ORIGINAL ARTICLE

Patterns of variation during adaptation in functionally linked loci

Diamantis Sellis[1,2] and Mark D. Longo[1]

[1]*Department of Biology, Stanford University,Stanford, California 94305*

[2]*E-mail: sellisd@gmail.com*

An understanding of the distribution of natural patterns of genetic variation is relevant to such fundamental biological fields as evolution and development. One recent approach to understanding such patterns has been to focus on the constraints that may arise as a function of the network or pathway context in which genes are embedded. Despite theoretical expectations of higher evolutionary constraint for genes encoding upstream versus downstream enzymes in metabolic pathways, empirical results have varied. Here we combine two complementary models from population genetics and enzyme kinetics to explore genetic variation as a function of pathway position when selection acts on whole-pathway flux. We are able to qualitatively reproduce empirically observed patterns of polymorphism and divergence and suggest that expectations should vary depending on the evolutionary trajectory of a population. Upstream genes are initially more polymorphic and diverge faster after an environmental change, while we see the opposite trend as the population approaches its fitness optimum.

KEY WORDS: Enzyme kinetics, fine-tuning, metabolic pathway, population genetics, selection.

Introduction

Genes work together to create phenotypes yet historically evolutionary analyses have tended to focus on one gene at a time. With the advent of high-throughput sequencing technology, research on the genome-scale has exploded. Despite this shift in scope, analyses have remained largely amechanistic and correlative. The bulk of genomic studies continue to focus on statistical relationships among genes and their correlated phenotypes while neglecting the rich implications of both functional linkage, which involves organizational and mechanistic aspects of gene interactions, and the more general influence of network context and topology. As an increasing body of literature indicates, such "system-level" properties may influence the ability of component genes to evolve, both in terms of adaptive change (see e.g., Flowers et al. 2007; Wright and Rausher 2010) and selective constraint (see e.g., Fraser et al. 2003).

By "functional linkage" we mean here the interaction of genes to produce phenotypic traits whether or not these genes are physically linked on chromosomes. Although genes are arranged on linear chains of DNA, they code for products that interact in vast networks across diverse scales of biological hierarchy. These networks are arranged in modular fashion with some gene products interacting more strongly with each other than with the gene products of other groups. Depending on the position of a gene within a module or of a module within the greater network, effects on phenotype, and hence fitness, may vary.

The importance of functional linkage and network structure to the phenotypic effects of a gene may be illustrated in the simple case of a linear metabolic pathway at steady state. Under these conditions, an enzyme's position in the pathway affects how dramatically a change in its activity will modify overall pathway flux (rate of metabolic production) (Kacser and Burns 1973; Savageau 1976). For any given reaction, an equilibrium constant (K_{eq}) describes the thermodynamically favored concentrations of substrate and product. Enzymatic control of the first reaction of a metabolic pathway is constrained by only one K_{eq}, but control of flux further downstream is constrained by all of the K_{eq} values of previous steps. Thus, all else being equal (e.g., expression levels), a functional change in an upstream enzyme would likely have a larger effect on the final product than an equivalent change at a downstream enzyme (Kacser and Burns 1973; Hartl et al. 1985).
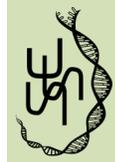
Given that natural selection is expected to target enzymes that influence flux (Watt and Dean 2000) and that upstream enzymes tend to have majority flux control, it has been hypothesized that genes coding upstream enzymes should be more evolutionarily constrained, and thus less polymorphic, than genes coding for downstream enzymes (Rausher et al. 1999; Cork and Purugganan 2004). Equivalently, when comparing orthologous pathways across species, it is expected that upstream genes should be more conserved than downstream genes.

More broadly, we should expect that in signaling and other networks the amount of polymorphism observed in a gene will be influenced by its location in the network. In the case of signaling pathways, changes in genes encoding upstream proteins should have greater effects on fitness and hence be more evolutionarily constrained than genes encoding downstream proteins. In the case of more complex network topologies, changes in genes encoding proteins that are more centrally embedded would lead to relatively greater functional effects, and hence these genes would be expected to be relatively more evolutionarily constrained than those encoding less central proteins.

Although the study of genetic variation has a long history in population genetics, few studies (e.g., Wright and Rausher 2010; Rausher 2013) have explored the theoretical implications of functional linkage or network context as they relate to selection and subsequent patterns of variation. Despite this overall paucity of theory, a number of empirical studies have analyzed natural patterns of polymorphism and divergence across multiple pathways and networks in a wide range of species. Results have differed dramatically with some researchers finding the expected pattern of disproportionate evolutionary constraint in upstream and centrally embedded genes and others finding either the opposite trend or no trend at all (Tables 1 and 2). Even in the simple case of a linear metabolic pathway there are numerous reasons to expect some discordance between empirical findings. Linear pathways often function in either direction depending on factors such as the physiological state, the cell-type or cell-cycle stage. Hence it is sometimes unclear what "position" an enzyme should be assigned in a pathway. Even delineating the limits of a pathway itself is not always straightforward (Schuster et al. 2000). Furthermore, pathways often have branched topologies rather than simple linear arrangements, and diverse biochemical pathways may share identical enzymes (Greenberg et al. 2008). These difficulties are compounded in multicellular organisms where the same genes may encode proteins that have completely different functions in different cell types (Piatigorsky 2007). Furthermore, multiple intermediate products may be under selective pressure rather than just the downstream products of the pathway (Watt 1986). Also, given that any pathway will itself constitute one component of a greater hierarchical metabolic topology, it is not clear at which level of this hierarchy selection would act most strongly. There

Table 1. Patterns of polymorphism (intraspecies variation).

	Taxa/species		System Analyzed	Independent variable	Reference
Follows expectations	Mammals:	*Homo sapiens*	Asparagine N-glycosylation metabolic pathway	Pathway rank	(Dall'Olio et al. 2012)
Opposite pattern	Insects:	*Bombyx morii, B. mandarina*	Melanin synthesis metabolic pathway	Pathway rank	(Yu et al. 2011)
	Insects:	*Drosophila*	Insulin/TOR signal transduction pathway	Pathway rank	(Alvarez-Ponce et al. 2012)
No polarity or complex pattern	Plants:	*Arabidopsis thaliana*	Phenylpropanoid metabolic pathway	Pathway rank	(Ramos-Onsins et al. 2008)

Expectations (as discussed in the text) are that upstream or more central genes should be less polymorphic than downstream or less central genes.

Table 2. Patterns of divergence (interspecies variation).

	Taxa/species	System analyzed	Independent variable	Reference
Follows expectations	Plants: *Zea mays, Antirrhinum majus, Ipomoea purpurea*	Anthocyanin biosynthesis metabolic pathway	Pathway rank	(Rausher et al. 1999)
	Plants: *Ipomoea trifida, I. purpurea*	Anthocyanin biosynthesis metabolic pathway	Pathway rank	(Rausher et al. 2008)
	Plants: *Solanum lycopersicum, Capsicum annuum, Gentiana lutea, Chrysanthemum x morifolium, Citrus spp., Daucus carota, Oryza sativa*	Carotenoid biosynthesis metabolic pathway	Pathway rank	(Livingstone and Anderson 2009)
	Plants: *D. carota, Chaerophyllum bulbosum*	Carotenoid biosynthesis metabolic pathway	Pathway rank	(Clotault et al. 2012)
	Plants: *O. sativa, Vitis vinifera, A. thaliana, Populus trichocarpa, Ricinus communis*	terpenoid biosynthesis metabolic pathway	Pathway rank	(Ramsay et al. 2009)
	Plants: *Pueraria montana, A. thaliana, Populus tremuloides*	Isoprene biosynthetic metabolic pathway	Pathway rank	(Sharkey et al. 2005)
	Mammals: *H. sapiens, Pan troglodytes, Macaca mulatta, Mus musculus, Rattus norvegicus*	Dopamine metabolic pathway	Pathway rank	(Ma et al. 2010)
	Mammals: *H. sapiens, P. troglodytes, Gorilla gorilla, Pongo pygmaeus, M. mulatta*	N-glycosylation metabolic network	Degree and closeness centrality	(Montanucci et al. 2011)
	Yeast: *Saccharomyces cerevisiae, S. paradoxus, S. bayanus, S. mikatae*	Genome-scale metabolic network	Closeness centrality, flux magnitude	(Vitkup et al. 2006)
	Mammals: *H. sapiens, P. troglogdytes, M. mulatta, M. musculus, R. norvegicus, Equus ferus, Canis lupus, Bos taurus*	Genome-scale metabolic network	Betweenness centrality	(Hudson and Conant 2012)

(Continued)

Table 2. Continued.

