lnks(n) connecting several nodes (k) to central node (i) and each other Ci = 2nI/k(k–1).	STRINGv9.05	Barabási and Oltvai. 2004	http://string-db.org/
Gene function	Gene function as defined by NCBI, where 25 categories are grouped into 5 meta categories.	NCBI COG db	Tatusov RL et al. 1997	http://www.ncbi.nlm.nih.gov/COG/
Position in operon	Position from the operon transcription initiation in terms of gene order.	RegulonDBv8.0	Salgado H et al. 2012	http://regulondb.ccg.unam.mx/
Regulation connectivity	A number of regulatory interaction (regulates or regulated by) a gene has (requiring confidence > 0.9 out of 1).	RegulonDBv8.0	Salgado H et al. 2012	http://regulondb.ccg.unam.mx/
Regulation clustering	Number of lnks(n) connecting several nodes (k) to central node (i) and each other Ci = 2nI/k(k–1).	RegulonDBv8.0	Salgado H et al. 2012	http://regulondb.ccg.unam.mx/
Expression level	Relative expression level of the native *E. coli* homolog.	*E.coli* Genome Project U. Wisconsin	Allen et al. 2003	http://www.genome.wisc.edu/functional/microarray.htm#dat
Gene length	Gene length in base pairs (bp).	Genobase version 8.0 Purdue University		http://ecoli.naist.jp/GB8/search/info.jsp?id=JW0001

Table S3. NCBI Cog Classificaation System.

Gene function	COG code	Classification
Metabolism	C	Energy production and conversion
	G	Carbohydrate transport and metabolism
	E	Amino acid transport and metabolism
	F	Nucleotide transport and metabolism
	H	Coenzyme transport and metabolism
	I	Lipid transport and metabolism
	P	Inorganic ion transport and metabolism
	Q	Secondary metabolites biosynthesis, transport and catabolism
Informational	J	Translation, ribosomal structure and biogenesis
	A	RNA processing and modification
	K	Transcription
	L	Replication, recombination and repair
	B	Chromatin structure and dynamics
Cell signaling	D	Cell cycle control, cell division, chromosome partitioning
	Y	Nuclear structure
	V	Defense mechanisms
	T	Signal transduction mechanisms
	M	Cell wall/membrane/envelope biogenesis
	N	Cell motility
	Z	Cytoskeleton
	W	Extracellular structures
	U	Intracellular trafficking, secretion, and vesicular transport
	O	Posttranslational modification, protein turnover, chaperones
Poorly characterized	R	General function prediction only
	S	Function unknown
Multi Function	any	Gmrs responsible for more than one function above

Table S4. Analysis of Deviance for the Protein Connectivity Hypothesis Test Model.

	df	**Deviance**	**Resid. Df**	**-2*ln*L**	**Pr(>Chi)**
NULL	NA	NA	34175	89715.44	NA
Gene function	4	86.89002	34171	89628.55	6.02E-18
Protein connectivity	1	4.122188	34170	89624.43	4.23E-02
Protein clustering	1	1.669507	34169	89622.76	1.96E-01
Frailty(replicate)	5.999413	131.4534	34163	89491.31	6.35E-26
Gene function:Protein connectivity	4.003341	24.91981	34159	89466.39	5.24E-05
Gene function:Protein clustering	4.000502	1.881494	34155	89464.5	7.58E-01
Protein connectivity:Protein clustering	1.000099	12.7787	34154	89451.73	3.51E-04
Gene function:Protein connectivity:Protein clustering	4.001175	17.79091	34150	89433.93	1.36E-03

Table S5. Analysis of Deviance for the Regulation: Connectivity Hypothesis Test Model.

	df	Deviance	Resid. Df	$-2lnL$	Pr(>Chi)
NULL	NA	NA	34175	89715.44	NA
Gene function	4	86.89002	34171	89628.55	6.02E-18
Regulation connectivity	1	4.233315	34170	89624.32	3.96E-02
Regulation clustering	1	4.220814	34169	89620.1	3.99E-02
Frailty(rep)	5.998922	131.4534	34163	89488.64	6.34E-26
Gene function:Regulation connectivity	4.002199	16.94	34159	89471.7	1.99E-03
Gene function:Regulation clustering	3.999948	14.3285	34155	89457.37	6.32E-03
Regulation connectivity:Regulation clustering	0.999989	10.20623	34154	89447.17	1.40E-03
Gene function:Regulation connectivity:Regulation clustering	4.000791	8.43213	34150	89438.74	7.70E-02

Table S6. Analysis of Variance Table Reduced Model: Effect of Intrinsic Growth on Fitnes.

	numDF	denDF	F-value	p-value
(Intercept)	1	4453	957.3313	<.0001
Protein connectivity	1	4453	25.6245	<.0001
Protein clustering	1	4453	1.1317	0.2875
Gene function	4	4453	16.8004	<.0001
(Intercept)	1	4453	972.8706	<.0001
Regulation connectivity	1	4453	7.2749	0.007
Regulation clustering	1	4453	29.9971	<.0001
Gene function	4	4453	14.0726	<.0001

CHAPTER THREE: *VIBRIO CHOLERAE* AND *STAPHYLOCOCCUS AUREUS* HOMOLOG DIVERGENCE BUT NOT PROTEIN CONNECTIVITY PREDICT FITNESS COST OF EXOGENOUS GENE EXPRESSION IN *ESCHERICHIA COLI*.

Authors: Artur Romanchuk, Kedar Karkare, Christina L. Burch, Corbin D. Jones

ABSTRACT

The dramatic molecular and phenotypic changes that result from horizontal gene transfer (HGT) can lead to rapid and radical bacteria evolution. While the palette of potentially transferable genes is enormous, it is clear that not all genes HGT equally well. The probability of successful HGT, that is probability a specific gene successfully transfers to a new host via a plasmid, depends on how minimal the fitness cost of the newly acquired gene is. The Complexity Hypothesis was proposed to explain why some genes are more likely to have a higher cost when transferred. The Complexity Hypothesis proposes a simplified HGT model that predicts that gene transferability should decline with increased number (complexity) of gene interactions and the rate of that decline should accelerate with increasing divergence. We investigated this proposed relationship with exogenously expressed *V. cholera* and *S. aureus* gene homologs, which have diverged to different degrees from *E. coli*, in an *E. coli* background and estimated fitness relative to a reference strain. Overall the relative fitness mean for *V. cholera* genes, the least diverged of the two species, was significantly higher than the mean for *S. aureus* genes. Similarly, our explicit tests of the Complexity Hypothesis suggested that divergence was the most significant factor affecting fitness. The number of protein-protein interactions, however, did not have a strong influence, nor was there a significant interaction

between divergence and number of protein-protein interactions as expected under the Complexity Hypothesis.

INTRODUCTION

Horizontal gene transfer (HGT) has shaped prokaryotic evolution (Barve and Wagner 2013, Lawrence and Hendrickson 2003, Lindasy 2014, Morita et al.2013). The worldwide collection of genes housed within microbes is enormous as the potential for innovative gene combinations and novel phenotypes. HGT is a key mechanism for shuffling this variation across genomes and generating novel gene combinations (Williams et al. 2012, Hao et al. 2010). Evolution via HGT fine-tunes enzymatic reactions (Schonknecht et al. 2013, Schonknecht et al. 2014, Treangen and Rocha 2011) and facilitates major ecological transitions (Syvanen 2012, Keeling and Palmer 2008, Moran and Jarvik 2010, Yoshida et al. 2010, Schaack et al. 2010).

The probability of successful HGT, that is, the probability for a specific gene to be retained following transfer to a new host, depends on whether the newly acquired gene can avoid loss due to genetic drift or selection. Natural selection will purge microbial populations of genes that do not provide a benefit or cause fitness costs (Jiang et al. 2013, Knight et al. 2013, Hellweger 2013). In many cases, the fitness impact of newly acquired genes depends on the environment, *e.g.* antibiotic resistance genes that rapidly spread via HGT in clinical settings (Jansen et al. 2014, Nielsen et al. 2014, Devirgiliis et al. 2013, Juhas M 2013). Regardless of ecological context, many recently transferred genes can be energetically or physiologically costly (Park and Zhang 2012, Chou et al. 2011,) or interact poorly with their new genomic backgrounds (Avrani et al. 2011, Engelberg-Kulka and Glaser 1999). Indeed as evidence of

the latter, observational and phylogenetic studies show that proteins that participate in few protein-protein interactions HGT more frequently than proteins with many protein interactions (Noda and Barona 2013, Gophna and Ofran 2011, Cohen et al. 2011, Wellner and Gophna 2008).

The Complexity Hypothesis was proposed by Jain et al. (1999) to explain the observed relationship between HGT success and low numbers of protein interactions (low connectivity). The Complexity Hypothesis envisions a simplified HGT model where a transferred gene replaces a native homolog. If a transferred protein interacts successfully with all of the proteins normally interacted with by the native homolog, then the HGT is predicted to have no fitness cost. However it is likely that the transferred protein is somewhat divergent from the native protein in amino acid identity and that this divergence causes some or all protein-protein interactions to fail. Fitness cost is therefore the result of failed interactions. Specifically, Jain et al. argues:

> *"assume that genes for thioredoxin*[1 interaction] *and ribosomal protein S5* [6 interctions] *have been horizontally transferred to separate Escherichia hosts, that both share a similar percent- age of protein identity, and that the probability of each protein successfully making a required interaction with another gene product is 0.25. According to this simplified model, the probability that a transferred thioredoxin could successfully interact with thioredoxin reductase would be 0.25, whereas the probability that a transferred S5 could be assembled into a small subunit is* 0.25^6 *(0.00024) or about 1,000 times less"*(Jain et al. 1999).

Thus the Complexity Hypothesis predicts that gene transferability should decline with

increasing protein connectivity and the rate of that decline should accelerate with increasing divergence. That is, the most divergent genes with many protein interactions are the most costly and least divergent genes with few interactions are the least costly.

Here we experimentally test the Complexity Hypothesis. We measure the fitness cost of a set of proteins transferred into *E. coli* from *Vibrio cholerae* and *Staphylococcus aureus*, which have diverged from *E. coli* by different degrees (Figure 3.1B). We use an inducible expression plasmid to express 89 *V. cholera* and 89 *S. aureus* gene homologs of *E. coli* genes that span a wide range of protein connectivity ("connectivity"; Figure 3.1A) one at a time, in *E. coli*. We measure the effect of each expressed homolog ("tested strain") on the *E. coli* relative growth rate through pairwise competition with a genetically marked reference *E. coli* strain ("reference strain") (Figure 3.1C). By pairing functionally conserved homologs we seperate the effect of gene function, which is a known confounding factor (Chapter 2, Boto 2010, Jain 2002), from sequence divergence. Using these data we investigate how homolog divergence and protein connectivity affect fitness impact of exogenously expressing the *V. cholera* and *S. aureus* genes.

MATERIAL AND METHODS

Bacterial Strains and Plasmids

We used BEI Resources *V. cholera* and *S. aureus* genetic strain collections (http://www.beiresources.org/). The methicillin-resistant *S. aureus* collection from BEI is a COL Gateway clone set consisting of 25 plates which contain 2343 sequence validated clones from *S. aureus* strain COL cloned into *E. coli* DH10B-T1 cells. *S. aureus* open reading frames were cloned into the BEI vector pDONR221 (Invitrogen, Carlsbad, California) with a native start codon but no native stop codon (pDEST17 will provide a stop codon). The BEI *V. cholerae*

Gateway clone set consists of 46 plates which contain 3813 sequence validated clones from *V. cholerae* strain El Tor N16961 cloned into E. coli DH10B-T1 cells. *V. cholerae* open reading frames were inserted into the BEI pDONR221 vector with native start codon and stop codons. Open reading frames from the BEI collections were subcloned into Invitrogen plasmid pDEST17 for expression in Invitrogen Stellar chemically competent cells (Invitrogen Cat#636763, Carlsbad, CA). Invitrogen Stellar competent cells containing pUCBB-eGFP (pPRS3a_1 eGFP gene) was used as the reference strain in pairwise competition assays.

