Evolution of complex phenotypes through successions of adaptive steps

Tin Yau Pang, Martin Lercher

Bioinformatik, Heinrich-Heine-Universität Düsseldorf

Abstract

The emergence of complex phenotype is a fascinating question of evolutionary biology, and we sought to understand preadaptation which facilitated the development of complex phenotypes, in the context of bacterial metabolic network. Genes coordinated for a phenotype are likely to cluster on the same place of the genome, which so allows horizontal gene transfer (HGT) to pass the phenotype to another bacterium. But for a complex phenotype, its genes are clustered on different places of the genome cannot be transferred adaptively; it is preadaptation, which refers to adaptive transfer of a segment relevant to a complex phenotype for other purposes, that allows it later to be recruited for the complex phenotype. To search for preadaptation in the evolutionary history of E. coli, we reconstructed the ancestral genomes from various strains, identified the transferred genes, grouped them into possible transferred segments, and analyzed the gains in nutritional phenotypes corresponding to the acquisitions of segments of metabolic genes. Properties of these HGT segments inferred from data are enumerated and compared with a model of HGT, which shows that: 1) HGT segments are likely to adaptive, and segments carrying reactions essential to phenotypic gains but non-adaptive are rare; 2) the landscape of segment transfer for complex phenotypes is directional and path-dependent; 3) cooperation between HGT segments to support various nutritional phenotypes are observed to
be more frequent than expected, which serves as an evidence to preadaptation in the evolution of bacterial metabolic network.

**Significance**

Evolution of a complex trait is an intriguing topic in evolution, as its sheer complexity seems incompatible with theory of adaptation and selection. An explanation to complex traits is preadaptation, where the seemingly nonadaptive intermediate traits of a complex trait are first selected for other purposes, and later recruited for the complex trait. In this work, we modelled the evolution of complex nutritional phenotypes in bacterial metabolic network; we defined segment transfer as the basic step of evolution, phenotypes requiring more than a single transfer are selected against and hence is complex. Preadaptation in this context is detected when genes acquired from different transfer events are cooperating for the same phenotype, which in this analysis is shown to be common.

**Introduction**

Horizontal gene transfer (HGT) allows the sharing of genes between bacteria and drives evolution (15). When a bacterium receives a foreign DNA segment, it will have a chance to use the genes on it to solve new challenges arose due to changes in its native environment; if it works, then the segment will further spread out and get fixed in the population. Acquisition of foreign genome segments brings not only the immediate gains such as new phenotypes, but it also broadens the space of phenotype that is reachable in the next HGT event. Such expansion of reachable space is important to evolution, as it allows bacteria to preadapt to complex phenotype, i.e., phenotype whose genes are clustered in different locations on the genome, it cannot be acquired adaptively but requires multiple transfer events. Preadaptation is responsible for the development of a complex phenotype, by which the bacterium acquires segments relevant
to the complex phenotype adaptively but for other purposes, until finally the originally complex phenotype is only one transfer away and can be acquired adaptively. In this study, we analyzed the adaptation of E. coli to various nutritional phenotypes; in particular, we sought to understand cooperation between transferred segments, because segment acquisition is the basic step of adaptation, and phenotypes that depends on multiple segments are likely to be complex in the first place and its development is likely to be a consequence of preadaptation.

We reconstructed the ancestral genomes and metabolic networks (14) to study preadaptation. The genes acquired by the ancestral genomes are identified and grouped into genomic segments that had possibly been transferred, based on the adjacency of genes found on the extant genomes. These genomic segments are then considered as the basic unit of HGT, as genes coordinated for a phenotype are likely to be driven to the same genomic cluster by the force of evolution (18). The phenotypes of a metabolic network is defined based on its capability to metabolize different combinations of nutritions in using flux balance analysis (FBA) (4), and the changes of nutritional phenotypes when one ancestral strain evolved into another is examined. For each new phenotype acquired or existing phenotype with biomass flux enhanced, parsimony FBA (pFBA, also called minimum total flux constraint) (6) is applied to determine the flux distribution, and identified the corresponding HGT segments to explain the reactions that are not inherited directly from the ancestor. In this way, the transferred segments and their phenotypes are defined, which allows us to study adaptation and preadaptation in details.