	Taxa/species		System analyzed	Independent variable	Reference
	Insects:	*Drosophila* (12 spec.)	Genome-scale metabolic network	Degree centrality	(Greenberg et al. 2008)
	Insects:	*D. melanogaster, D. simulans*	RAS-mediated signal transduction pathway	Pathway rank	(Riley et al. 2003)
	Mammals and birds:	*H. sapiens, M. musculus, B. taurus, Monodelphis domestica, Ornithorhynchus anatinus, Gallus gallus*	Genome-scale signal transduction network	Degree centrality	(Alvarez-Ponce et al. 2011)
	Yeast, worm, insect:	*S. paradoxus, Caenorhabditis briggsae, Drosophila pseudoobscura*	Genome-scale protein–protein interaction network	betweenness and closeness centrality	(Hahn and Kern 2005)
	Yeast:	*S. cerevisiae, S. paradoxus*	Genome-scale protein–protein interaction network	Degree centrality	(Hahn et al. 2004)
	Yeast:	*S. cerevisiae, Candida albicans, Schizosaccharomyces pombe*	Genome-scale protein–protein interaction network	Degree centrality	(Fraser et al. 2003)
Opposite pattern	Mammals:	*H. sapiens, P. troglodytes, G. gorilla, Pongo pygmaeus, M. mulatta*	N-glycosylation metabolic pathway	Pathway rank	(Montanucci et al. 2011)
	Insects:	*Drosophila* (12 spec.)	Insulin/TOR signal transduction pathway	Pathway rank	(Alvarez-Ponce et al. 2009)
	Worm:	*Caenorhabditis* (8 spec.)	Insulin/TOR signal transduction pathway	Pathway rank	(Jovelin and Phillips 2011)
	Yeast:	(10 spec.)	HOG signal transduction pathway	Pathway rank	(Wu et al. 2010)
	Mammals and birds:	*M.musculus, B. taurus, M. domestica, O. anatinus, G. gallus*	Insulin/TOR signal transduction pathway	Pathway rank	(Alvarez-Ponce et al. 2011)
	Mammals:	*H. sapiens,* 8 mammalian genomes	Genome-scale signal transduction pathway	Pathway rank	(Alvarez-Ponce 2012)

(Continued)

Table 2. Continued.

	Taxa/species	System analyzed	Independent variable	Reference
No polarity or complex pattern	Plants: *A. thaliana, A. lyrata*	Phenylpropanoid metabolic pathway	Pathway rank	(Ramos-Onsins et al. 2008)
	Plants: *y: Oryzeae* (5 sp.), *Chikusichloa aqautica, Luziola leiocarpa, Rhynchoryza subulata, Ehrharta erecta*	Gibberellin metabolic pathway	Pathway rank	(Yang et al. 2009)
	Mammals: *H. sapiens, P.troglogdytes, M. mulatta, M. musculus, R. norvegicus, E. ferus, C. lupus, B. taurus*	Genome-scale metabolic network	Degree centrality	(Hudson and Conant 2011)
	Bacteria: *Escherichia coli, Haemophilus influenze*	Genome-scale metabolic network	Centrality degree	(Hahn et al. 2004)
	Mammals: *H. sapiens, P. troglodytes, G. gorilla, P. pygmaeus, M.mulatta*	N-glycosylation metabolic network	betweenness centrality	(Montanucci et al. 2011)

Expectations (as discussed in the text) are that upstream or more central genes should be less divergent than downstream or less central genes.

is also the possibility that the magnitude and targets of selection are changing through time during adaptation. Finally, differences among genes such as mutation rate, differential expression levels, or posttranslational modifications may all contribute to evolutionary constraints independent of a gene's position along a pathway.

While any number of the above complications may explain discrepancies between empirical findings and theoretically expected patterns of variation, we argue that there is an even more basic explanation that has not been fully explored. We develop a simulation combining a forward population genetic model that can describe population variation and a model of first-order enzyme kinetics of a linear pathway that links genotypes to fitness. We model adaptation in an initially monomorphic haploid population that evolves after a change in the environment. We find that the dynamics of adaptation involve fine-tuning of fitness in a process that is quite different from single gene adaptation dynamics. As a result the pattern of divergence and polymorphism strongly depends on the history of the evolving population, and particularly on the stage of the adaptive process. During adaptation, two opposite patterns of variation are observed in our model despite equal mutation rates across all genes of the pathway. At the early stages of the adaptive walk, the genes coding for upstream enzymes are more polymorphic. As the population approaches the optimum, the selective pressure switches from directional to stabilizing and the genes coding for downstream enzymes become the most polymorphic. Patterns of divergence (i.e., genetic differences between populations) change in a similar way, with substitutions initially appearing in the upstream genes and subsequently in the downstream genes. Using this minimal model we are therefore able to reproduce a range of patterns corresponding to those that have been empirically reported.
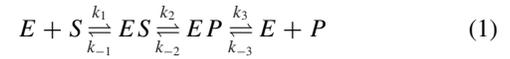
Methods

We implement a Wright–Fisher infinite alleles model of a haploid genome of n loci that describes the evolutionary dynamics of a linear metabolic pathway. The model has two components: a lower level that describes the development of phenotypes from genotypes and a higher level that describes the evolution of a population composed of these phenotypes. At the lower level, phenotypes are based on an enzyme kinetics model of a metabolic pathway composed of n enzymes. Each enzyme is encoded by one gene, and the set of genes encoding all the enzymes of a pathway comprise that pathway's genotype. The steady-state concentrations of substrates and products depend on the kinetics of all constituent enzymes and are calculated for each unique genotype. The absolute fitness of each genotype is determined by the rate of production of the final product of its corresponding metabolic pathway. At the population level, our model implements mutation,

selection, and drift and tracks resultant gene frequency changes across generations.
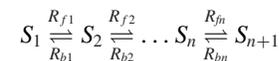
METABOLIC PATHWAY MODEL
The kinetics of an enzymatic reaction can be modeled as a three-step process of association, catalysis, and dissociation:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} EP \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} E + P \qquad (1)$$

where E is the enzyme, S the substrate, P the product and k_i and k_{-i} are the forward and backward rates of a reaction. The evolution of the system can be modeled by mutating all six rate parameters and changing the enzyme concentration. We know that mutations can produce enzymes with multiple combinations of parameters affected and with variable magnitudes. We therefore aggregate all the rate parameters and enzyme concentration in a forward $R_f = (V_{max}/K_m)_f$ and a backward parameter $R_b = (V_{max}/K_m)_b$. This involves an assumption that enzymes are at sufficiently high concentration that they are substantially below saturation. Thus, we approximate (1) by:

$$S \underset{R_b}{\overset{R_f}{\rightleftharpoons}} P$$

and link n such reactions with the product of one being the substrate of the next one (both products and substrates now denoted by S) to create a linear pathway

$$S_1 \underset{R_{b1}}{\overset{R_{f1}}{\rightleftharpoons}} S_2 \underset{R_{b2}}{\overset{R_{f2}}{\rightleftharpoons}} \ldots S_n \underset{R_{bn}}{\overset{R_{fn}}{\rightleftharpoons}} S_{n+1}$$

composed of n enzymes (E_i, $i = 1 \ldots n$) catalyzing n reversible reactions. The reversibility parameter $\alpha_i = 1/K_{eq} = R_{bi}/R_{fi}$ links the two parameters of each reaction, and depends only on the thermodynamics of the reaction (i.e., it is not mutable) (Wright and Rausher 2010). We assume the same reversibility parameter α for all reactions in the linear pathway. According to the law of mass action, the rate of change of the substrate concentration $[S_i]$ is:

$$\frac{d[S_i]}{dt} = [S_{i-1}]R_{fi} - 1 + [S_{i+1}]R_{bi} - [S_i](R_{bi-1} + R_{fi}) \quad (2)$$