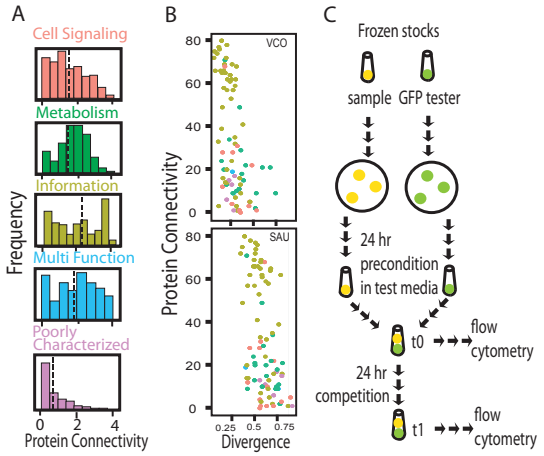

Figure 3.1. Experimental approach to test the effects of divergences and connectivity on relative fitness of transferred genes. A) Histograms showing medians and distributions of protein connectivity (x-axis number of protein- protein interactions) for genes from each cellular function category in *E.coli* (pink – Cell Signaling, green – Metabolism, light green – Information, blue – Multi Function, purple – Poorly Characterized). B) Protein divergence and connectivity of tested genes from *V. cholearai* (VCO) and *S. aureus* (SAU). C) Experimental design showing sample preparation, mixing, sampling and experimental timeline.

Cloning and Genome Manipulation

I cloned 89 homologous genes from each of the BEI (http://www.niaid.nih.gov/) *V. cholera* and *S. aureus* genetic strain collections into the Invitrogen pDEST17 expression plasmid using the Gateway system (Life technologies, Carlsbad CA, LR clonase II enzyme mix Cat#11791020) to achieve IPTG inducable expression. The pDONR221 entry clone contains recombination cloning sites, attL1 and attL2 to facilitate gene transfer into a destination vector

pDEST17, along with M13 forward and reverse priming sites for sequencing and a Kanamycin resistance gene for selection. Recombination into pDEST17 was facilitated through an attL substrate with an attR substrate (pDEST17) to create an attB-containing expression clone with an Ampicillin resistance gene for selection. Cloning LR reaction was performed in accordance with Gateway protocol. Prior to cloning pDONR221 plasmids were purified using column purification according to Qiagen mini prep protocol (Qiagen, Venlo Linburg, plasmid mini kit Cat#12123). Resulting clones were grown overnight in LB media with Ampicillin (60ug/ml). Freezer stocks were made by combining equal volumes overnight growth with 80% glycerol solution and stored at -80C.

Pairwise Competition Assays

Pairwise competition assays were conducted using the serial transfer protocol common to microbial competition experiments (*e.g.* Travisano et al. 1995). Assays were initiated by transferring a single colony from plated frozen stock into 100ul of LB media containing ampicillin (60ug/ml) and 0.1M IPTG in a single 200ul snap-cap PCR tube. The resulting culture was incubated with shaking for 24 hours at 37ºC, this served as the precondition period for both the *V. cholerae* and *S. aureus* samples and the GFP containing reference strain. At the end of the initial 24 hour precondition period each sample strain culture was mixed in equal volumetric amounts with a reference strain culture, diluted 100 fold into 100ul of fresh LB media containing Ampicillin (60ug/ml) and 0.1M IPTG. Although strains were mixed in equal volumes the resulting cell concentrations deviated from the 50:50 proportion due to fitness cost of the cloned homolog during the precondition period (most different ratio was 20% of tested strain to 80% of GFP tester strain). The resulting culture was incubated with shaking for 24

hours at 37ºC, this served as the competition period. Immediately following the preconditioning period (t_0) and the competition period (t_1), cultures were diluted 10-fold in 10mM $MgCl_2$, and the sample (eGFP-) and tester (eGFP+) cells were enumerated using flow cytometry. Relative selection rate (r) was measured as difference in the Malthusian parameter where M_t is the Malthusian parameter is the natural log ratio of sample cells/eGFP reference cells at time t as described in Travisano and Lenski (1996). Four replicate competition assays were conducted for each of the 80 *V. cholera* and *S. aureus* homologous genes.

Flow cytometry

Flow cytometry measurements were carried out on the BD Accuri C6 flow cytometer, data was collected over 3000 events under slow 8ul/s flow conditions using the Accuri C6 software. Cultures were run in a pairwise fashion, for example if the *E. coli* strain containing the *V. cholera* homolog of the rpsF gene was measured first, the *E. coli* strain containing the *S. aureus* homolog of the rpsF gene was measured next. All samples were kept at room temperature to prevent E. coli cell shrinkage and a corresponding decrease in eGFP detection efficiency.

Statistical analysis

I used a Welch's pairwise t-test and general linear model framework (likelihood ratio test ANCOVA) to examine the impact of protein divergence and connectivity on the fitness effect of HGT. All statistical analysis were conducted using the lme4 package in R (version 3.0.0). The selection rate was modeled as a function of the fixed effects specified in Results and of the random effects of gene identity and assay date.

RESULTS

Visual inspection of the resulting data revealed several general patterns. Qualitatively comparing the trajectories of individual replicates for all tested strains between t_0 and t_1 relative to the reference strain showed low variance among replicates and moderate variance among tested strains (Figure S1). Most tested genes rise in frequency by the second time point (Figure 3.2), indicating a high cost of GFP expression in the reference strain. Across all genes there was generally low variance among replicates and moderate variance among genes in log(fitness) of the tester strain relative to the reference strain (calculated as selection rate; see Methods). The overall shape of log(fitness) distributions among the sampled *S. aureus* and *V. cholera* genes were also similar (Figure 3.2B).

We investigated the mechanistic causes of fitness differences among genes using Welch two sample t-test and analysis of covariance (ANCOVA). First, we performed a two sample t-test using species as groups to determine whether coming from either a highly divergent (*S. aureus*), or a less divergent (*V. cholerae*) species (in terms of protein divergence), was a predictor of tested gene relative fitness. Figure 3.3 shows that overall the log(relative fitness) mean for *V. cholera* genes was higher than the mean for *S. aureus* genes ($t = -3.2542$, $df = 88$, $p = 0.001614$) suggesting that the source strain matters. While this result is consistent with our expectation, it does not reveal the underlying cause of the observed difference between strains.

A standard linear modeling approach to determine which genomic parameters contribute to the difference between *V. cholerae* and *S. aureus* homologs is complicated by covariance in our dataset. We used pairs of homologs that are shared across three divergent bacterial species—*E.coli, V. cholera,* and *S. aureus*—and whose function is presumably evolutionarily conserved. Considering that the *V. cholera* and *S. aureus* genes used are paired in terms of their

function, they cannot be treated independently under the constraints of a standard linear model.

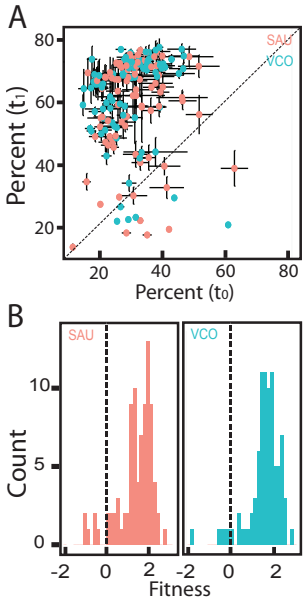

Figure 3.2. Consistently a majority of tested genes increased in density over the competition period, which suggests fitness advantage relative to reference strain. A) A global view of percent tested gene density in the competing population on day one (t_0) and day two (t_1). Most gene constructs strains carrying a tested gene comprise ~ 20% of the competition pool at t_0 and ~ 60% of the competition pool 24 hours later at t_1. A small fraction of tested strains decreased in density over time when compared to the reference strains. B) Histogram of pairwise competition fitnesses measured as relative selection rate, which is the difference of the Malthusian parameters *m* for each tested and reference strain pair (Malthusian parameter = $ln(sd_1/sd_0)$ where sd_t is strain density at time *t*). Most tested strains grow faster than reference strain and the distributions of relative fitnesses for genes sampled from *V. cholerai* and *S. aureus* are similar.

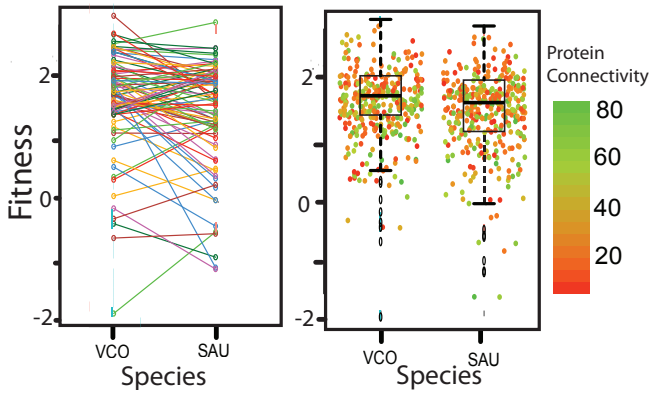

Figure 3.3. Compared to the more divergent *S. aureus, V. cholerai* genes are less costly when expressed in an *E. coli* background. (Left) Welch Two Sample t-test (t = -3.2542, df = 88, p = 0.001614) of fitness measured via selection rate (*r*) shows that overall *V. cholerae* genes increase in density significantly more than *S. aureus* genes. The positive slope of most lines connecting homologous genes indicate that on average most *V. cholerae* genes are more beneficial however there are few exception. (Right) Box and whisker plot of fitness with an overlaid scatter plot of each gene colored according to gene protein connectivity. There appears to be no obvious relationship in either species between fitness and connectivity.

Given the possible non-independence between homologs and that the comparison of difference in fitness between *V. cholerea* and *S. aureus* homologs is the true unit of our analysis, we used ANCOVA to analyze our dataset (Table 3.1). We used a series of nested models to make specific tests of the Complexity Hypothesis predictions. Beginning with the model log(fitness) = gene + day, where both gene and day are random effects, we then used likelihood ratio tests to assess the significance of sequential additions of the fixed effects divergence, connectivity, and a divergence*connectivity interaction.

Table 3.1. ANCOVA Table: Species Covariance Model Effect on Relative Fitness.

	numDF	deviance	estimate	Chi^2	p-value
divergence	1	1689.4	-0.27824	12.454	0.000417
protein connectivity	1	1676.9	0.000591	3.2677	0.07066
divergence:protein connectivity	2	1673.7	-0.00958	2.9498	0.08589

As predicted by the Complexity Hypothesis and as expected from the comparison between species, divergence had a statistically significant negative impact on fitness (Table 3.1; estimate = -0.278, df = 1, c^2 = 12.454, p = 0.00041). The effect of connectivity was near zero (estimate = 0.0006) and the interaction between divergence and connectivity was negative (estimate = -0.0096). Although the relative magnitudes and direction of the latter effect is consistent with the Complexity Hypothesis, neither connectivity nor the interaction was significant (df = 1, c^2 = 3.2677, p = 0.07066 and df = 2, c^2 = 2.9498, p = 0.08589 respectively).

We also examined the relationship between our measures of log(fitness) and Jain's measure if interaction probability (Jain et al, 1999), which attempts to capture the probability that a transferred gene will *successfully* make all of its protein-protein interactions. Here, interaction probability is calculated as

$$p_i = (1-d)^k,$$

where *d* is amino acid divergence between the transferred gene and the native homolog and *k* is the number of protein interactions made by the native gene. Contrary to Jain's prediction, p_i had no effect on log(fitness) for both *V. cholerae* genes and *S. aureus* genes (Figure 3.4B df = 1, $c^2 = 3.0274$, p = 0.08187).

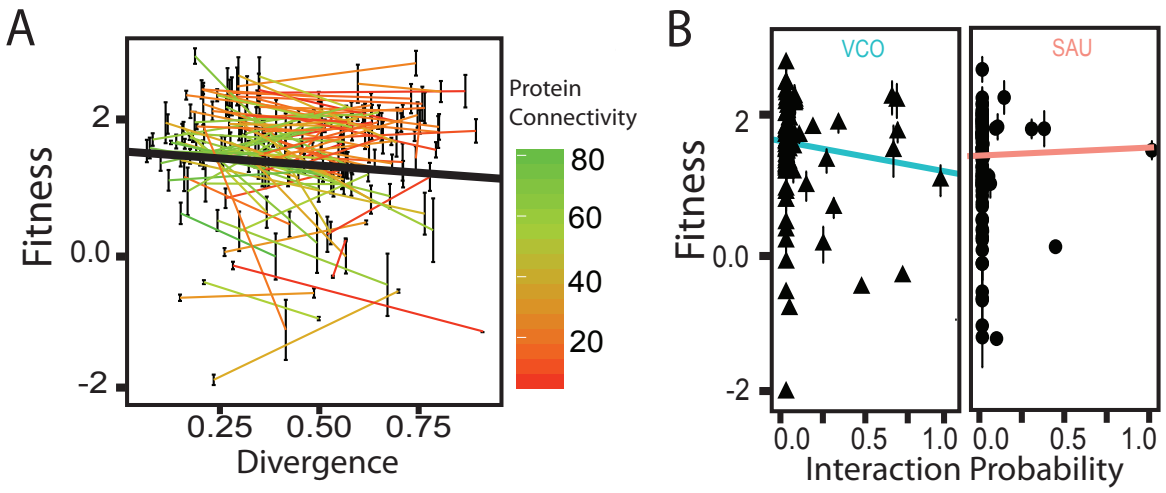

Figure 3.4. Increased protein divergence negatively covaries with fitness. A) Overall scatter plot of fitness (*r*) vs. divergence indicates fitness cost of increasing divergence. Lines connect homologous genes from *V. cholerae* and *S. aureus*. Lines are color colored in terms of protein connectivity. B) Jain et al's interaction probability (p_i) for the sampled *S. aureus* and *V. cholerae* genes compared to fitness. Contrary to the Complexity Hypothesis low interaction probability is not a good predictor of high fitness cost.