**Results**

We reconstructed the phylogenetic tree of 53 E. coli strains that have published metabolic networks (14), along with 17 strains of other species that served as outgroup in the tree (see Table S1 for details of strains). The orthologous genes of these 53+17 genome sequences are mapped, 1,334 universal genes that have no paralogs are aligned, and their phylogenetic tree is
constructed from the alignment (see Materials and Methods for details, and Fig S1 for the phylogenetic tree of these 70 strains). The resultant phylogenetic tree with 70 strains is rerooted into the topology that separates the 17 outgroup strains from the E. coli strains, and the entire subtree of outgroup is removed; analysis on 200 bootstrap maximum-likelihood trees shows that every internal node of the tree are supported by a bootstrap value of at least 60% (see Fig 1 for E. coli tree with bootstrap values). Each of the remaining 52 internal nodes of the 53 E. coli tree is then considered an ancestral strains, and their genomes are constructed with a maximum parsimony algorithm based on the genomes of the 53 extant strains that minimizes the number of gene gain and loss events across the phylogeny (see Materials and Methods for details of ancestral genome reconstruction). Flux balance analysis (FBA) performed on these metabolic networks over 654 nutritional environments provided in Ref (14) shows that 30 extant and 44 ancestral metabolic networks are functional, i.e., having a biomass flux greater than 1E-6, in some environments. Out of the 23 non-functional extant strains, 10 of them are expected to be non-functional in the 654 phenotypic conditions as they require the addition of extra nutrients (14), and the remaining 13 is a result of inconsistent ortholog-reaction mapping across different metabolic models. Every pair of neighbouring ancestral or extant strains on this phylogenetic tree is considered as an evolutionary step, and hereafter if FBA is involved, then only evolutionary steps with both strains functional are considered. The genes gained and lost in each step are identified and later grouped into possible HGT segments, and FBA is performed to examine the phenotypic gains brought by the segments.

**HGT segment structure derived from repeatedly co-transferred genes**

We analyzed the structure of repeatedly co-transferred genes. To start with, the gain and lost history of each gene in the 52 evolutionary steps is represented by a binary vector of 104 elements, one element for gain and one for lost event in each step; if a gene is gained (lost) in
Figure 1: Phylogenetic tree of the 53 E. coli strains; the number at each clade shows its bootstrap value, and so do the colour of the branches.
an evolutionary step, then the corresponding vector element 1, and otherwise 0. The score of association and avoidance of each gene pair is calculated by 1) align their vectors as a 104x2 matrix, 2) enumerate the occurrence of the four patterns (0,0), (0,1), (1,0) and (1,1) among the 104 rows, 3) calculate the p-value of the right and left tail of Fishers exact test and 4) the score of association and avoidance of a gene pair are defined as the logarithm of the p-value of the right and left tail respectively, where the more negative is the score, the stronger is the association or avoidance. We did not directly interpret this p-value as the significance of association / avoidance, which is because in Fishers exact test, it assumes that each event is independent of any other, but in our case the gain or loss of a gene in one step affects the state of that gene in the following step. Instead we compared it with a null mode for co-transferred gene pairs, which is made first by permuting the presence-absence configuration of each gene in the 53 extant strains and then constructing the ancestral genomes based on these permuted extant genomes; the distribution of score of association and avoidance for this null model is also calculated (see Fig S2 for distribution of scores of association and avoidance for data and null model). The shape of the distribution of association scores between data and null model are different from each other, as there are more gene pairs significantly associated in real transfer events than that in the null model; however, the distribution of scores of avoidance between data and null model are similar with each other, which means that genes in general do not tend to avoid each other.

Gene pairs with high statistical support are likely to be real co-transferring pairs, and we derived their distance distribution to help understand the structure of HGT segments. Let us denote \( N_d(s) \) and \( N_n(s) \) to be the number of gene pairs with score below \( s \) in the data and null model; we defined the significance level of association given a cutoff score \( s \) as \( N_n(s)/N_d(s) \), which is the ratio between number of pairs in the null model with score below \( s \) over the number of pairs with the same cutoff in the data. From Fig S2, one derives that the cutoff for significance level 0.05 (0.005) is -5.0 (-6.8). To estimate the genomic separation of each of these significantly
associated gene pairs at the time they were added to their host ancestral genomes, we searched through the 53 extant genomes and obtained the separations (distance between the start position of the two genes) of every gene pair in each of the genome sequences where both genes are present, and considered the minimum of those values to be the estimation of the genomic distance between the gene pair in the ancestral genome. Fig 2 shows the genomic separation of gene pairs collected at significance level 0.05 (solid curve) and 0.005 (broken curve). Both curves show a kink at 30kb, and hence in the following analysis we assume that a typical HGT segment has a maximum capacity of 30kb, where genes with separation greater than 30kb on a genome sequence can only be transferred in separate HGT events.