The concentration of the first substrate ($[S_1]$) and of the final product ($[S_{n+1}]$) are constant ($\frac{d[S_1]}{dt} = 0$, $\frac{d[S_{n+1}]}{dt} = 0$). Their values are set such that their ratio is half of the equilibrium ratio that ensures that there is flux through the pathway ($S_1 = 1$, $S_{n+1} = \frac{1}{2} \cdot \prod_{i=1}^{n} \alpha_i$). Thus, we are able to study the evolution of the enzymes in the pathway as a single module, assuming that enzymes upstream and downstream of the pathway do not evolve.

Fitness is determined by the rate of production (J) of the final product ($[S_{n+1}]$) when all metabolite concentrations are at steady state:

$$J = [S_n]R_{fn} - [S_{n+1}]R_{bn} \qquad (3)$$

The absolute fitness (W) is given by a Gaussian function centered on the optimum rate of production (J_{opt}) with standard deviation $\sigma_w = 0.2$,

$$W = e^{-\frac{(J-J_{opt})^2}{2\sigma_w^2}} \qquad (4)$$

The choice of a fixed optimum and steady-state values assumes the dynamics of steady-state metabolism.

POPULATION GENETIC SIMULATION

We model the evolution of a population of size N as a Wright–Fisher process by explicitly simulating genotypes in an infinite alleles model. Each enzyme is encoded by a gene G_i and mutations are modeled by a Poisson process with a mutation rate, μ, for each gene. We model the magnitude of mutation effects, m, as displacements added to the forward parameter, R_{f_i}, of the mutated gene. The backward rate parameter R_{b_i} changes accordingly as α is constant. We restrict the values of $R_{f_i} + m$ between 0 and 1. For all simulations m is sampled from a normal distribution centered at R_{f_i} and with standard deviation σ_μ. In the description of the landscape topology we use a uniform distribution of mutation sizes ($m \sim U(0, m_{max})$). We explore a range of values for the population genetic parameter $\Theta = 2N\mu$ ($\Theta = 10^{-3}, 10^{-2}, 0.1, 1, 10$), the standard deviation of distribution of mutation effects $\sigma_\mu = 0.01, 0.05, 0.1, 0.5$, the reversibility parameter $\alpha = 0.001, 0.01, 0.05, 0.5, 1$ and the strength of selection $\sigma_w = 0.01, 0.1, 0.2, 0.5$.

SIMULATION PARAMETERS

In all simulations we assume no recombination between the genes, and $J_{opt} = 0.5$. Results in the main text refer to simulations of a two-gene pathway, though we also simulated four-gene genotype evolution to verify that similar patterns emerge in longer pathways (Supporting Information). We perform two sets of 10,000 simulation runs starting from a monomorphic population with random combinations of rates (R_{f_i}) such that the fitness of the initial genotype is either far ($0.19 < W < 0.21$) or close to the optimum ($0.79 < W < 0.81$). The population size is set to 10^3 and the total simulation time is 10^4 generations. To study patterns of polymorphism across genes we measure the number of alleles present in the population at each locus at specified time intervals ($t = 10, 10^2, 10^3, 10^4$) in the simulations starting far from the optimum. For the study of patterns of divergence during adaptation we calculate the time point when the population first reaches the optimum. We consider the population adapted when the mean fitness reaches the neutral regime, the point at which fitness gains are offset by drift. This is usually considered to be when $\bar{W} > 1 - \frac{1}{N}$, but because the neutral regime boundary is not necessarily sharp we use $\bar{W} > 1 - \frac{5}{N}$ as our cutoff. For the simulations where the population successfully reached adaptation, we calculate the

total number of mutations for each gene in the most fit genotype. For the study of patterns of divergence after the population has reached the optimum we perform a set of 1000 simulations with the same parameters as before, but for $3 \cdot 10^5$ generations so there is enough time to observe the change in the divergence pattern.

Simulations are written in C++ using the GNU scientific library (Galassi et al. 2009). Source code is available at http://multigene.sourceforge.net/ and http://dx.doi.org/10.5061/dryad.bf597. Simulation documentation is included in the Supporting Information.

Results

To understand variation patterns in metabolic pathways we explore the geometry of the fitness landscape and perform simulations of populations adapting to a new environment. We study the patterns of polymorphism within each population during adaptation and the patterns of divergence of an adapted genotype from its ancestral state.

FITNESS LANDSCAPE

We can use the shape of the fitness landscape in concert with population genetic theory to get some intuition about the expected dynamics of adaptation and subsequent patterns of variation. In the case of mutation limited regimes $\Theta \ll 1$, the subsequent lack of variation in the population would make the search for patterns of polymorphism irrelevant as there would be no variation to examine. We thus focus on the case where adaptation is not mutation limited ($\Theta \gg 1$) such that multiple mutations are present in the population at the same time. New mutations are sampled from a uniform distribution of values, with maximum value 1 ($m \sim U(0, 1)$) that allows a genotype to reach all values of R_{f_i} with a single mutation in the corresponding gene. We now focus on a genotype coding a pathway of two enzymes that has a fitness landscape that is easy to visualize and understand. The fitness landscape is a surface in a 3 dimensional space that includes the rates R_{f_1}, R_{f_2}, and fitness W as axes. The landscape is smooth with a single ridge of high fitness (Fig. 1A). The value of the optimal flux J_{opt} determines the location of the ridge. Large values of J_{opt} correspond to landscapes with a high fitness ridge located at the large values of R_{f_i} and small values of J_{opt} correspond to landscapes with a high fitness ridge located at the small values of R_{f_i}. When reversibility (α) is low, the rate of the upstream gene R_{f_1} is the most important determinant of fitness, as small changes of R_{f_1} cause large fitness changes (Fig. 1). When α is large, the landscape is more symmetric and both enzymes have similar contributions to fitness. In a monomorphic population located far from the optimum ($R_{f_1} = 0.15$, $R_{f_2} = 0.15$, $W = 0.2$) (Fig. 1B) the rate of beneficial mutations in the two genes is very
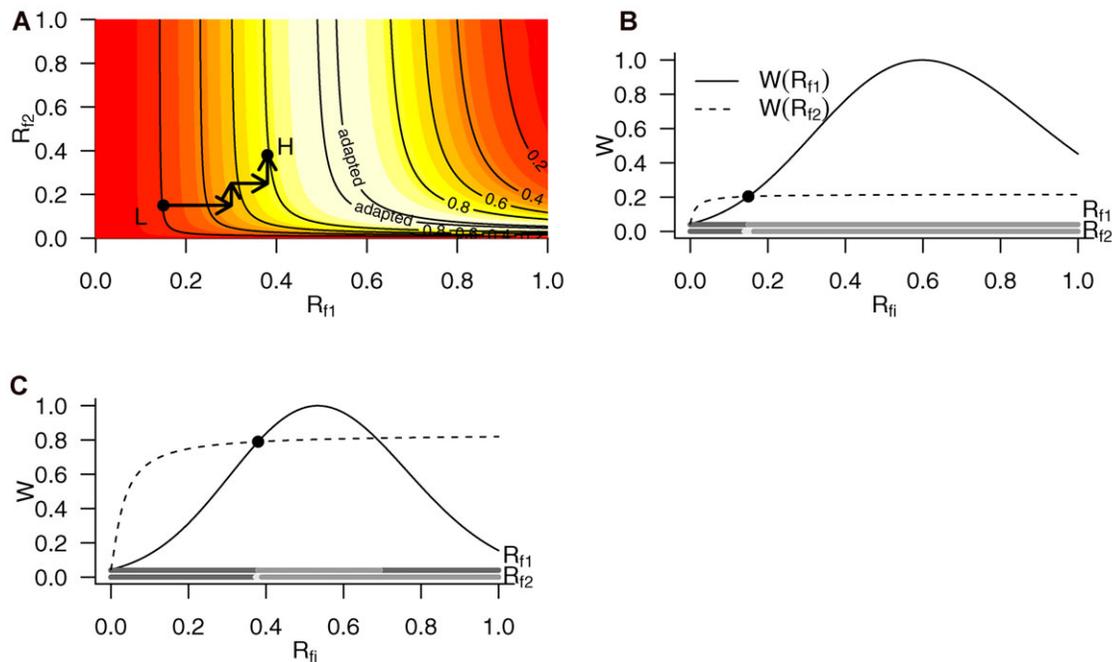
Figure 1. (A) Fitness landscape of a pathway with two enzymes. Each point in the surface corresponds to a unique genotype. Light color corresponds to genotypes with high fitness, dark to genotypes with low fitness. The value of the absolute fitness is marked on the fitness isoclines. Mutations in each gene displace the genotype either horizontally or vertically. An example trajectory is shown with black dots corresponding to the successive genotypes and black arrows to the mutations linking them. Note that mutations in the second gene, corresponding to movement along the y-axis, are mostly parallel to the fitness isoclines. (B) Two one-dimensional slices of the fitness landscape are superimposed. The continuous line corresponds to fitness along a horizontal line passing through point "L" in Figure 1A and the dashed line to fitness along a vertical line passing through point "L." The bars along the x-axis correspond to regions of beneficial (light) neutral (white) or deleterious (dark) mutations. (C) As in B, but for point "H" in Figure 1A.