DISCUSSION

Our results are consistent with one major prediction of the Complexity Hypothesis that the cost of HGT should increase with divergence. We explicitly tested the two main predictions made by the Complexity Hypothesis—a negative relationship between divergence and HGT fitness and an interaction between divergence and protein connectivity in determining HGT fitness. Although the Complexity Hypothesis makes statements about the influence of gene function on HGT fitness, here we avoid that confounding factor by testing pairs of divergent homologs. Although our sample of tested genes is enriched for Informational genes among the most highly connected genes, our sample spans multiple

cellular functional categories. Three main patterns regarding fitness are apparent in our data: (1) the overall less divergent *V. cholerae* genes decrease fitness less. (2) On a per gene basis more divergent homologs lower fitness more than less divergent homologs regardless of species origin. (3) When we used the p_i measure proposed by Jain et al (1999) to look at within species variation in divergence and connectivity, there was no support for the Complexity Hypothesis. Although not significant experiment wide, a subset of genes hint on the interaction between divergence and protein connectivity predicted by the Complexity Hypothesis (Figure 3.5).

Our results point to differences in relative strengths of the two mechanistic processes proposed to guide HGT evolution under the Complexity Hypothesis**.** Our data show that divergence is much stronger selective agent of HGT gene fitness than connectivity. This is not surprising because amino acid divergence can influence a protein's ability to perform its respective chemical reactions, cause a protein to catalyze inappropriate substrates, and affect its interactions with of its protein partners. Which of these possible problems governs the patterns here is not revealed by our experiment, although we strongly suspect that it is a catalytic defect rather than a structural defect.

Current data suggest that genetic incompatibilities resulting from HGT are often rooted in mechanistic inefficiencies of protein chemistry as suggested by the Complexity Hypothesis (Baltrus 2013**).** Others have argued that a shift in molecular stochiometry contributes to HGT fitness cost as cells are forced to waste molecular resources and the substrate flux through genetic pathways is perturbed by exongenous expression of the newly acquired gene (Papp et al). Our data does not explicitly test this idea, but because every tested gene is exogenously

expressed at the same level, and every *E.coli* native homolog is not naturally expressed at the same level, we inadvertently sample fitness cost of altered expression.

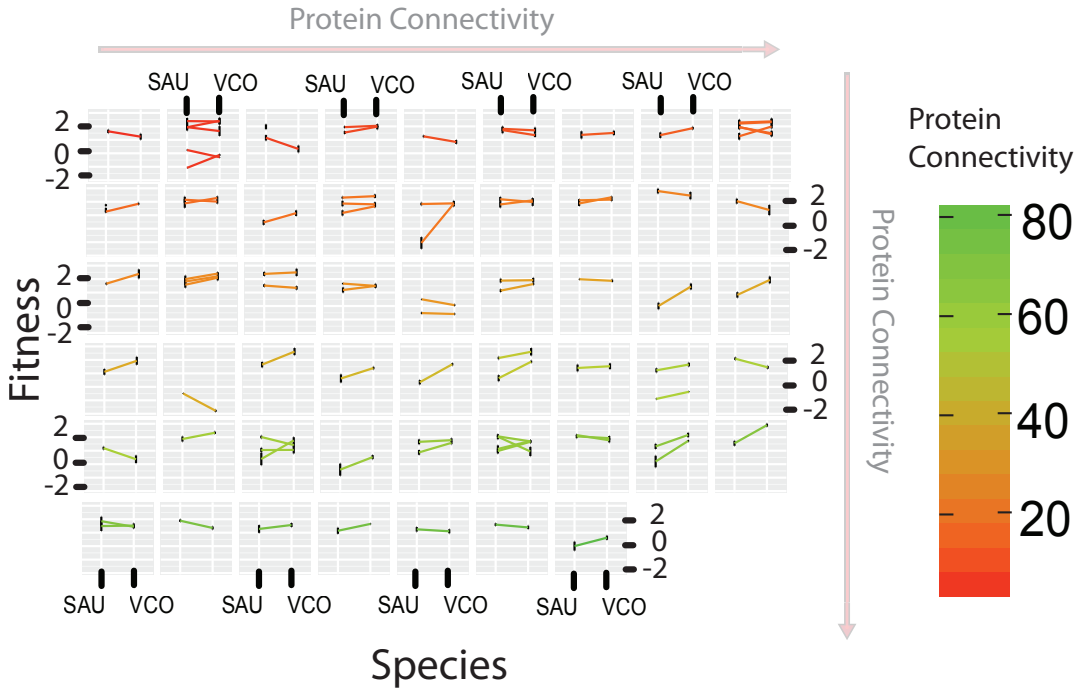

Figure 3.5. Although many genes (especially low connected) show limited interaction between fitness cost and divergence, there are examples of severe interaction as well. Plots of fitness vs. divergence for all 89 genes spanning every protein interaction category (across boxes connectivity increases from left to right and top to bottom). Lines connect homologous genes from the two examined species, and color-coding reflects the number of protein interactions for each gene.

We do not see any evidence that difference in expression level affects HGT fitness cost (df = 1, estimate = -1.484e-05 , c^2 = 1.1355, p = 0.2866). The fitness effect of overexpression appears small and similar to protein connectivity, although estimating the precise magnitude of fitness the cost associated with overexpression is likely needed.

Comparative data clearly indicates paucity of highly connected and highly divergent genes in the past HGT events, but the experimental data presented here and that of others has failed to demonstrate strong selection imposed by genetic complexity. Most studies rely on

comparative approaches because these provide evolutionary information regarding a large numbers of genes. Comparative studies are compromised by the ability to infer HGT events, incomplete sampling, and time. Thus, the evolutionary signal inferred from these studies lacks the precision compared as quantifiable selection coefficients estimated via experimental studies of HGT divergence, overexpression and protein interaction architecture . Experimental studies, however, suffer from difficulties in designing sensitive experimental approaches, which adequately explore the Complexity Hypothesis and other sources of HGT cost. Future work is needed to design better analyses and better experiments to reconcile these two divergent datasets.

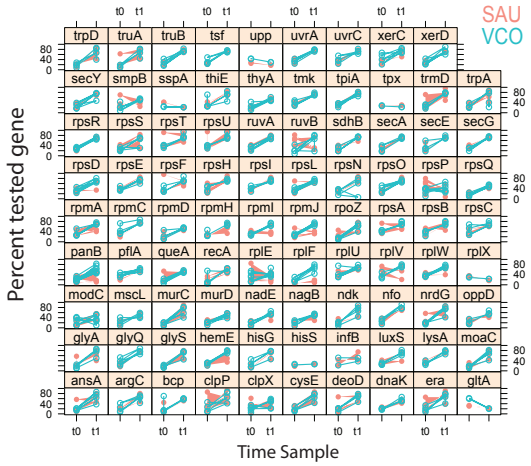

Figure S1. Within gene variance is small than between gene variance. Plots show percent tested gene trajectories over time. Most tested genes increase in frequency over time.

REFERENCES

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., Lindell, D., 2011. Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature 474, 604–608. doi:10.1038/nature10172

Baltrus, D.A., 2013. Exploring the costs of horizontal gene transfer. Trends Ecol. Evol. (Amst.) 28, 489–495. doi:10.1016/j.tree.2013.04.002

Barve, A., Wagner, A., 2013. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500, 203–206. doi:10.1038/nature12301

Boto, L., 2010. Horizontal gene transfer in evolution: facts and challenges. Proc. Biol. Sci. 277, 819–827. doi:10.1098/rspb.2009.1679

Chou, H.-H., Chiu, H.-C., Delaney, N.F., Segrè, D., Marx, C.J., 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332, 1190–1192. doi:10.1126/science.1203799

Cohen, O., Gophna, U., Pupko, T., 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28, 1481–1489. doi:10.1093/molbev/msq333

Devirgiliis, C., Zinno, P., Perozzi, G., 2013. Update on antibiotic resistance in foodborne Lactobacillus and Lactococcus species. Front Microbiol 4, 301. doi:10.3389/fmicb.2013.00301

Engelberg-Kulka, H., Glaser, G., 1999. Addiction modules and programmed cell death and antideath in bacterial cultures. Annu. Rev. Microbiol. 53, 43–70. doi:10.1146/annurev.micro.53.1.43

Gophna, U., Ofran, Y., 2011. Lateral acquisition of genes is affected by the friendliness of their products. Proc. Natl. Acad. Sci. U.S.A. 108, 343–348. doi:10.1073/pnas.1009775108

Hao, W., Richardson, A.O., Zheng, Y., Palmer, J.D., 2010. Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. Proc. Natl. Acad. Sci. U.S.A. 107, 21576–21581. doi:10.1073/pnas.1016295107

Hellweger, F.L., 2013. Escherichia coli adapts to tetracycline resistance plasmid (pBR322) by mutating endogenous potassium transport: in silico hypothesis testing. FEMS Microbiol. Ecol. 83, 622–631. doi:10.1111/1574-6941.12019

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U.S.A. 96, 3801–3806.

Jansen, G., Barbosa, C., Schulenburg, H., 2014. Experimental evolution as an efficient tool to dissect adaptive paths to antibiotic resistance. Drug Resist. Updat. doi:10.1016/j.drup.2014.02.002

Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R., Marraffini, L.A., 2013. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. PLoS Genet. 9, e1003844. doi:10.1371/journal.pgen.1003844

Juhas, M., 2013. Horizontal gene transfer in human pathogens. Crit. Rev. Microbiol. doi:10.3109/1040841X.2013.804031

Keeling, P.J., Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618. doi:10.1038/nrg2386

Knight, G.M., Budd, E.L., Lindsay, J.A., 2013. Large mobile genetic elements carrying resistance genes that do not confer a fitness burden in healthcare-associated meticillin-resistant Staphylococcus aureus. Microbiology (Reading, Engl.) 159, 1661–1672. doi:10.1099/mic.0.068551-0

Lawrence, J.G., Hendrickson, H., 2003. Lateral gene transfer: when will adolescence end? Mol. Microbiol. 50, 739–749.

Lindsay, J.A., 2014. Staphylococcus aureus genomics and the impact of horizontal gene transfer. Int. J. Med. Microbiol. 304, 103–109. doi:10.1016/j.ijmm.2013.11.010

Moran, N.A., Jarvik, T., 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328, 624–627. doi:10.1126/science.1187113

Morita, M., Yamamoto, S., Hiyoshi, H., Kodama, T., Okura, M., Arakawa, E., Alam, M., Ohnishi, M., Izumiya, H., Watanabe, H., 2013. Horizontal gene transfer of a genetic island encoding a type III secretion system distributed in Vibrio cholerae. Microbiol. Immunol. 57, 334–339. doi:10.1111/1348-0421.12039

Nielsen, K.M., Bøhn, T., Townsend, J.P., 2014. Detecting rare gene transfer events in bacterial populations. Front Microbiol 4, 415. doi:10.3389/fmicb.2013.00415

Noda-García, L., Barona-Gómez, F., 2013. Enzyme evolution beyond gene duplication: A model for incorporating horizontal gene transfer. Mob Genet Elements 3, e26439. doi:10.4161/mge.26439

Papp, B., Pál, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197. doi:10.1038/nature01771

Park, C., Zhang, J., 2012. High expression hampers horizontal gene transfer. Genome Biol Evol 4, 523–532. doi:10.1093/gbe/evs030

Schaack, S., Gilbert, C., Feschotte, C., 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. (Amst.) 25, 537–546. doi:10.1016/j.tree.2010.06.001

Schönknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Bräutigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., Carr, K., Wilkerson, C., Rensing, S.A., Gagneul, D., Dickenson, N.E., Oesterhelt, C., Lercher, M.J., Weber, A.P.M., 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science 339, 1207–1210. doi:10.1126/science.1231707

Schönknecht, G., Weber, A.P.M., Lercher, M.J., 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36, 9–20. doi:10.1002/bies.201300095

Syvanen, M., 2012. Evolutionary implications of horizontal gene transfer. Annu. Rev. Genet. 46, 341–358. doi:10.1146/annurev-genet-110711-155529

Travisano, M., Lenski, R.E., 1996. Long-term experimental evolution in Escherichia coli. IV. Targets of selection and the specificity of adaptation. Genetics 143, 15–26.