**Phenotypic property of an HGT segment**

Transfer of metabolic genes allows bacteria to share reactions that may lead to phenotypic gains, which can be acquisition of a new nutritional phenotype or enhancement of the biomass flux of an existing phenotype. To evaluate the gains in an HGT event, we grouped the transferred metabolic genes in different evolutionary steps into 30kb segments, and enumerated the phe-
notypic gains brought by each HGT segments; they are then compared with a model of HGT. Given a subset $S$ of genes transferred in an evolutionary step, a possible configuration with minimal number of segments is decided in the following way: 1) identify the starting positions of the genes in all the extant genomes they are present; 2) pick a random gene $g_A \in S$ and put it in the picked gene set $P$; 3) for each not-picked gene (represented by the set $S \setminus P$), enumerate the possible 30kb segments in the 53 extant strains that can cover the starting position of itself and all picked genes in $P$; 4) pick the one not-picked gene with the highest segment count and move add it to $P$; 5) repeat step 3 - 4, test each remaining not-picked gene in $S \setminus P$ by enumerating the segments in the extant genomes that can accommodate itself and the picked genes, and pick the one with the highest segment count that supports its grouping with the previously picked genes; 6) when no more genes can be picked, the picked genes in $P$ are then grouped into an HGT segment and removed from $S$, $P$ is emptied (set to $\emptyset$), step 2 - 5 is repeated to group another segment from the remaining genes, until every gene is assigned to a segment.

After grouping the transferred genes in an evolutionary step into minimal number of 30kb segments, FBA is applied to calculate the biomass flux in the 654 nutritional conditions (J4) on the host metabolic network with one of these segments added. Consequently, by comparing the biomass flux of the metabolic network with a segment added and metabolic network without any segment added, one can deduce the acquired and enhanced nutritional phenotypes that come with each segment. Specifically, a nutritional phenotype is acquired if the biomass flux at its nutritional environment is below 1E-6 before segment addition but gets beyond 1E-6 after the addition; a nutritional phenotype is enhanced if the biomass flux is greater than 1E-6 originally and increased by more than 1% after segment addition. The benefits introduced by a segment to its host metabolic network is then quantified by the number of new metabolic genes, number of new reactions, number of acquired new phenotypes and number of enhanced existing phenotypes.
For comparison, we simulated an HGT model with real genomic segments and one with randomized segments. For every donor-recipient-pair chosen from the functional extant strains, let subset $S$ include the complementary metabolic genes that appear only in the donor but not in the recipient; these genes are grouped into 30kb segments in the following procedure: 1) pick a random gene from $S$ and include it in $P$; 2) consider each not-picked complementary metabolic gene in subset $S \setminus P$ whether it has the potential to be in the same segment with the picked genes, i.e., its starting position and that of the other picked genes can be covered by a 30kb segment on the donor genome; 3) pick one of the potential genes randomly and include it in $P$; 4) repeat step 2 - 3 to pick another gene from unpicked genes in $S \setminus P$; 5) when no more genes can be picked, assign the picked genes in $P$ to a segment, picked genes in $P$ are removed from $S$, and then $P$ is emptied (set to $\emptyset$); 6) repeat step 1 - 5 on the remaining genes in $S$, until every gene is assigned to a segment. Going through this procedure will group the complementary metabolic genes in the donor into real genomic segments; and random permutation of the genes in these real genomic segments creates randomized segments who has the same gene number distribution but with random gene content.

This HGT model with randomized segments can serve as a null model to understand segment properties, i.e., how the number of reactions, acquired phenotypes and enhanced phenotypes are different from each other between real genomic segments and randomized segments; nonetheless it cannot serve as a null model to the number distribution of genes of a segment, and hence we considered the following scenario to derive a number distribution of genes in a segment if all genes are randomly distributed in the genome. For a donor-recipient-pair, if the complementary metabolic genes are randomly distributed in the genome, then their number distribution on a segment will follow Poisson distribution with $\lambda$ equal to the density of complementary metabolic genes. To estimate this density, we counted the number of complementary metabolic genes in the donor of a donor-recipient-pair, and divided it by the number
of segments in the donor (genome length divided by 30kb) to get the density for the pair; averaging this density over all possible functional donor-recipient-pairs results in a mean density $\lambda = 0.3215$ complementary metabolic genes per segment.