similar. However, beneficial mutations in the upstream gene have a much larger fitness effect than beneficial mutations of similar magnitude in the downstream gene. We thus expect that adaptation far from the optimum will be driven by adaptive mutations in the upstream gene. In a monomorphic population located close to the optimum ($R_{f_1} = 0.38$, $R_{f_2} = 0.38$, $W = 0.8$) (Fig. 1C) the rate of beneficial mutations is different between the upstream and downstream genes. In this case most new mutations in the upstream gene are strongly deleterious and will tend to reduce the amount of polymorphism in the upstream gene. The downstream gene is not as constrained in this manner. In a very fit monomorphic population (located on top of the ridge), most mutations in the downstream gene are neutral, while most mutations in the upstream gene are deleterious.

Combining the above, in a population that is in the process of climbing a hill toward a fitness peak (directional selection), beneficial mutations are available to drive adaptation and the amount of polymorphism is determined by the supply of new beneficial mutations (Desai and Fisher 2007). On the contrary, in a population that is already located on the peak of the landscape (stabilizing selection), most new mutations will be deleterious and therefore the polymorphisms present in the population will

be primarily determined by a selection-mutation balance (Wright 1938). During adaptation to a new environment we expect to see a transition from the directional selection regime to the stabilizing selection regime. Given that this transition would not necessarily happen synchronously for all enzymes of the pathway (Supporting Information Fig. S5) due to the underlying hierarchical structure of our model, we should expect to see evidence of a "fine-tuning" mode of adaptation resulting in a temporal partitioning of variation across the genes in the pathway. In this fine-tuning mode of adaptation, variation should initially be driven by large mutational steps in the upstream gene followed by smaller effect mutations in the downstream gene, leaving characteristic signatures of polymorphism and divergence.

We visualize the modes of adaptation on the genotype space by plotting the difference of the fixation probabilities of mutations in the upstream and downstream genes. We assume a population size of 10^3 and m $\sim$ U[0,1] and calculate fitness W_{ij} for a dense grid of genotypes (R_{f_1}, R_{f_2}). For each genotype $R_{f_1} = i$, $R_{f_2} = j$ we calculate the average selection coefficients ($\bar{s}_m$) of beneficial mutations ($\Pr_{ben} = \Pr[W_m > (W_{ij} + 5/N)]$) in the upstream and downstream genes. We then calculate the product of the average probability of fixation of mutations $\Pr_{fm} = \frac{1-e^{2\bar{s}_m}}{1-e^{-2\bar{s}_m N}}$ for
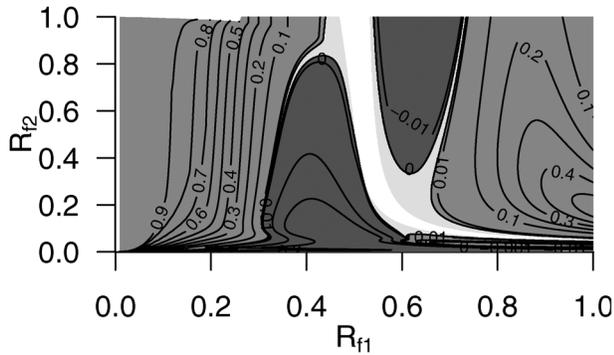
Figure 2. Far from the optimum adaptation proceeds with large effect mutations in the upstream gene (light shade area), while close to the optimum ridge it proceeds by small effect mutations in the downstream gene (dark shade area). Contour lines represent the difference between invasion probabilities of a new mutation in the upstream and downstream gene. The white area represents the neutral region where the difference in probabilities is less than $5/N$. The invasion probability at each point in the graph is calculated as the probability of a genotype acquiring a beneficial mutation with a fitness effect large enough to overcome drift and fix in a monomorphic population. We assume $m \sim U(0, 1)$ and $N = 10^3$.

each gene and the probability of a mutation being beneficial for each gene ($\mathrm{Pr}_{fixation}^{upstream} = \mathrm{Pr}_{fm}^{upstream} \mathrm{Pr}_{ben}^{upstream}$ or $\mathrm{Pr}_{fixation}^{downstream} = \mathrm{Pr}_{fm}^{downstream} \mathrm{Pr}_{ben}^{downstream}$), which corresponds to the fixation probability of mutations in each gene in a monomorphic population of genotype i, j. We then plot the difference $\mathrm{Pr}_{fixation}^{upstream} - \mathrm{Pr}_{fixation}^{downstream}$ that indicates whether adaptive mutations in the upstream or downstream gene will tend to sweep in the population (Fig. 2).

We observe three regions: one "coarse-grained" region in which genotypes containing new mutations of large effect have a greater probability of invading, a fine-tuning region in which genotypes containing new mutations of smaller effect have a greater probability of invading, and a neutral region in which the difference in these probabilities is less than $5/N$. In the first coarse-grained region, genotypes have low fitness and the most likely mutant to invade a monomorphic population will be one carrying a mutation in the upstream gene. In the fine-tuning region, genotypes have high fitness and the most likely mutant to invade will be one with a downstream gene mutation. In populations with smaller size, the neutral regime around the fitness peak is larger, reducing the fine-tuning ability of the adapting population.