Travisano, M., Mongold, J.A., Bennett, A.F., Lenski, R.E., 1995. Experimental tests of the roles of adaptation, chance, and history in evolution. Science 267, 87–90.

Treangen, T.J., Rocha, E.P.C., 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7, e1001284. doi:10.1371/journal.pgen.1001284

Vergote, G.J., Vervaet, C., Van Driessche, I., Hoste, S., De Smedt, S., Demeester, J., Jain, R.A., Ruddy, S., Remon, J.P., 2002. In vivo evaluation of matrix pellets containing nanocrystalline ketoprofen. Int J Pharm 240, 79–84.

Wellner, A., Gophna, U., 2008. Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. Mol. Biol. Evol. 25, 1835–1840. doi:10.1093/molbev/msn131

Williams, D., Gogarten, J.P., Papke, R.T., 2012. Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol Evol 4, 1223–1244. doi:10.1093/gbe/evs098

Yoshida, S., Maruyama, S., Nozaki, H., Shirasu, K., 2010. Horizontal gene transfer by the parasitic plant Striga hermonthica. Science 328, 1128. doi:10.1126/science.1187145

CHAPTER FOUR: BIGGER IS NOT ALWAYS BETTER - TRANSMISSION AND FITNESS BURDEN OF ~1MB *PSEUDOMONAS SYRINGAE* MEGAPLASMID pMPPla107.

Authors: Artur Romanchuk*, Corbin D. Jones, Kedar Karkare, Autumn Moore, Brian A. Smith, Chelsea Jones, Kevin Dougherty, David A. Baltrus
Resubmitted with revisions to PLASMID March 14, 2014
*writer and plasmid transfer experiment design.

ABSTRACT

Horizontal gene transfer (HGT) is a widespread process that enables the acquisition of genes and metabolic pathways in single evolutionary steps. Previous reports have described fitness costs of HGT, but have largely focused on the acquisition of relatively small plasmids. We have previously shown that a *Pseudomonas syringae* pv. *lachrymans* strain recently acquired a cryptic megaplasmid, pMPPla107. This extrachromosomal element contributes hundreds of new genes to *P. syringae* and increases total genomic content by approximately 18%. However, this early work did not directly explore transmissibility, stability, or fitness costs associated with acquisition of pMPPla107. Here we show that pMPPla107 is self-transmissible across a variety of diverse Pseudomonad strains, on both solid agar and within shaking liquid medium cultures, with conjugation dependent on a type IV secretion system. To our knowledge, this is the largest self-transmissible megaplasmids known outside of *Sinorhizobium*. This megaplasmid can be lost from all novel hosts although the rate of loss depends on medium type and genomic background. However, in contrast, pMPPla107 is faithfully maintained within the original parent strain *(Pla*107) even under direct negative

selection during laboratory assays. These results suggest that *Pla*107 specific stabilizing mutations have occurred either on this strain's chromosome or within the megaplasmid. Lastly, we demonstrate that acquisition of pMPPla107 by strains other than *Pla*107 imparts severe (20%) fitness costs under competitive conditions *in vitro*. We show that pMPPla107 is capable of transmitting and maintaining itself across multiple *Pseudomonas* species, rendering it one of the largest conjugative elements discovered to date. The relative stability of pMPPla107, coupled with extensive fitness costs, makes it a tractable model system for investigating evolutionary and genetic mechanisms of megaplasmid maintenance and a unique testing ground to explore evolutionary dynamics after HGT of large secondary elements.

INTRODUCTION

Horizontal gene transfer mediated by large secondary elements, which results in the movement of genomic regions directly between organisms without reproduction, substantially alters evolutionary dynamics within microbial populations (Gogarten and Townsend 2005, Diaz-Ricci and Herna´nde 2000). HGT facilitates rapid phenotypic evolution (Pennisi 2004), the incorporation of new metabolic pathways and network expansions, and can lead to ecological speciation across microbes (Lawrence 1996, Shapiro et al. 2012, Lawrence 1998). Similarly, HGT is a major mechanism behind the acquisition of antibiotic resistance genes and new virulence determinants, which facilitate the emergence of new pathogen types (Grad et al., 2011, Sørensen 2005, Baltrus et al. 2011, Ashbolt et al. 2013, Broaders et al. 2013, Warnes et al. 2012).

Megaplasmids are the largest contiguous regions that can undergo HGT across bacterial cells. The term chromid has been proposed to designate a subset of megaplasmids that share a

number of important characteristics with the main chromosome including similar GC content and suites of loci which are diverged from, yet potentially provide redundant functions to, existing housekeeping genes. These genomic patterns have led to the speculation that secondary chromosomes in genera such as *Burkholderia* and *Vibrio* began as chromids (Cooper et al. 2010), although distinct transitional forms have yet to be clearly identified. However, unlike the chromosome, chromids maintain plasmid replication and partitioning systems and the bulk of the genes that they carry may be neutral or serve accessory functions. Importantly, genus-specific genes are overrepresented on chromids and this lack of genus-independent genetic diversity suggests existence of barriers to megaplasmid transfer outside of closely related species. Given that secondary chromosomes may be "evolutionary test beds" (Cooper et al. 2010) and chromids contain an abundance of unique genes and pathways, HGT of large secondary elements has the potential to dramatically alter evolutionary trajectories within microbial species (Wolf and Koonin et al. 2012).

To be maintained at high frequencies within populations, regions transferred via HGT must provide fitness benefits, minimize fitness costs, or disperse into new host cells at high rates (Gomes et al. 2013, Ponciano 2006). However, benefits of HGT are often partially counterbalanced by metabolic or physiological costs as natural selection, prior to acquisition, has not had an opportunity to fine tune interactions between transferred regions and their new genomic contexts (Baltrus et al. 2012). Previous studies have investigated both the positive and negative effects of HGT using relatively small plasmids (<15kb) (De Gelder 2008, Harrison and Brockhurst 2012, although see Platt et al. 2011), but much of this work provides an incomplete view as plasmids can range up to 2Mb (Smillie et al. 2010). Because increasing plasmid size appears to be negatively correlated with self-transmissibility, and plasmids may fundamentally

change in gene content and composition above an imprecisely defined threshold (Harrison et al. 2010, Harrison and Brockhurst 2012), it remains unclear if and how selection pressures scale with plasmid size and how frequently such costs act as a barrier to horizontal transfer.

The megaplasmid pMPPla107 was first identified by genome sequencing a phylogenetically diverse suite of *Pseudomonas syringae* strains (Baltrus et al. 2011). Evidence of pMPPla107-like plasmids has only been found within a cluster of closely related *P. syringae* strains isolated during a disease outbreak on cucumbers in Japan in the 1960s with phylogenetic data suggesting acquisition was a relatively recent HGT event (Baltrus et al. 2011). This secondary element contains all the hallmarks of previously identified chromids, including the presence of many hypothetical proteins, potential duplicate versions of housekeeping genes, and tRNA genes highly similar to those in other *Pseudomonas* species (Table 4.4). Furthermore, self-transmissibility is suggested by the presence of a putative type IV secretion system that is most similar in sequence to *Legionella* Dot/Icm (Baltrus et al. 2011, Joseph et al. 1999).

In this work we demonstrate that pMPPla107 is self-transmissible across a diverse range of Pseudomonads. We also show that, despite significant deleterious effects on growth, pMPPla107 can be stably maintained within new genomic backgrounds when exposed to a range of growth conditions. That presence of this megaplasmid in natural host strains is strongly correlated with slower growth *in vitro* and *in planta* compared to closely related strains that lack pMPPla107 (Baltrus et al. 2011), suggests that these costs have not been completely compensated for evolution. Due to the size of pMPPla107 and shared characteristics with previously discovered chromids, this system provides a unique and experimentally tractable opportunity to study how large-scale HGT alters evolutionary dynamics within bacterial populations.

Hypothesis

Plasmids are a mobile gene pool that can be accessed by multiple unrelated bacterial species. Megaplasmids comprise a rich reservoir of genes available for transfer across bacteria, but which, under certain conditions, may also significantly lower fitness of the recipient bacteria. The large size of pMPPla107 and the nature of genes it carries enables laboratory tests of a key hypothesis concerning megaplasmid transmission dynamics and maintenance: the major barrier to spread of pMPPla107 across viable hosts is its deleterious fitness impact on the host cell rather than transfer rates.

Approach

We quantify the pMPPla107 transfer rates to several related *Pseudomonas syringae* pathovars, *P. stutzeri,* and *Escherichia coli.* We further quantify loss rate of pMPPla107 across several genetic backgrounds in different growth media types. Finally, we compare fitness impact of the pMPPla107 megaplasmid between close relative of the parent *P. syringae* strain and a divergent species, *P. stutzeri.*

MATERIAL AND METHODS

Culture conditions

All experiments were carried out at 27^{o}C. Liquid and solid medium consisted of either low salt Lysogeny (Lennox Luria-Bertani broth (LB) or M9 minimal medium, prepared according to Sambrook & Russel (2001), and supplemented with additional carbon source and antibiotics where appropriate. For growth of *P. stutzeri* strains, saltwater LB (SWLB) was instead of LB as a growth medium (Sikorski et al. 1998). All liquid cultures were incubated on a rotary shaker (200 rpm). Antibiotics were used as necessary in the following concentrations:

20 μg/mL tetracycline (Tet), 50 μg/mL kanamycin (Kan), 50 μg/mL rifampicin (Rif), 10 μg/mL gentamycin (Gm). Sterile 10mL $MgCl_2$ was used for all dilutions and cell suspensions.

Bacterial strains and plasmids

The ~976 kb plasmid pMPPla107 (Baltrus et al. 2011), was originally described in *Pseudomonas syringae* strain MAFF301305 (also known as *Pla*107). A draft assembly sequence for pMPPla107 can be found at GenBank accession CM000959.1. All strains used in the study are listed in Table 4.1. All cloning plasmids used in gene knockouts and complementation experiments, as well as their relevant phenotypic characteristics, are listed in Table 4.1. Splicing overlap extension PCR was used to construct both the megaplasmid tagging construct and the *dotB* deletion construct. For construction of the *dotB* deletion strain, regions upstream of putative *dotB* were first amplified by primers DBL134/DBL136 with regions downstream first amplified by DBL135/DBL137. Resulting fragments were then spliced using bridge PCR and amplification with Gateway tailing primers DBL3 (5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTCC) and DBL4 (5'-GGGGACCACTTTGTACAAGAAAGCTGGGTG). Successfully bridged fragments were recombined into the Gateway donor plasmid pDONR207 and then into the destination vector pMTN1907 where they could be conjugated into *P. syringae* pv. *lachrymans* 107 (Baltrus et al.. 2012). The megaplasmid tagging construct was created in the same way, except primer sets for amplification were DAB1185/DAB1287 and DAB1186/DAB1288. See supporting information for further information regarding construction of the *dotB*- mutant and tagging construct.

Table 4.1.
Pseudomonas syringae strains and plasmids.