Thus, if adaptation is not mutation limited ($\Theta \gg 1$), during an adaptive walk starting with a monomorphic population far from the optimum multiple adaptive mutations will be present in the population at the same time and most of them will be in the upstream gene. The transition from directional to stabilizing selection is expected to leave a characteristic signature in the order of substitutions across the genes of a metabolic pathway,

with mutations first fixing in the upstream gene. A population evolving in the coarse-grained region of adaptation is expected to be more polymorphic in the upstream gene that will also have a higher rate of divergence. On the contrary, a population evolving in the fine-tuning region is expected to have a reduced amount of polymorphism in the upstream gene and a higher rate of polymorphism in the downstream gene. Thus, as long as an adaptive walk crosses the boundary between the two regions, we expect to see a transition in the observed patterns of variation. Because the maximal contributions to flux of downstream genes in a directional pathway will always be subject to the equilibrium constants of all upstream enzymes (as described in the introduction), patterns of polymorphism, and divergence are expected to remain qualitatively similar for pathways of two or more genes.

PATTERNS OF POLYMORPHISM

The picture of variation that emerges from a static analysis of the fitness landscape is validated by simulating adaptive walks (see Supporting Information: 1. Parameter exploration). The simulated populations are asexual, genes are linked and we repeat the simulations for pathways of 2 and 4 genes. Simulations start from monomorphic populations with low fitness and at specific time intervals we measure the number of alleles in the population at each locus. We find for a wide range of parameter values (Θ: $0.1 - 10$, σ_μ: $0.05 - 0.5$, α: $0.001 - 0.05$, σ_w: $0.1 - 0.5$ see Supporting Information: 2. Parameter exploration) that the upstream gene is more polymophic at the beginning of the adaptive walk while the downstream is more polymorphic when the population reaches the optimum. In the representative case of 2 genes and a conservative choice of $\Theta = 0.1$, in the beginning of the adaptive walk (time $t = 10^2$ generations) the number of alleles at the upstream gene is significantly higher than the number of alleles at the downstream gene (Wilcoxon signed rank test, $P = 1.2 \cdot 10^{-10}$), while at time $t = 10^4$ the opposite is true ($P = 2 \cdot 10^{-16}$). More generally, across the whole pathway the upstream genes are the most polymorphic at the beginning of the adaptive walk, but, as the population approaches the optimum, the trend is reversed and the downstream genes become the most polymorphic, both as measured by the number of alleles (Fig. 3, Supporting Information Fig. S7) and heterozygosity (Supporting Information Figs. S3 and S4).

The pattern of phenotypic variance ($\mathrm{Var}(R_i) = \mathrm{E}[(p_i R_i)^2] - (\mathrm{E}[p_i R_i])^2$) of the reaction parameters (R_{f_i}) also shows a parallel shift to polymorphism (Fig. 3B). At generation 10 the phenotypic variance is low in both genes although it is marginally higher in the upstream gene than the downstream gene (Wilcoxon signed rank tests P-value: 0.063). At generation 100 the phenotypic variance of both genes is high and is significantly higher in the upstream gene than in the downstream (P-value: 0.018). After 10^3 generation the pattern reverses and the downstream gene has
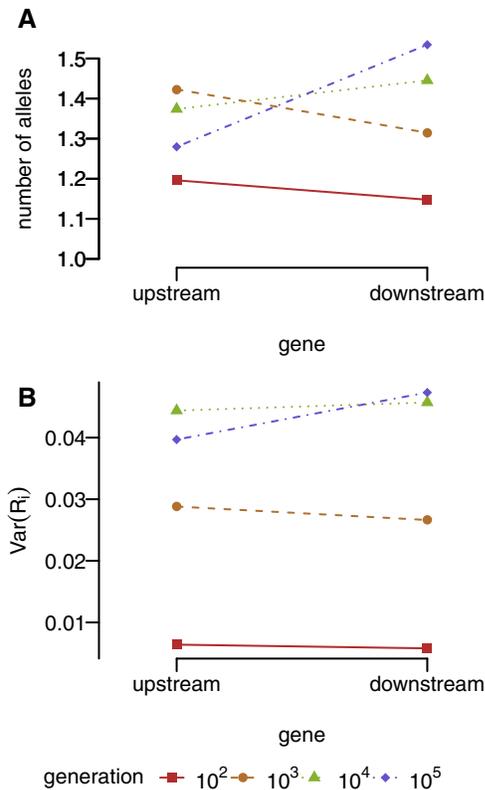
Figure 3. The degree of polymorphism in genes coding for enzymes across a linear metabolic pathway depends on which phase of adaptation is studied. (A) During the initial phase of adaptation (coarse grained) the upstream genes are more polymorphic, while during fine tuning the downstream genes become the most polymorphic ones. (B) The pattern of phenotypic variance also changes depending on the phase of adaptation. During directional selection the upstream genes are the most variable, i.e the variance is distributed along the trajectory of adaptation (cf. Fig. 1). During stabilizing selection the downstream enzymes are more variable.

a higher phenotypic variance (*P*-values: 0.013, $1.2 \cdot 18^{-26}$ at 10^3 and 10^4 generations).

PATTERNS OF DIVERGENCE

We then focus on the patterns of divergence during adaptation. We perform two series of simulations starting from a monomorphic population with random combinations of rates (R_{f_i}). For one set of simulations we include as starting states only combinations of rates that correspond to low fitness ($0.19 < W < 0.21$) and for the other we include combinations that correspond to high fitness values ($0.79 < W < 0.81$) such that the fitness of the starting genotype is either far or close to the optimum. As soon as the population reaches the optimum we calculate the total number of substitutions (fixed mutations) for each gene in the most fit genotype (Fig. 4A). Thus the mutations we study at the end of the

walk are mainly mutations accumulated during an adaptive walk and not neutral drift.

The distribution of fitness effects (ΔW) and of mutational sizes (ΔJ) decrease along the pathway (Fig. 4A, B). The upstream gene has significantly larger ΔJ than the downstream when starting both close and far from the optimum (Wilcoxon rank sum test, close: $P = 3.0 \cdot 10^{-13}$, far: $P < 2.2 \cdot 10^{-16}$). The same holds true for ΔW (close: $P < 2.2 \cdot 10^{-16}$, far: $P < 2.2 \cdot 10^{-16}$). Median step sizes in terms of flux are smaller in simulations starting closer to the optimum and in the downstream gene the spread (error bars) is larger. The smaller step sizes result from the fact that large step sizes close to the fitness peak would result in overshooting the optimum. The error bars are larger because the sample size of mutations is smaller: for the same number of simulations it takes fewer steps to adapt if the population starts closer to the optimum. Although the upstream gene contributes the most to fitness in all cases, its relative contribution to fitness is diminished in the simulations starting closer to the optimum. This is evident in the fact that the ratio of average contribution to fitness of upstream versus downstream genes is higher in populations starting far from the optimum than those starting close and is a reflection of the fine-tuning process implicit in the fitness landscape. We quantify the contribution of a gene to fitness ΔW by calculating the fitness changes of a genotype at each fixed mutation in the particular gene in the most fit genotype at the generation where the population reaches adaptation. The difference (and ratio) in the median ΔW across simulations between the upstream and downstream gene is 0.129 (3.026) for simulations starting far from the optimum and 0.07 (2.46) for simulations starting close to the optimum.

The total number of substitutions is larger in the simulations starting far from the optimum, as the genotypes require more steps to adapt (Fig. 4, Supporting Information Fig. S4). However, the ratio of average substitutions in the simulation starting far from the optimum to the average substitutions in simulations starting close to the optimum decreases along the genes of the pathway (far: 2.5, close: 1.3). Also, relatively more substitutions occur in the downstream genes as the populations approach the fitness peak (Fig. 4D, Supporting Information Fig. S6). The median fitness increase ΔW is 0.19 in the upstream gene and 0.06 in the downstream gene in simulations starting far from the optimum and for simulations starting close to the optimum, ΔW is 0.12 and 0.05 for mutations in the upstream and downstream genes correspondingly. Given that these substitutions increase fitness, these observations imply a selective process that favors fixation of genes depending on both their position in the pathway and distance from the fitness optimum.