Strain	Species	Pathovar	Markers	Notes	Citation
*Pla*107	*P. syringae*	*lachrymans*	Rif^R	Rif^R isolate of MAFF 301315	*Baltrus et al 2011*
DAB462	*P. syringae*	*lachrymans*	Rif^R	Rif^R isolate of MAFF YM8003	This paper
DAB837	*P. syringae*	*lachrymans*	Rif^R, Kan^R, Cam^R, *eyfp*	DAB462 with Tn7 from AKN84	This paper
*Pto*DC3000	*P. syringae*	*tomato*	Rif^R	Rif^R isolate of NCPPB 4369	*Cuppels 1986*
DAB812	*P. syringae*	*tomato*	Rif^R, Kan^R, Cam^R, *eyfp*	*Pto*DC3000 with Tn7 from AKN86	This paper
DAB885	*P. syringae*	*lachrymans*	Rif^R, Tet^R, Suc^S	*Pla*107 with pDAB326 integrated into megaplasmid pMPPla107	This paper
DAB328	*P. syringae*	*lachrymans*	Rif^R, Kan^R, Tet^R, Suc^S	DAB885 selected for Kan^R	This paper
DAB895	*P. syringae*	*lachrymans*	Rif^R, Kan^R, Cam^R, *eyfp*, Tet^R, Suc^S	DAB837 conjugated with megaplasmid from DAB885	This paper
DAB908	*P. syringae*	*tomato*	Rif^R, Kan^R, Cam^R, *eyfp*, Tet^R, Suc^S	DAB812 conjugated with megaplasmid from DAB885	This paper
23a24	*P. stutzeri*			environmental isolate	*Sikorski et al 1999*
DBL332	*P. stutzeri*		Rif^R	Rif^R isolate of 23a24	This paper
DBL386	*P. stutzeri*		Rif^R, Kan^R, Cam^R, *eyfp*	DBL332 with Tn7 from AKN86	This paper
DBL365	*P. stutzeri*		Kan^R Tet^R, Suc^S	DBL332 conjugated with megaplasmid from DAB328	This paper
DBL390	*P. stutzeri*		Rif^R, Gm^R, *lacZ*	DBL332 with pUC18-mini-Tn7T-Gm-*lacZ*	This paper
DBL408	*P. stutzeri*		Rif^R, Kan^R, Cam^R, *eyfp*, Tet^R, Suc^S	DBL408 conjugated with megaplasmid from DAB885	This paper
DBL453	*P. stutzeri*		Rif^R	DAB365 with MTN1907 flipped out	This paper
DBL492	*P. stutzeri*		Rif^R, Nal^R	Nal100 resistant isolate of DBL332	This paper
DBL493	*P. stutzeri*		Rif^R, Gm^R	DBL453 with pMAR2xT7 integrated into megaplasmid	This paper
DBL494	*P. stutzeri*		Rif^R, Gm^R, Nal^R	DBL492 conjugated with megaplasmid from DBL493	This paper
DBL610	*P. syringae*	*lachrymans*	Rif^R, Gm^R	DAB462 conjugated with megaplasmid from DBL494	This paper
DBL713	*P. putida*		Kan^R, Suc^S, Naturally Tet^R Cam^R	DBL306 conjugated with megaplasmid from DBL328	This paper
DBL306	*P. putida*		Naturally Tet^R Cam^R	KT2440 ATCC 47054	*Heim et al 2002*
DBL714	*P. fluorescens*		Rif^R, Kan^R Tet^R, Suc^S	DBL73 conjugated with megaplasmid from DBL328	This paper
DBL73	*P. fluorescens*		Rif^R	Rif^R isolate of SBW25	This paper
DBL72	*P. fluorescens*			from Paul Rainey	*Joyce et al 2012*
DBL675	*P. syringae*	*lachrymans*	Rif^R, Tet^R, Suc^S	*Pla*107 with pDBL47 integrated into megaplasmid pMPPla107	This paper
DBL695	*P. syringae*	*lachrymans*	Rif^R	*Pla*107 with pDBL47 flipped out, leaving *dotB* deletion	This paper
DBL700	*P. syringae*	*lachrymans*	Rif^R, Tet^R, Suc^S	DBL695 with pDAB326 integrated into *dotB*- megaplasmid	This paper
ar1448a-1	*P. syringae*	*phaseolicola*	Rif^R	Rif^R isolate of Pph1448a	This paper
ar1448a-2	*P. syringae*	*phaseolicola*	Rif^R, Kan^R, Cam^R, *eyfp*	*Pph*1448a with Tn7 from AKN86	This paper
ar50225-2	*P. aptata*		Rif^R, Kan^R, Cam^R, *eyfp*	DSM50225 with Tn7 from AKN86	This paper
ar50225-1	*P. aptata*		Rif^R	Rif^R isolate of DSM50225	This paper
Plasmid					
pDAB321			GmR	pDONR207 with megaplasmid tagging construct	This paper
pDAB326			TetR, KanR	MTN1907 with megaplasmid tagging construct (LR from pDAB321)	This paper
pDONR207			GmR	Gateway Entry vector	Invitrogen
MTN1907			TetR, KanR	Gateway Destination vector	*Baltrus et al 2012*
AKN84			KanR, CamR, *eYFP*	Tn7 vector with *eYFP*	*Lambertson et al 2004*
AKN86			KanR, CamR, *eCFP*	Tn7 vector with *eCFP*	*Lambertson et al 2004*
pUC18-mini-Tn7T-Gm-*lacZ*			GmR, *lacZ*	Tn7 vector with *lacZ*	*Choi et al. 2006*
pTNS2			AmpR	Transposase plasmid for Tn7	*Choi et al. 2006*
pDBL46			GmR	pDONR207 with *dotB* deletion construct	This paper
pDBL47			TetR, KanR	MTN1907 with *dotB* deletion construct (LR from pDBL46)	This paper

^{R}resistance marker to particular antibiotic or substance
^{S}susceptibility marker to particular antibiotic or substance

Conjugative plasmid transfer

Transfer rates were measured by counting the number of recipient strains containing pMPPla107 24-hours post exposure to the donor strain using the procedure described in De Gelder *et al.* 2008. In all cases except tests of the *dotB* mutant (DAB700), the donor strain was DAB885. To transfer plasmids between strains, 2 mL of overnight-grown cultures of the pMPPla107 donor (grown in 10 µg/mL Tet) and a kanamycin resistant recipient (grown in 50 µg/mL Kan) were centrifuged, the supernatants removed, and the pellets resuspended in 150 µL of sterile 10mM $MgCl_2$. Subsequently mating partner cell suspension were combined (50 µL of

each), centrifuged a second time followed by supernatant removal. The resulting pellet was resuspended in 50 μL of sterile 10mM $MgCl_2$, with that total volume pipetted onto a 47-mm diameter polycarbonate filter (Isopore membrane filter 0.4mm HTBP, Millipore). This filter was then placed on top of a sterile LB plate containing no antibiotics. After overnight incubation, the entire cell mass was harvested and suspended in 300 μL of sterile 10mM $MgCl_2$, from which dilutions were made to separate LB Tet/Kan, LB Tet, and LB Kan plates selective for transconjugants, donors, and recipient cells, respectively. We have randomly sampled cells from >25 independent assays across strains using diagnostic PCR and have never identified a recipient cell that displays transconjugant phenotypes but which lacks the megaplasmid. Transfer experiments using the *dotB* deletion mutant (DBL700) were carried out in the same manner using two independently created mutant strains and with DAB837 as a recipient. In all cases where Tet/Kan resistant colonies appeared during *dotB*- conjugations, the colonies were confirmed as kanamycin resistant DBL700 cells using Tn7 site specific PCR to discriminate these cells from DAB837. For liquid conjugation experiments, O/N cultures of donor and recipient cells were mixed 1:1, diluted 1:100 in KB medium containing 50 μg/mL rifampicin, and grown O/N on a rotary shaker at 27^{o}C. For these experiments, donor cells were typically DAB885 while recipient cells were either DAB837 or DBL390. The next day, dilutions of the O/N culture were plated on KB agar containing combinations of antibiotics that selected for the recipient strain (kanamycin for DAB837 and gentamycin for DBL390) as well as the megaplasmid (tetracycline). Megaplasmid acquisition was confirmed by diagnostic PCRs for both the megaplasmid and recipient strain and as well as phenotypes when possible (blue clones on Xgal for DBL390).

Plasmid loss experiments

pMPPla107 was phenotypically marked with tetracycline resistance by integrating a suicide plasmid (pDAB326) through conjugative mating and recombination. Integration of pDAB326 also allows for negative selection during growth on 5% sucrose medium via *sacB* (Baltrus et al. 2012). To quantify the rate of loss of pMPPla107 for a variety of strains, each strain was grown O/N in liquid culture at 27^{o}C in the assay medium with tetracycline. Cells were then washed 1x with 10mM $MgCl_2$ and diluted 1:10 into a master mix that contained the assay medium without tetracycline. This master mix was dispensed into replicate cultures within a 96 well plate and allowed to grow until cultures appeared turbid to the eye (typically 1-3 days depending on strain and media type). At this point, each replicate culture was diluted and plated on KB plates with no selection (to obtain total population sizes) as well as KB agar plates containing 5% sucrose (to obtain counts of plasmid free cells). Each assay was carried out with at least 6 independent cultures per replicate assay, with at least 3 independent assays for each strain*medium combination. Numbers in Table 4.3 represent averages pooled across these independent assays. Because experiments were initiated by a 1:10 dilution from a turbid stationary phase culture, approximately 3.3 doublings occurred during each these experiments. We have adjusted percentages of plasmid loss by dividing overall percentage of sucrose resistant cells at the end of each experiment by 3.3, in order to calculate plasmid loss per cell division. We acknowledge that strain specific selection pressures could bias percentages reported in Table 4.2 but have no reason to believe that such biases significantly skew comparisons across strains reported here because under these conditions all strains grow at approximately the same rates and to similar final densities (Lau et al. 2013, Sota et al. 2007, De Gelder et al. 2007, Heuer et al. 2007).

As elimination of the pDAB326 phenotypic markers could occur via mutation or recombination of the *sacB* cassette, we tested the correlation between phenotype loss and megaplasmid loss in at least 24 single colonies chosen from independent cultures having undergone the same conditions as rate loss experiments described above. PCR was used to verify presence of the megaplasmid for each colony using primer set 1 and primer set 2 from Baltrus *et al.* 2011. These individual colonies were also streaked to KB plates with and without tetracycline to phenotypically confirm megaplasmid loss. The megaplasmid was considered lost if the parent strain was susceptible to tetracycline, sucrose resistant, and failed in PCR reactions for two separate primer sets.

Plasmid cost estimation through direct competition

Competitive fitness assays, carried out as described in Baltrus et al. 2008 with modifications as described below, were used to quantify the fitness cost of pMPPla107. Schematics for each type of competition are provided in Supplemental Figures 1 and 2.

To prepare strains for competition, all strains were streaked out on LB/SWLB agar medium. Strains were first grown overnight from these plates at 27^{o}C in liquid LB/SWLB medium containing rifampicin. Strains containing the megaplasmid were grown in LB medium further supplemented with tetracycline or gentamycin depending on the assay and selection for megaplasmid maintenance. After the initial overnight, strains were diluted 1:100 and grown for an additional period of time (24 hours for *P. syringae,* 48 hours for *P. stutzeri*) at 27^{o}C in liquid LB/SWLB medium at which point competitions were carried out. For each competition, a test strain was directly competed against a control strain in the same culture, with relative fitness measured as the change in ratio of test:control over the course of the assay. In all cases, control

strains lacked the megaplasmid. Competitions for both megaplasmid + and – strains occurred at the same time and used independent samples from the same control strain inoculum.

To perform competitions, strains were first washed 1x with 10mM MgCl2, and mixed in 5:1 test:control ratios. This particular ratio was used because, when competitions were initiated with alternative ratios (1:1), we often found that megaplasmid strain counts were below the level of detection when the end point was sampled. A 1:100 dilution of these strain mixtures were used to found each replicate competition culture, yielding approximately 6.7 generations of growth during the assay. Dilutions of the these strain mixtures were plated LB/SWLB agar plates containing no antibiotics other than rifampicin in order to assess the ratio of test/control strains at time zero. Competitions were then grown for either 24 (*P. syringae)* or 48 hours (*P. stutzeri*) at 27^{o}C at which point dilutions were plated on KB/SWLB medium containing rifampicin supplemented with 40 μg/mL Xgal as appropriate. For *Pla*8003, test strains consisted of paired isolated that either lacked or contained pMPPla107 tagged with gentamycin resistance (megaplasmid -, DAB462; megaplasmid +, DBL610). The control strain for *Pla*8003, DAB837, was created by phenotypically marking *Pla*8003 with a kanamycin resistance allele through Tn7 transposition (Lambertsen et al. 2004). Once colonies had grown to sufficient size, each plate was replica plated to KB medium supplemented with kanamycin to quantify size of the control strain population. For *P. stutzeri,* test strains consisted of paired isolates that either lacked or contained a megaplasmid tagged with tetracycline resistance (megaplasmid -, DBL332; megaplasmid +, DBL365). The control strain DBL390 was phenotypically marked with gentamycin resistance and *lacZ* at its Tn7 site. Discrimination between white and blue colonies after plating on SWLB medium supplemented with Xgal enabled quantification of population sizes for both test and the control strains without replica plating. In every one of our assays, the

megaplasmid free test strains had slightly lower fitness than the control strains. To account for this bias, relative fitness within each assay block was normalized so that the megaplasmid free strains had a value of 1, with averages of competitions adjusted by the same value. An ANOVA was then performed on these adjusted values with experimental block as a random variable, and strain as a fixed variable.

Plasmid effect on intrinsic growth rate

To independently investigate plasmid cost values in *P. stutzeri* and *P. syringae*, growth experiments were carried out in M9 medium supplemented with glycerol (10mM) as the sole carbon source. Growth curves of plasmid carrying strains and their plasmid free ancestors were quantified and compared over 48 hour period in a Molecular Devices Spectramax 340PC plate reader. Strains were first conditioned by growing each in a 96 well plate for 48 hours in M9 medium at room temperature. Each strain was then diluted 1:100 in fresh M9 medium with at least 7 replicate cultures per strain per plate and grown for an additional 48 hours at room temperature. All empty wells contained sterile M9 and acted as a temperature buffers and cross-contamination controls for the entire plate. The slope of the growth curve at the transition point from lag phase to exponential phase (Vmax – maximum velocity at that point) was used to calculate fitness.

Blast and KEGG Annotation of Proteins Within pMPPla107

The Kyoto encyclopedia of genes and genomes (KEGG - http://www.genome.jp/kegg/) was queried with all publically available protein annotations from ORFs found on pMPPla107 (GenBank: CM000959.1). The overall number of proteins matched to a subset of metabolic

pathways is reported within Table 4.3. BLASTp was then used to search the NR database with protein sequences for a subset of genes where pathway membership could be predicted through KEGG annotation. The annotations and characteristics for a subset of housekeeping genes found on the megaplasmid along with best sequence matches are reported in Table 4.4.

RESULTS

The pMPPla107 megaplasmid is self-transmissible

Mobilization assays show that pMPPla107 possesses high transfer rates from its parent strain to all *P. syringae* strains tested, with equally high transfer rates to *P. stutzeri* (Figure 4.1). We also successfully conjugated pMPPla107 to both *P. putida* KT2440 and *P. fluorescens* SBW25, although we have not quantified exact rates (Table 4.1). We found that pMPPla107 transfers fairly well in liquid medium, although transconjugant proportions are approximately $1/10^{th}$ of what was found on solid agar (data not shown). No successful megaplasmid transfers were ever observed to occur between *P. syringae* and *E.coli* strain MG1655 in three independent trials. Within tested *P. syringae* strains, *Pto*DC3000 is the most divergent from the donor while *Pla*8003 is least divergent (Baltrus et al. 2011). Given similar observed transfer rates across all tested *Pseudomonas* strains, it does not appear as though the ability of pMPPla107 transfer is correlated with evolutionary relatedness between the recipient and the donor pseudomonad strains. However not transfer occurred to highly unrelated *E. coli* strain suggesting possible impact of extreme divergence on conjugation efficiency.

pMPPla107 contains a putative type IV secretion system that could enable conjugation. To test this possibility we deleted a gene, highly similar to the *dotB* ATPase, from the

megaplasmid and measured conjugation from the parent strain into *Pla*8003 using the same assays as described above. No true transconjugants were recovered after three independent trials, with an average recipient population size of $5x10^9$ cells per trial. On occasion we found cells which phenocopied transconjugants for antibiotic resistance, but in every case tested these were shown to be kanamycin resistant mutants of the donor strain DBL700 (which are likely enabled by the presence of a kanamycin resistance gene in pMTN1907 normally inactive in *Pseudomonas* (Baltrus et al. 2012)). As this mutant completely eliminates transfer across highly similar strains we conclude that the putative type IV secretion system is required for conjugation.

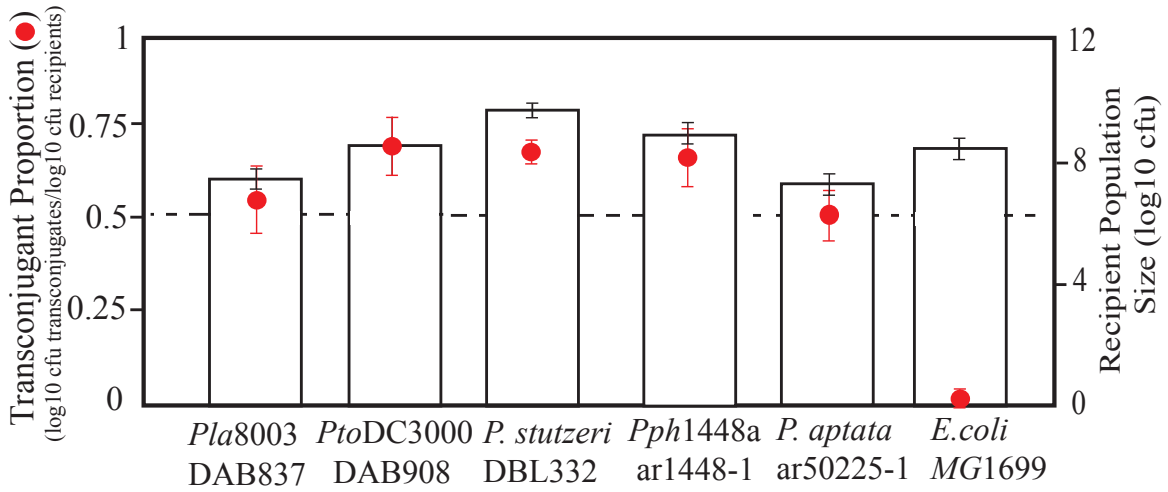

Figure 4.1 The pMPPla107 conjugates with high efficiency to pseudomonad strains but not *E. coli.* Box plot showing transfer of pMPPla107 megaplasmid to 6 *Pseudomonas* strains in comparison to *Escherichia coli*. Solid bars indicate population size of recipient bacteria (log_{10} c.f.u. - right axis) and red circles indicate proportion of pMPPla107 transconjugants after 24 hours post contact.

Acquisition of the pMPPla107 megaplasmid by naive cells leads to fitness loss

We quantified pMPPla107 fitness cost in two divergent Pseudomonad species: *P. syringae Pla*8003 and *P. stutzeri*. In both cases, acquisition of pMPPla107 leads to a ~20% loss in competitive fitness (Figure 4.2A). Although megaplasmid transfer can occur between test and

control strains under these competitive conditions, we found that such transfers do not occur at high enough rates to significantly affect fitness estimates or explain the differences between strains (<1% of cells, data not shown). Interestingly, visual examination of growth curves revealed that megaplasmid presence extends the lag phase and only slightly reduces the slope of the exponential phase within *P. stutzeri* and *P. syringae* (Figure 4.2B).

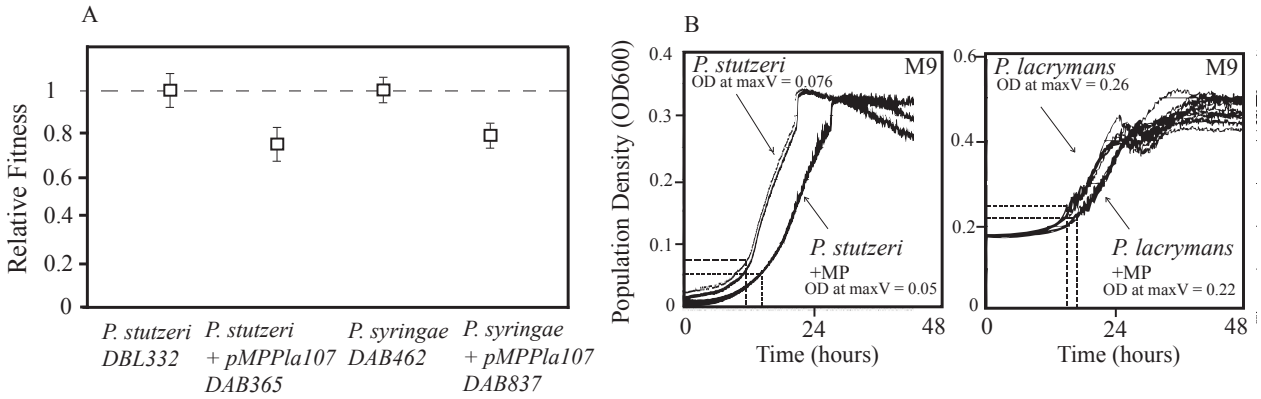

Figure 4.2 The pMPPla107 lowers host fitness of *Pla*8003, a close relative to the donor strain *P. syringae*, and divergent pseudomonas *P. stutzeri*. A) Pairwise competition results between strains of *P. syringae* or *P. stutzeri* that contain or lack the megaplasmid. In each case, megaplasmid acquisition lowers competitive fitness by ~20% (*P. syringae* p-value =0.0229, F = 43.07, df = 1; *P. stutzeri* p-value =0.0022, F = 120.95, df = 1). Relative fitness is normalized so that value of megaplasmid free strains is 1 within each replicate experiment. B) Example growth curves of pMPPla107 containing *P. stutzeri* (DAB365) and *P. lacrymans* (DAB837) cells in comparison to plasmid free cells. Megaplasmid free cells reach maximum velocity (maxV = lag phase to exponential phase transition) sooner than cells containing the pMPPla107.

Generally low pMPPla107 loss rate depends on medium type and host

After acquisition, pMPPla107 can be lost from the host population across all recipient strains under a variety of media types (Table 4.2, complete, M9 media supplemented with glucose, M9 media supplemented with succinate, and M9 media supplemented with a combination glucose and succinate). Megaplasmid loss rates are generally low however dependent on host and medium type. Comparatively, megaplasmid loss rates are highest in rich KB medium and lowest in glucose medium, with intermediate loss in succinate. However, loss of pMPPla107 does appear to partially depend on interactions between bacterial host genetic

background and medium type as plasmid loss in KB and succinate medium differs by an order of magnitude across *Pto*DC3000 and *Pla*8003. Moreover, while we never confirmed loss of the megaplasmid from *P. stutzeri* strains (DBL365 or DBL408) during the course of these assays, we have subsequently seen loss during culturing of these strains on KB medium in the absence of antibiotic selection for the megaplasmid. Despite screening four times as many independently evolved sucrose resistant colonies from *Pla*107 as any of the other strains and continually screening during routine laboratory culture, we have not been able to isolate a megaplasmid free version of this strain (Table 4.2).

Table 4.2.
Megaplasmid loss without stabilizing selection: proportion of megaplasmid free cells in three different environments.

Strain	Medium	Average % sucrose resistant	St Dev % sucrose resistant	# of Assays	PCR
*Pto*DC3000	KB	2.757%	1.948%	3	N/A
DAB908	M9 + Glucose	0.031%	0.049%	3	N/A
	M9 + Succinate	2.597%	1.370%	5	N/A
	M9 + Glucose, Succinate	0.470%	0.492%	4	0/24
*Pla*8003	KB	0.065%	0.094%	3	N/A
DAB895	M9 + Glucose	0.003%	0.002%	3	N/A
	M9 + Succinate	0.012%	0.009%	3	N/A
	M9 + Glucose, Succinate	0.007%	0.001%	4	18/24
*Pla*107	KB	0.110%	0.146%	3	N/A
DAB328	M9 + Glucose	0.012%	0.010%	3	N/A
	M9 + Succinate	0.245%	0.267%	3	N/A
	M9 + Glucose, Succinate	0.010%	0.015%	4	100/100
P. stutzeri	SW-LB	0.159%	0.208%	6	N/A
DBL408	M9 + Glucose	0.071%	0.042%	4	N/A
	M9 + Succinate	0.089%	0.077%	6	N/A
	M9 + Glucose, Succinate	0.023%	0.027%	6	24/24*

*can be lost occasionally during passage on KB media

DISCUSSION

Despite the widespread and dramatic fitness effects of HGT on microbial evolution, a general understanding for how fitness costs scale with size of the transferred region remains unclear. In the face of such costs, megaplasmids can avoid loss from bacterial populations through three major strategies: by maintaining high transfer rates to naive cells, by increasing

host fitness, and by carrying toxin-antitoxin cassettes which create their own selection pressures (Loh et al. 2013, Unterholzner et al. 2013,Kophmann et al. 2013).