In contrast to the selective pressures driving diversity in the short-term, continued divergence in populations in the long-term (assuming no change in the environment) is shaped by neutral
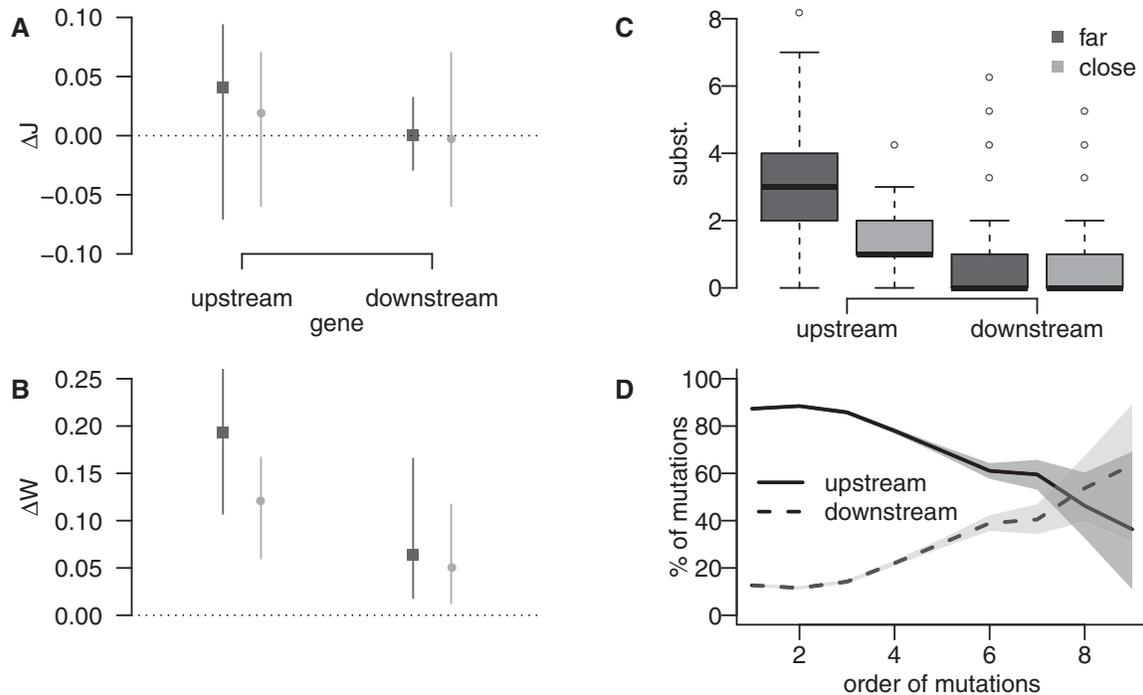
Figure 4. Distribution of mutational step sizes (ΔJ), fitness effects (ΔW), and total number of mutations in adapted genotypes calculated from 10,000 adaptive walks in a linear pathway of two genes. (A, B) Starting far from the optimum mutations in the upstream gene have significantly larger effect on flux and fitness, but this pattern is attenuated in simulations starting close to the optimum. (C) The difference in the number of substitutions in the upstream and downstream genes is larger in simulations starting far from the optimum. Points correspond to median values and error bars to first and third quartiles. (D) Simulations starting far from the optimum. The percentage of mutations fixed in the upstream and downstream genes depends on the order in which mutations fixed. Shaded regions correspond to 95% binomial sampling confidence interval.

forces. We study this scenario by moving our focus from individual genotypes that have just reached the optimum to populations that approach the optimum and then drift on top of the neutral plateau at the end of an adaptive walk. We perform 10^3 simulations starting far from the optimum ($0.19 < W < 0.21$) and running for $3 \cdot 10^5$ generations. During the simulations we calculate the average population divergence for each gene by weighting the divergence (number of fixed mutations) of each genotype by its frequency in the population excluding the very low frequencies ($x < 0.01$) (Fig. 5 and Supporting Information S8). We observe that the upstream gene diverges early during the adaptive walk. Once the population reaches the optimum, however, the rate of divergence slows down significantly until the increase in divergence is linear due to the rare neutral fixation of mutations. The downstream gene also shows an initial fast increase in divergence that also becomes linear after reaching the optimum. However, the rate of divergence (slope) for the downstream gene is larger as it is less constrained, which contributes to the inversion of the pattern of divergence across genes.

Discussion

In this study, we explore the emergence of variation patterns in metabolic pathways through the interplay of evolutionary and enzyme kinetic dynamics. We do not explicitly model variation but only the necessary requirements for variation to emerge. To this end, we link two models at different levels and timescales. At the level of the individual genotype, we implement an enzyme kinetic model for a linear metabolic pathway and calculate steady state concentrations of pathway substrates and products. At the population level, we implement a forward population genetic model of genotypes encoding enzymes of the pathway. The two levels are connected by the fitness of a genotype, which is determined by the rate of production of the last product of its corresponding metabolic pathway in relation to a fixed optimum. In starting our simulations with genotype parameters seeded by random values which place their phenotypes some distance from the optimum, we simulate a situation in which populations must adapt to a change in environment.

This modeling approach involves two important types of interactions, epistatic interactions between genes and interactions
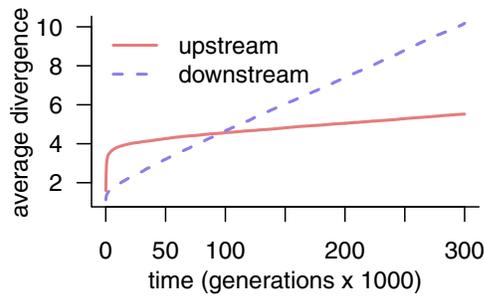
Figure 5. Average population divergence of upstream and downstream genes as a function of time, across 1000 simulations. During adaptation to a large environmental change mutations fix quickly in the upstream gene leading to a very high rate of divergence. As soon as the population adapts the upstream gene only fixes mutations linearly. Similarly the downstream gene initially diverges quickly, however not as fast as the upstream gene. After the population reaches the optimum, its neutral rate of divergence is higher and the pattern of divergence eventually changes with the downstream gene becoming more divergent than the upstream. We define the average population divergence for the upstream gene as $D_{upstream} = \sum^{k} d_i^{upstream} p_i$, where $d_i^{upstream}$ is the number of mutations fixed since the beginning of the simulation in the upstream gene of genotype i, p_i is the frequency of genotype i and k is the number of genotypes with frequency larger than 0.01. The average population divergence is defined analogously for the downstream gene.

between competing genotypes. Indeed, individual gene fitness within any particular genotype is meaningless given that the flux on which fitness is based is a system-level property dependent on the functional linkage/kinetics of all constituent enzymes. In this sense, all interactions between genes in our model can be considered epistatic. Interactions between genotypes arise as we do not make the common assumption that the most fit genotype will deterministically outcompete all others. Rather, we allow multiple genotypes in our simulated populations to evolve together. This process results in complex population dynamics such as clonal interference (Barrick and Lenski 2013) that contribute to the patterns of variation we find.

By implementing the above models in simulations of adaptive walks starting from monomorphic states with low fitness and by sampling genetic variation along evolutionary trajectories, we are able to qualitatively reproduce all the patterns of variation that have been observed in empirical studies. The differential amounts of variation observed stem from the asymmetrical action of selection related to the reversibility α of the pathway. In a directional pathway, with all else equal, the upstream genes are more important in the sense that equivalent changes in upstream genes have a larger effect on overall pathway flux and, therefore, fitness (Hartl et al. 1985). While the population is in a mostly directional selection regime, genes coding for upstream enzymes

have a higher probability of being polymorphic than genes coding for downstream enzymes. In this phase, adaptation is driven by multiple, independent partial selective sweeps in the upstream genes. However, once the population adapts, selection is mostly stabilizing and we observe the opposite pattern, with higher levels of polymorphism and rates of fixation in the downstream genes. Intermediate levels of polymorphism and divergence with no polarity in terms of pathway directionality may be found between these two selection extremes.