We find that, under laboratory conditions, transfer of the pMPPla107 occurs at comparatively high rates throughout strains across the genus *Pseudomonas.* Unlike many other conjugative plasmids, transfer can occur on solid medium as well as within shaking liquid cultures. This suggests a strong conjugation pilus, such that the conjugation machinery can overcome the physical forces of agitation during liquid culture. We have shown that deletion of a gene resembling *DotB* ATPase from the megaplasmid eliminates self-transmissibility, which suggests that the type IV secretion system carried by pMPPla107 is essential for conjugation. We further show that, while pMPPla107 is self-transmissible across diverse Pseudomonads, conjugation to a common laboratory strain of *E. coli* does not freely occur. It is unclear at present what the barrier to this transmission is. Possibilities include failure of the megaplasmid to establish a viable pilus with *E. coli* cells, failure to replicate within *E. coli* cells, and the potential for toxicity once acquired by *E.coli*. Alternatively *E.coli* may produce unknown compounds toxic to the donor *P. syringae* strain.

Acquisition of pMPPla107 is costly under laboratory conditions as two tested strains containing the megaplasmid grow more slowly in isolation and during competition assays. That the deleterious effects on growth appear to predominantly affect lag phase suggests that ribosome occupancy is a possible limiting factor for growth (Shachrai et al. 2010, Stoebel et al. 2010). Alternatively, these fitness costs could be due to cytotoxicity of improperly folded proteins, resource drain due to transcription and translation of horizontally transferred genes and pathways, or direct detrimental interactions between chromosomal pathways and megaplasmid

encoded proteins (Starikova et al. 2013, San Millan et al. 2013, Park & Zhang 2012, Diaz-Ricci et al. 2000).

Although pMPPla107 megaplasmid negatively affects fitness, the magnitude of this effect is not much larger than what is observed for smaller plasmids (Park and Zhang 2012, Levin and Cornejo 2009), so that competitive fitness costs do not appear directly correlated with size. The consistency of fitness effects across divergent *Pseudomonas* strains further suggests that costs are not due to disruption of cellular networks by specific protein interactions, as the magnitude of these would presumably scale with DNA sequence divergence from the original parent strain. Lastly, we note that fitness costs within strains that have recently acquired pMPPla107 under laboratory conditions parallel growth phenotypes within *Pla*107*,* which naturally contains this megaplasmid (Baltrus et al. 2011). Therefore, it is possible that there has not been enough time for evolution to compensate for fitness costs in strains that naturally contain pMPPla107 or that other selective forces (i.e. high rates of transfer) dominate under natural conditions.

Because most strains can lose pMPPla107 under laboratory conditions, it does not appear that active toxin-antitoxin systems play a major role in maintaining this megaplasmid immediately after acquisition. However, we have not been able to isolate a megaplasmid free version of the original parent strain and it therefore remains a possibility that such toxin-antitoxin system are strain specific or only operate under certain environmental conditions. To this point, megaplasmid stability appears substantially lower in strain *Pto*DC3000 across medium types compared to both *Pla*8003 and *P. stutzeri.* It is also possible that mutations that stabilize pMPPla107 have occurred on the *Pla*107 chromosome or on the megaplasmid itself, or the *Pla*107 strain has evolved to be naturally more tolerant of extrachromosomal elements. If

true, this scenario could represent, to our knowledge, the first identified transitional state between chromid and secondary chromosome.

Table 4.3.
KEGG Pathway analysis ofselect ORFs from pMPPla107.

KEGG Pathway	# of Megaplasmid Genes Identified
Purine metabolism	7
Pyrimidine metabolism	6
DNA replication	6
Base excision repair	4
Nucleotide excision repair	3
Mismatch repair	7
Homologous recombination	7
Nicotinate/nicotinamide metabolism	3

Table 4.4.
A Subset of the "Housekeeping" Loci Present on Both the Main Chromosome and pMPPla107.

Annotation	Protein	Closest BlastP Hit	E value	Max ID
COG0305 Replicative DNA helicase	DnaB	*Nitrococcus mobilis*	1.00E-96	42%
COG0847 DNA polymerase III, epsilon subunit	DnaQ	*Glaciecola nitratirducens*	2.00E-51	42%
COG0587 DNA polymerase III, alpha subunit	DnaE	*Pseudomonas stutzeri*	0	51%
COG0358 DNA primase (bacterial type)	DnaG	*Pseudomonas stutzeri*	4.00E-78	38%
COG0050 Elongation factor Tu	EfTu	*Pseudomonas stutzeri*	0	88%
COG2974 DNA recombination-dependent growth factor	RdgC	*Thiocapsa marina*	3.00E-35	28%
COG0550 Topoisomerase IA	TopA	*Desulfavibrio*	1.00E-18	25%
COG0057 Glyceraldehyde-3-phosphate dehydrogenase	GapA	*Burkholderia phytofirmans*	7.00E-28	50%
COG0568 DNA-directed RNA polymerase, sugma subunit	RpoH	*Hyphomonas neptunium*	2.00E-16	27%
COG0568 DNA-directed RNA polymerase, sugma subunit	RpoS	*Maritimibacter alkaphilus*	6.00E-11	24%
COG0568 DNA-directed RNA polymerase, sugma subunit	RpoH	*Mariprofundus ferroxydans*	7.00E-14	28%

This work demonstrates both transmissibility and fitness costs associated with pMPPla107 and demonstrates that the Type IV secretion system harbored on this megaplasmid is essential for transmission. We show that pMPPla107 is self-transmissible, with transfer occurring at relatively high rates in liquid and on solid media. We also demonstrate that megaplasmid acquisition decreases host fitness under laboratory conditions across divergent strain backgrounds, and that loss of pMPPla107 can be selected for in all strain backgrounds except the original parent strain. These results suggest that high transfer rate could be a viable short-term evolutionary strategy ensuring maintenance of pMPPla107. Maintenance of

pMPPla107 by the *Pseudomonas* strain in which it is naturally found highlights the potential for mutational events to offset costs of HGT over evolutionary time scales. HGT can often facilitate ecological shifts by introducing novel genes into existing bacterial genomes, and a general first evolutionary step after transitional HGT events might indeed be acquisition of mutations that ameliorate inherent fitness costs (Davis et al. 2009, Poon et al. 2009, Katju and Lynch 2006). In the future, we will determine if compensatory mutations facilitate coevolution between the bacterial host genomes and megaplasmids in ways and rates that differ from those previously documented for small plasmids. Overall, the pMPPla107 system represents an exceptional and tractable model system with which to investigate costs of HGT as well as evolutionary dynamics as recently acquired regions are integrated into conserved cellular networks.

ACKNOWLEDGEMENTS

We would like to thank Mara Duncan for support in experimental execution, advice and use of her equipment. We would also like to thank Lisa Bono for help with manuscript preparation.

REFERENCES

Ashbolt, N.J., Amézquita, A., Backhaus, T., Borriello, P., Brandt, K.K., Collignon, P., Coors, A., Finley, R., Gaze, W.H., Heberer, T., Lawrence, J.R., Larsson, D.G.J., McEwen, S.A., Ryan, J.J., Schönfeld, J., Silley, P., Snape, J.R., Van den Eede, C., Topp, E., 2013. Human Health Risk Assessment (HHRA) for environmental development and transfer of antibiotic resistance. Environ. Health Perspect. 121, 993–1001. doi:10.1289/ehp.1206316

Baltrus, D.A., Guillemin, K., Phillips, P.C., 2008. Natural transformation increases the rate of adaptation in the human pathogen Helicobacter pylori. Evolution 62, 39–49. doi:10.1111/j.1558-5646.2007.00271.x

Baltrus, D.A., Nishimura, M.T., Dougherty, K.M., Biswas, S., Mukhtar, M.S., Vicente, J., Holub, E.B., Dangl, J.L., 2012. The molecular basis of host specialization in bean pathovars of Pseudomonas syringae. Mol. Plant Microbe Interact. 25, 877–888. doi:10.1094/MPMI-08-11-0218

Baltrus, D.A., Nishimura, M.T., Romanchuk, A., Chang, J.H., Mukhtar, M.S., Cherkis, K., Roach, J., Grant, S.R., Jones, C.D., Dangl, J.L., 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 7, e1002132. doi:10.1371/journal.ppat.1002132

Broaders, E., Gahan, C.G.M., Marchesi, J.R., 2013. Mobile genetic elements of the human gastrointestinal tract: potential for spread of antibiotic resistance genes. Gut Microbes 4, 271–280. doi:10.4161/gmic.24627

Choi, K.-H., Schweizer, H.P., 2006. mini-Tn7 insertion in bacteria with single attTn7 sites: example Pseudomonas aeruginosa. Nat. Protocols 1, 153–161. doi:10.1038/nprot.2006.24

Cooper, V.S., Vohr, S.H., Wrocklage, S.C., Hatcher, P.J., 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput. Biol. 6, e1000732. doi:10.1371/journal.pcbi.1000732

Davis, B.H., Poon, A.F.Y., Whitlock, M.C., 2009. Compensatory mutations are repeatable and clustered within proteins. Proc. Biol. Sci. 276, 1823–1827. doi:10.1098/rspb.2008.1846

De Gelder, L., Ponciano, J.M., Joyce, P., Top, E.M., 2007. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. Microbiology (Reading, Engl.) 153, 452–463. doi:10.1099/mic.0.2006/001784-0

De Gelder, L., Williams, J.J., Ponciano, J.M., Sota, M., Top, E.M., 2008. Adaptive plasmid evolution results in host-range expansion of a broad-host-range plasmid. Genetics 178, 2179–2190. doi:10.1534/genetics.107.084475

Diaz Ricci, J.C., Hernández, M.E., 2000a. Plasmid effects on Escherichia coli metabolism. Crit. Rev. Biotechnol. 20, 79–108. doi:10.1080/07388550008984167

Diaz Ricci, J.C., Hernández, M.E., 2000b. Plasmid effects on Escherichia coli metabolism. Crit. Rev. Biotechnol. 20, 79–108. doi:10.1080/07388550008984167

Gogarten, J.P., Townsend, J.P., 2005a. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. doi:10.1038/nrmicro1204

Gogarten, J.P., Townsend, J.P., 2005b. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. doi:10.1038/nrmicro1204

Gomes, A.L.C., Galagan, J.E., Segrè, D., 2013. Resource competition may lead to effective treatment of antibiotic resistant infections. PLoS ONE 8, e80775. doi:10.1371/journal.pone.0080775

Grad, Y.H., Lipsitch, M., Aiello, A.E., 2012. Secular trends in Helicobacter pylori seroprevalence in adults in the United States: evidence for sustained race/ethnic disparities. Am. J. Epidemiol. 175, 54–59. doi:10.1093/aje/kwr288

Harrison, E., Brockhurst, M.A., 2012a. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 20, 262–267. doi:10.1016/j.tim.2012.04.003

Harrison, E., Brockhurst, M.A., 2012b. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 20, 262–267. doi:10.1016/j.tim.2012.04.003

Heim, S., Ferrer, M., Heuer, H., Regenhardt, D., Nimtz, M., Timmis, K.N., 2003. Proteome reference map of Pseudomonas putida strain KT2440 for genome expression profiling: distinct responses of KT2440 and Pseudomonas aeruginosa strain PAO1 to iron deprivation and a new form of superoxide dismutase. Environ. Microbiol. 5, 1257–1269.

Heuer, H., Fox, R.E., Top, E.M., 2007. Frequent conjugative transfer accelerates adaptation of a broad-host-range plasmid to an unfavorable Pseudomonas putida host. FEMS Microbiol. Ecol. 59, 738–748. doi:10.1111/j.1574-6941.2006.00223.x

Katju, V., Lynch, M., 2006. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol. Biol. Evol. 23, 1056–1067. doi:10.1093/molbev/msj114

Kopfmann, S., Hess, W.R., 2013. Toxin-antitoxin systems on the large defense plasmid pSYSA of Synechocystis sp. PCC 6803. J. Biol. Chem. 288, 7399–7409. doi:10.1074/jbc.M112.434100

Kwon, S.-Y., Choi, Y.-J., Kang, T.-H., Lee, K.-H., Cha, S.-S., Kim, G.-H., Lee, H.-S., Kim, K.-T., Kim, K.-J., 2005. Highly efficient protein expression and purification using bacterial hemoglobin fusion vector. Plasmid 53, 274–282. doi:10.1016/j.plasmid.2004.11.006

Lambertsen, L., Sternberg, C., Molin, S., 2004a. Mini-Tn7 transposons for site-specific tagging of bacteria with fluorescent proteins. Environ. Microbiol. 6, 726–732. doi:10.1111/j.1462-2920.2004.00605.x

Lambertsen, L., Sternberg, C., Molin, S., 2004b. Mini-Tn7 transposons for site-specific tagging of bacteria with fluorescent proteins. Environ. Microbiol. 6, 726–732. doi:10.1111/j.1462-2920.2004.00605.x

Lau, B.T.C., Malkus, P., Paulsson, J., 2013. New quantitative methods for measuring plasmid loss rates reveal unexpected stability. Plasmid 70, 353–361. doi:10.1016/j.plasmid.2013.07.007

Lawrence, J.G., Ochman, H., 1998. Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. U.S.A. 95, 9413–9417.