Tracking the dynamics of adaptation in terms of the fitness landscape, we observe that at the early stages of an adaptive walk multiple mutations in the upstream genes increase in frequency and an elongated cloud of genotypes initially spreads mostly along the R_{f_1} axis, contributing to high phenotypic variance at the upstream genes (Figs. 1, B). As the population approaches the optimum ridge, the shape, and size of the genotype cloud changes. It becomes larger (higher phenotypic variance) as the top of the fitness peak is flatter, and spreads more along the axes of downstream genes.

Relative differences in levels of polymorphism between upstream and downstream genes over the course of the simulation are a function of differences in rates of beneficial and neutral mutations. The propoportion of random mutations that can be beneficial during adaptation for all genes decreases as the population approaches a local optimum, though these rates differ according to pathway position. The initially higher polymorphism of the upstream genes when the population is far from the optimum is driven by multiple independent beneficial mutations, as it is easy to become slightly better in a maladaptive environment. Beneficial mutations in the downstream genes have the same probability of occurrence, but the magnitude of their fitness increase is less, due to their generally smaller contribution to flux, and hence they do not increase in frequency to the same degree as the upstream genes. As the population moves closer to the optimum, the upstream genes become more constrained (due to a higher rate of deleterious mutations), and thus less polymorphic. Downstream genes are less constrained at this "fine-tuning" stage and thus develop higher levels of polymorphism. Finally, on top of the fitness peak, the trend toward higher polymorphism in the downstream genes continues as a result of their greater freedom to drift neutrally than the now highly constrained upstream genes.

Similarly, during the adaptive walk mutations fix initially with higher probability in the upstream genes, which contributes to higher levels of divergent variation between lineages. Far from the optimum, adaptation proceeds substantially by substitutions in the upstream enzyme while, near the optimum, substitutions in the downstream gene increase in prevalence. Very close to the optimum, downstream genes continue to accumulate neutral and nearly neutral substitutions, whereas at the upstream gene purifying selection restricts the accumulation of substitutions. At the

end of an adaptive walk, the distribution of mutational sizes and fitness effects of substitutions in adapted genotypes decreases from upstream to downstream genes. These findings regarding substitutions are in agreement with Wright and Rausher (2010), suggesting that the predictions of both approaches may be robust to variation in details of the models. One apparent exception is that Wright and Rausher (2010) did not find that the distribution of mutational sizes decreases from upstream to downstream. However, they quantified mutational sizes in a different way. We measure the mutation size in terms of effect on flux ΔJ, whereas Wright and Rausher (2010) measured it in terms of effect on the kinetic parameter k. They found that it was approximately equal across genes. However, because the effect of a given change in k on flux decreases from upstream to downstream genes (because of the parallel decline in flux control), the predictions of the two models are similar.

Fisher (1930) used the manipulation of a microscope as a metaphor for adaptation. He argued that changes of large effect applied randomly, such as movement of the microscope lenses by any angle or distance, will most likely distort the image we see, while smaller changes would be safer to perform. By analogy, and in concert with his geometric theory of adaptation, he argued that the smallest possible mutations should be the ones favored by natural selection. Kimura (1983) refined expectations regarding the distribution of fitness effects by taking into account the fact that drift would remove mutations of small effect from an adapting population. Thus mutations of intermediate fitness effect would be expected to dominate the adaptive process. Orr (1998) further refined our understanding of adaptation by recognizing that the intermediate fitness value would be relative to the distance from the optimum and would diminish as adaptation proceeds. As a consequence, the distribution of fitness effects of beneficial mutations of an adapted genotype would be an exponential with the largest mutations fixing in the first adaptive steps and the smaller effect mutations fixing at the final steps. Wright and Rausher (2010) found a similar pattern evident in the number of substitutions found in upstream versus downstream genes of a metabolic pathway, and attributed this result to an interplay between dominant fitness effects of upstream genes and neutrality. We extend this result by incorporating short time scale dynamics and polymorphism. Doing so we find that adaptation proceeds in a fine-tuning mode. Returning to the microscope analogy, just as we first have to adjust the coarse focus knob and then use the fine focus knob to get a clear image, selection may be expected to first "adjust" upstream genes while "fine tuning" with downstream genes. Further, the specifics regarding the optimal "knobs" to manipulate (= position in a pathway to "adjust") will depend on how close we are to begin with to a "clear image." Patterns of both divergence and polymorphism follow from the evolutionary analog of this fine-tuning dynamic.

While there are clearly a great number of factors that may affect actual distributions of genetic variation in nature, our results may serve as a simple null model for expectations. Even holding all other complexities equal, expectations regarding levels of polymorphism and divergence in a functionally linked and hierarchical pathway must consider the distance of populations from local fitness optima. A population or species that has recently experienced a large change in environment can be expected to show higher levels of polymorphism and substitutions in upstream versus downstream genes of metabolic pathways under selection. Conversely, if we were to find elevated levels of polymorphism or substitutions in upstream genes that may indicate a population is currently adapting to a major change in selective pressure. Elevated levels of polymorphism or divergence in downstream genes may be an indication of constraint of upstream genes and neutral drift, as is commonly assumed, or may be an indication of a "fine-tuning" process of adaptation. Genome-wide scans might benefit from incorporating expectations of such an asymmetric mode of adaptation to make better sense of sequence data. Perhaps the best test of these expectations would be in analyzing genetic change found in experimental evolution experiments such as those of (Tenaillon et al. 2012; Herron and Doebeli 2013; Kvitek and Sherlock 2013; Lang et al. 2013). Indeed, in a recent review of evolutionary experiments, (Hindré et al. 2012) highlighted the importance of compensatory mutations in downstream genes after the first adaptive mutations in genes with a central location in the regulatory network. Our model allows for both this sort of "compensatory" fine-tuning after "over-shooting" as well during a gradual approach to the fitness peak.

There are several limitations to our approach that should be noted. First, while our findings have clear relevance to studies of clonally reproducing haploid populations such as bacteria and yeast (reviewed in Hindré et al. (2012)), it is unclear how relevant our findings would be to diploid, sexually reproducing populations. Future extensions of our model such as the incorporation of diploidy and recombination may shed light on this question. Similarly, while our results should help to explain patterns of variation in linear metabolic pathways, further work will be necessary to determine the effects of more complex topologies, such as how changes in parts of more comprehensive topologies (e.g., full regulatory or metabolic networks) may ripple forth to affect remote regions of the system. Another thing to keep in mind about our approach is that we are modeling steady state metabolism with the assumption that enzymes are far from saturation. In nature, however, other metabolic dynamics could be important determinants of fitness. These include rare occurrences with large consequences such as the need for rapid energy production to evade predators (Watt and Dean 2000).

The formalism of our model may be applied to a large number of evolving directional processes besides metabolic pathways.

The underlying principle is general and similar dynamics should be expected in hierarchical or directional processes where changes at a higher level or at the first step are more important and influential for the final product. We would thus expect that the evolution of diverse processes such as protein folding, ontogenetic development, ritual mating behavior, and cascades in gene regulatory networks or even nonbiological processes such as assembly lines and computer software development might follow similar dynamics and show parallels in their variation patterns.

LITERATURE CITED
Alvarez-Ponce, D. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. BMC Evol. Biol. 12:192.

Alvarez-Ponce, D., M. Aguadé, and J. Rozas. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. Genome Res. 19:234–242.

———. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. Genome Biol. Evol. 3:87–101.

Alvarez-Ponce, D., S. Guirao-Rico, D. J. Orengo, C. Segarra, J. Rozas, and M. Aguadé. 2012. Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in *Drosophila melanogaster*. Mol. Biol. Evol. 29:123–132.

Barrick, J. E., and R. E. Lenski. 2013. Genome dynamics during experimental evolution. Nat. Rev. Genet. 14:827–839.