Lawrence, J.G., Roth, J.R., 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143, 1843–1860.

Levin, B.R., Cornejo, O.E., 2009. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. PLoS Genet. 5, e1000601. doi:10.1371/journal.pgen.1000601

Loh, J.M.S., Proft, T., 2013. Toxin-antitoxin-stabilized reporter plasmids for biophotonic imaging of Group A streptococcus. Appl. Microbiol. Biotechnol. 97, 9737–9745. doi:10.1007/s00253-013-5200-7

Loper, J.E., Hassan, K.A., Mavrodi, D.V., Davis, E.W., 2nd, Lim, C.K., Shaffer, B.T., Elbourne, L.D.H., Stockwell, V.O., Hartney, S.L., Breakwell, K., Henkels, M.D., Tetu, S.G., Rangel, L.I., Kidarsa, T.A., Wilson, N.L., van de Mortel, J.E., Song, C., Blumhagen, R., Radune, D., Hostetler, J.B., Brinkac, L.M., Durkin, A.S., Kluepfel, D.A., Wechter, W.P., Anderson, A.J., Kim, Y.C., Pierson, L.S., 3rd, Pierson, E.A., Lindow, S.E., Kobayashi, D.Y., Raaijmakers, J.M., Weller, D.M., Thomashow, L.S., Allen, A.E., Paulsen, I.T., 2012. Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genet. 8, e1002784. doi:10.1371/journal.pgen.1002784

Park, C., Zhang, J., 2012. High expression hampers horizontal gene transfer. Genome Biol Evol 4, 523–532. doi:10.1093/gbe/evs030

Pennisi, E., 2004. Microbiology. Researchers trade insights about gene swapping. Science 305, 334–335. doi:10.1126/science.305.5682.334

Platt, T.G., Bever, J.D., Fuqua, C., 2012. A cooperative virulence plasmid imposes a high fitness cost under conditions that induce pathogenesis. Proc. Biol. Sci. 279, 1691–1699. doi:10.1098/rspb.2011.2002

Ponciano, J.M., De Gelder, L., Top, E.M., Joyce, P., 2007. The population biology of bacterial plasmids: a hidden Markov model approach. Genetics 176, 957–968. doi:10.1534/genetics.106.061937

Poon, A.F.Y., Swenson, L.C., Dong, W.W.Y., Deng, W., Kosakovsky Pond, S.L., Brumme, Z.L., Mullins, J.I., Richman, D.D., Harrigan, P.R., Frost, S.D.W., 2010. Phylogenetic analysis of

population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol. Biol. Evol. 27, 819–832. doi:10.1093/molbev/msp289

San Millan, A., Heilbron, K., MacLean, R.C., 2014. Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. ISME J 8, 601–612. doi:10.1038/ismej.2013.182

Shachrai, I., Zaslaver, A., Alon, U., Dekel, E., 2010. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. Mol. Cell 38, 758–767. doi:10.1016/j.molcel.2010.04.015

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J., 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336, 48–51. doi:10.1126/science.1218198

Sikorski, J., Graupner, S., Lorenz, M.G., Wackernagel, W., 1998. Natural genetic transformation of Pseudomonas stutzeri in a non-sterile soil. Microbiology (Reading, Engl.) 144 (Pt 2), 569–576.

Sikorski, J., Rosselló-Mora, R., Lorenz, M.G., 1999. Analysis of genotypic diversity and relationships among Pseudomonas stutzeri strains by PCR-based genomic fingerprinting and multilocus enzyme electrophoresis. Syst. Appl. Microbiol. 22, 393–402. doi:10.1016/S0723-2020(99)80048-4

Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P.C., de la Cruz, F., 2010. Mobility of plasmids. Microbiol. Mol. Biol. Rev. 74, 434–452. doi:10.1128/MMBR.00020-10

Sørensen, K.B., Canfield, D.E., Teske, A.P., Oren, A., 2005. Community composition of a hypersaline endoevaporitic microbial mat. Appl. Environ. Microbiol. 71, 7352–7365. doi:10.1128/AEM.71.11.7352-7365.2005

Sota, M., Tsuda, M., Yano, H., Suzuki, H., Forney, L.J., Top, E.M., 2007. Region-specific insertion of transposons in combination with selection for high plasmid transferability and stability accounts for the structural similarity of IncP-1 plasmids. J. Bacteriol. 189, 3091–3098. doi:10.1128/JB.01906-06

Starikova, I., Al-Haroni, M., Werner, G., Roberts, A.P., Sørum, V., Nielsen, K.M., Johnsen, P.J., 2013. Fitness costs of various mobile genetic elements in Enterococcus faecium and Enterococcus faecalis. J. Antimicrob. Chemother. 68, 2755–2765. doi:10.1093/jac/dkt270

Stoebel, D.M., Dykhuizen, D.E., 2010. Waste and yet want not. Mol. Cell 38, 625–626. doi:10.1016/j.molcel.2010.05.028

Unterholzner, S.J., Poppenberger, B., Rozhon, W., 2013. Toxin-antitoxin systems: Biology, identification, and application. Mob Genet Elements 3, e26219. doi:10.4161/mge.26219

Vogel, J.P., Andrews, H.L., Wong, S.K., Isberg, R.R., 1998. Conjugative Transfer by the Virulence System of Legionella pneumophila. Science 279, 873–876. doi:10.1126/science.279.5352.873

Warnes, S.L., Highmore, C.J., Keevil, C.W., 2012. Horizontal transfer of antibiotic resistance genes on abiotic touch surfaces: implications for public health. MBio 3. doi:10.1128/mBio.00489-12

Wolf, Y.I., Makarova, K.S., Yutin, N., Koonin, E.V., 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct 7, 46. doi:10.1186/1745-6150-7-46

CHAPTER FIVE: CONCLUSIONS

To understand how phenotypes evolve we must understand evolution of protein assemblages underlining phenotypes. Increasingly, gradual evolution is called into doubt as a causal agent for the entirety of adaptive change. Extreme genetic changes occur regularly and these often result in phenotypic adaptation. For example dramatic changes in gene regulation and also horizontal gain of new genes can both lead to radical adaptive change, (Feng et al. 2014, Rapa and Labbate 2013, Fernandez and Backert 2014, Schonknecht et al 2014, Englstadter and Moradigaravand 2014) although radical change can be extremely deleterious to organism's fitness.

With this in mind my work explores fitness cost of radical genomic change in bacteria. In my work I have shown that as expected fitness cost of horizontally transferred genes (HGT) can be mediated through organism's genetic architecture (protein-protein interactions), divergence and ecological plasmid dynamics. In my dissertation I identify a special set of genes, which lower host's fitness greatly during homologous HGT regardless of their protein-protein interaction architecture (Chapter 2). These genes are responsible for regulating genetic information (DNA) within the cell and are collectively called Informational genes. I also demonstrate that HGT fitness cost of a subset of divergent *E. coli* homologs (including informational genes) is mainly driven by divergence (Chapter 3). I finally explore horizontal transfer dynamics of a magaplasmid capable of transferring ~1000 ORFs and show that large deleterious but rapidly conjugating megaplasmids are resistant to selective loss from host bacterial populations (Chapter 4).

I was surprised to find little support for the Complexity Hypothesis (R.A. Jain 1999) which links fitness cost of divergent HGT to disruption of protein interactions architecture (presumably due to divergence in protein regions responsible for said interactions). Considering the limits of the Complexity Hypothesis - that HGT of complete homologs is a lower bound of the Complexity Hypothesis just like HGT of completely novel non-homologous genes is it's upper bound. No fitness cost should have been observed for transfer of non-divergent homologs. Instead exogenous expression of non-divergent homologs was deleterious for genes with low protein connectivity and less so for genes with high protein connectivity. These data indicate that at least at the lower limit of divergence the Complexity Hypothesis provides false predictions regarding protein connectivity.

This work does support the claim of the Complexity Hypothesis that gene function of exogenously expressed genes affects fitness cost of HGT. In our case Metabolic genes were overall less costly and Informational genes were most costly. Perhaps constant levels of certain proteins are required continuously throughout the cell's lifecycle. Metabolic proteins may be constantly in demand, providing energy and materials for cell growth. Informational proteins may be required sporadically during times of cell division only. This unbalanced demand structure may be why exogenous continuous expression of metabolic proteins decreased fitness less. Clear counter examples are ribosomal genes. These Informational proteins are likely to be in constant demand by the cell because they are responsible for producing all other proteins. However consistent with the line of reasoning presented here ribosomal genes, unlike most other Informational genes, were among the least costly exogenously expressed genes.

Exogenous expression of divergent homologs in the *E.coli* background did reveal the expected pattern between increased fitness cost and increased protein divergence. Fitness cost

was not sensitive to the interaction between divergence and protein connectivity as predicted by the Complexity Hypothesis. Possibly indicating that large fitness cost of divergent genes is due to their unbalancing effect on molecular stochiometry. These results support the ideas proposed under the Balance Hypothesis (Papp et al. 2003), mainly that protein function inefficiencies cause cells to increasingly waste molecular resources. Similarly to the Complexity Hypothesis under the Balance Hypothesis highly connected divergent proteins are expected to be most deleterious when exogenously expressed. However unlike the Complexity Hypothesis under the Balance Hypothesis highly connected divergent proteins are deleterious precisely because they bind all of their partners and thus decrease efficiency of numerous proteins in the cell.

Regardless of protein interaction architecture effect on fitness and gene function effect on fitness HGT vehicle ecology can shield genes from purifying selection. My findings reveal a deleterious megaplasmid that is virtually incurable from bacteria host populations. Although the pMPPla107 megaplasmid decreases hosts fitness on average by 20% the high efficiency with which this megaplasmid conjugates constantly allows it to re-infect plasmid free cells. Thus the nearly 1000 ORFs carried by this plasmid are constantly exposed to purifying selection imposed by the bacterial host genome and ecology. Such hyper-infectious dynamic suggest that over time through action of mutation and natural selection pMPPla107 ORFs will evolve to ameliorate their originally deleterious fitness effects. As such the hyper-infectious megaplasmid pMPPla107 in particular, and others like it, comprise a powerful evolutionary pool.

I am confident that future experimental investigations into fitness effects of homologous and non-homologous HGT coupled with deep understanding of plasmid ecology will strengthen the conclusions and reconcile the inconsistencies in the data presented here. The result will be more complete understanding of how radical evolutionary events complement more gradual

evolutionary processes in creating and maintaining adaptive potential of all living forms. More explicitly further understanding of how protein interaction architecture impacts adaptive evolution will finally reveal how incredibly complex molecular systems (*i.e.* the ribosome) have evolved from simpler assemblages of only a few molecular partners.

REFERENCES

Engelstädter, J., Moradigaravand, D., 2014. Adaptation through genetic time travel? Fluctuating selection can drive the evolution of bacterial transformation. Proc. Biol. Sci. 281, 20132609. doi:10.1098/rspb.2013.2609

Feng, S., Powell, S.M., Wilson, R., Bowman, J.P., 2014. Extensive gene acquisition in the extremely psychrophilic bacterial species Psychroflexus torquis and the link to sea-ice ecosystem specialism. Genome Biol Evol 6, 133–148. doi:10.1093/gbe/evt209

Fernandez-Gonzalez, E., Backert, S., 2014. DNA transfer in the gastric pathogen Helicobacter pylori. J. Gastroenterol. doi:10.1007/s00535-014-0938-y

Papp, B., Pál, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197. doi:10.1038/nature01771

Rapa, R.A., Labbate, M., 2013. The function of integron-associated gene cassettes in Vibrio species: the tip of the iceberg. Front Microbiol 4, 385. doi:10.3389/fmicb.2013.00385

Schönknecht, G., Weber, A.P.M., Lercher, M.J., 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36, 9–20. doi:10.1002/bies.201300095