Clotault, J., D. Peltier, V. Soufflet-Freslon, M. Briard, and E. Geoffriau. 2012. Differential selection on carotenoid biosynthesis genes as a function of gene position in the metabolic pathway: a study on the carrot and dicots. PLOS One 7:e38724.

Cork, J. M., and M. D. Purugganan. 2004. The evolution of molecular genetic pathways and networks. BioEssays 26:479–484.

Dall'Olio, G. M., H. Laayouni, P. Luisi, M. Sikora, L. Montanucci, and J. Bertranpetit. 2012. Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. BMC Evol. Biol. 12:98.

Desai, M. M., and D. S. Fisher. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176:1759–1798.

Fisher, R. 1930. The genetical theory of natural selection. The Clarendon Press, Oxford, U.K.

Flowers, J. M., E. Sezgin, S. Kumagai, D. D. Duvernell, L. M. Matzkin, P. S. Schmidt, and W. F. Eanes. 2007. Adaptive evolution of metabolic pathways in *Drosophila*. Mol. Biol. Evol. 24:1347–1354.

Fraser, H., D. Wall, and A. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol. Biol. 3:11.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi. 2009. GNU scientific library reference manual. 3rd ed. Network Theory Ltd. Available at http://www.network-theory.co.uk/gsl/manual/.

Greenberg, A. J., S. R. Stockwell, and A. G. Clark. 2008. Evolutionary constraint and adaptation in the metabolic network of *Drosophila*. Mol. Biol. Evol. 25:2537–2546.

Hahn, M. W., and A. D. Kern. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22:803–806.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J. Mol. Evol. 58:203–211.

Hartl, D. L., D. E. Dykhuizen, and A. M. Dean. 1985. Limits of adaptation: the evolution of selective neutrality. Genetics 111:655–674.

Herron, M. D., and M. Doebeli. 2013. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. PLoS Biol. 11:e1001490.

Hindré, T., C. Knibbe, G. Beslon, and D. Schneider. 2012. New insights into bacterial adaptation through in vivo and in silico experimental evolution. Nature reviews. Microbiology 10:352–365.

Hudson, C., and G. Conant. 2011. Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. BMC Evol. Biol. 11:89.

Jovelin, R., and P. C. Phillips. 2011. Expression level drives the pattern of selective constraints along the insulin/TOR signal transduction pathway in *Caenorhabditis*. Genome Biol. Evol. 3:715–722.

Kacser, H., and J. A. Burns. 1973. The control of flux. Symp. Soc. Exp. Biol. 27:65–104.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

Kvitek, D. J., and G. Sherlock. 2013. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. PLoS Genet. 9:e1003972.

Lang, G. I., D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock, D. Botstein, and M. M. Desai. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500:571–574.

Livingstone, K., and S. Anderson. 2009. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. J. Hered. 100:754–761.

Ma, X., Z. Wang, and X. Zhang. 2010. Evolution of dopamine-related systems: biosynthesis, degradation and receptors. J. Mol. Evol. 71:374–384.

Montanucci, L., H. Laayouni, G. M. Dall'Olio, and J. Bertranpetit. 2011. Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. Mol. Biol. Evol. 28:813–823.

Orr, H. A. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52:935–949.

Piatigorsky, J. 2007. Gene sharing and evolution: the diversity of protein functions. Harvard Univ. Press, Cambridge, MA.

Ramos-Onsins, S. E., E. Puerma, D. Balañá-Alcaide, D. Salguero, and M. Aguadé. 2008. Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana*. Mol. Ecol. 17:1211–1223.

Ramsay, H., L. H. Rieseberg, and K. Ritland. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol. Biol. Evol. 26:1045–1053.

Rausher, M. D. 2013. The evolution of genes in branched metabolic pathways. Evolution 67:34–48.

Rausher, M. D., R. E. Miller, and P. Tiffin. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol. Biol. Evol. 16:266–274.

Rausher, M. D., Y. Lu, and K. Meyer. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J. Mol. Evol. 67:137–144.

Riley, R. M., W. Jin, and G. Gibson. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*. Mol. Ecol. 12:1315–1323.

Savageau, M. A. 1976. Biochemical systems analysis: a study of function and design in molecular biology. Addison-Wesley Pub. Co., Advanced Book Program, Reading, MA.

Schuster, S., D. A. Fell, and T. Dandekar. 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. 18:326–332.

Sharkey, T. D., S. Yeh, A. E. Wiberley, T. G. Falbel, D. Gong, and D. E. Fernandez. 2005. Evolution of the isoprene biosynthetic pathway in kudzu. Plant Physiol. 137:700–712.

Tenaillon, O., A. Rodríguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S. Gaut. 2012. The molecular diversity of adaptive convergence. Science 335:457–461.

Vitkup, D., P. Kharchenko, and A. Wagner. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.

Watt, W. B. 1986. Power and efficiency as indexes of fitness in metabolic organization. Am. Nat. 127:629–653.

Watt, W. B., and A. M. Dean. 2000. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu. Rev. Genet. 34:593–622.

Wright, S. 1938. The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA 24:253–259.

Wright, K. M., and M. D. Rausher. 2010. The evolution of control and distribution of adaptive mutations in a metabolic pathway. Genetics 184:483–502.

Wu, X., X. Chi, P. Wang, D. Zheng, R. Ding, and Y. Li. 2010. The evolutionary rate variation among genes of HOG-signaling pathway in yeast genomes. Biol. Direct 5:46.

Yang, Y.-H., F.-M. Zhang, and S. Ge. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. BMC Evol. Biol. 9:206.

Yu, H.-S., Y.-H. Shen, G. X. Yuan, Y. G. Hu, H. E. Xu, Z. H. Xiang, and Z. Zhang. 2011. Evidence of selection at melanin synthesis pathway loci during silkworm domestication. Mol. Biol. Evol. 28:1785–1799.

Associate Editor: D. Roze
Handling Editor: T. Lenormand

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Supplementary information for: Patterns of variation during adaptation in functionally linked loci.

Table S1. Class allele.
Table S2. Class genotype.
Table S3. Most important input for method genotype: mutate.
Table S4. Most important input for method genotype: fW.
Table S5. Class population.
Table S6. Most important input for method population: mutate.
Table S7. Class environment.
Table S8. Most important input for method environment: pulse.
Table S9. Class incubator.
Table S10. Most important input for method incubator: run.
Table S11. The main parameters of the simulation.
Table S12. Columns in tsX.dat files.
Table S13. Columns in tableX.dat files.
Table S14. Columns in afX.dat files.
Table S15. Columns in sfsX.dat files.
Table S16. Patterns of variation for non-mutation limited adaptation (large values of Θ).
Table S17. Patterns of variation for mutation limited adaptation (small values of Θ).
Table S18. Patterns of variation for mutations of large effect size (large values of σ_μ).
Table S19. Patterns of variation for mutations of small effect size(small values of σ_μ).
Table S20. Patterns of variation in very directional pathways (small values of α).
Table S21. Patterns of variation in more reversible pathways (large values of α).
Table S22. Patterns of variation for small values of σ_ω.
Table S23. Patterns of variation for large values of σ_ω.
Figure S1. Class inheritance diagram.
Figure S2. Effect of reversibility α and standard deviation of fitness function σ_ω on the shape of the fitness landscape.
Figure S3. Changing pattern of polymorphism during an adaptive walk.
Figure S4. Changing pattern of polymorphism during an adaptive walk in a metabolic pathway of 4 genes.
Figure S5. The rate of beneficial μ_b and deleterious μ_d new mutations depends on the genotype (R_{f_1}, R_{b_1}) and distribution of mutation rates.
Figure S6. As in Figure 4 but for a linear pathway comprised of 4 genes.
Figure S7. The degree of polymoprhism in genes.
Figure S8. Change in the rate of divergence for each gene across metabolic pathway of four genes, as in Figure 5.