Panel A of Fig 3 shows the cumulative number distribution of genes in a segment for data (blue), HGT model (red) and Poisson distribution with $\lambda = 0.3215$ and renormalized after zero point removed (black). The distribution curves of the data and HGT model are visually similar, and Wilcoxon’s rank sum test also cannot reject the hypothesis that they have the same mean (two-sided, p-value = 0.2321); it also shows that real segments transfer more genes than expected from Poisson distribution. Panel B, C and D show the cumulative number distribution of new reactions, acquired new phenotypes and enhanced existing phenotypes respectively, for the data (blue), HGT model with real segments (red) and HGT model with randomized segments (green). In these three cases the distributions of the data are statistically similar with the distributions of the HGT model with real segments and Wilcoxon’s rank sum test cannot reject the hypothesis that they have the same mean values (two-sided, p-values = 0.8433, 0.6383 and 0.7153 respectively). Furthermore, the number distributions of new reactions and new phenotypes of the data are different from that of the HGT model with randomized segments (supported by two-sided Wilcoxon’s rank sum test at p-value = 0.0227 and 3.458E-11), but not for the number distribution of enhanced existing phenotypes as Wilcoxon’s rank sum test cannot reject the hypothesis that they are similar (two-sided, p-value = 0.0947). As a reference, Fig S3 shows probability density distributions of the cumulative distributions shown in Fig 3. By comparing the number distribution of gene usage in acquiring new phenotypes and enhancing existing phenotypes in the data, one finds that transferring new phenotypes (blue) uses more genes than enhancing existing phenotypes (turquoise) (Fig S4); the majority of phenotype-enhancements involve only one gene, and therefore the larger number of reactions in real segment transfers does not boost their ability in enhancing phenotypes.
Figure 3: Cumulative distribution of number of genes brought by a segment addition (Panel A) observed in data (blue), HGT model (red) and renormalized Poisson distribution with $\lambda=0.3215$ and zero point removed (black). Cumulative distribution of number of new reactions (Panel B), transferred new phenotypes (Panel C) and enhanced existing phenotypes (Panel D) in data (blue), HGT model with real segments (red) and HGT model with randomized segments (green). See Supp. Materials for their probability distributions.
The segments of metabolic genes defined in the data, HGT model with real / randomized segments are compared with real operons. From DOOR: Database of prOkaryotic OpeRons (11), we downloaded the operons of 45 strains that are part of our 53 extant strains and present in the database, and mapped the genes in each operon to our orthologs. Then for each segment that has been considered, we searched for the operon that has the highest number genes common with the segment. Heatmap of number distribution of common genes sharing with best matched operon at different segment size is shown in Fig. S5. It shows that for each segment in data and HGT model with real segments, it is likely to find a corresponding operon that has most of its gene orthologs, but this is not the case HGT model with randomized segments. This shows that the segments identified using our method are likely to be true operons.

Potential for a beneficial segment to be adaptively acquired

The success of incorporating a transferred segment into host genome depends on the likelihood for the metabolic pathway of the segment to connect to the host metabolic network and benefit the new host, and here let us consider how can a segment be beneficial and adaptively acquired. Let us define a segment to be either beneficial and non-beneficial, and a beneficial segment can either be adaptable and non-adaptable. A segment is adaptable (and beneficial) to a recipient strain if the addition of the segment immediately allows the recipient to acquire new phenotypes or enhance existing phenotypes, which is favoured by natural selection in the right condition. A segment is non-adaptable (but beneficial) if the addition of the segment does not lead to acquisition or enhancement of any phenotype, but provides some of the reactions for possible phenotypic gains and could be added with preadaptation. Finally, a segment is non-beneficial if it is not beneficial, which can happen if a segment provides reactions that already exist, reactions irrelevant to any of the 654 phenotypes considered or dead-end reactions.

We identified transferrable segments in functional extant strains, and tested each of these
segments by adding it to other functional extant strains. For each one of the functional extant strains (denoted as $s_a$), all of its non-universal metabolic genes are identified and grouped into 30kb segments in the same way as segments on the complementary metabolic genes on the donor of a donor-recipient-pair in the HGT model are grouped. Each of the segments in $s_a$ are then transferred to other extant strains that have any of the genes in the segment missing. For a recipient strain $s_b$, the new reactions brought by the addition of the segment are enumerated; FBA is performed over 654 nutritional conditions (14) on the metabolic network of $s_b$ with the segment added to enumerate the phenotypic gains (acquisition of new phenotypes / enhancement of existing phenotypes). If the addition of the segment results in any phenotypic gains, then the segment is beneficial and adaptable. Otherwise we tested whether it is beneficial but nonadaptable, or non-beneficial by: 1) take the union metabolic network of $s_a$ and $s_b$; 2) perform FBA over 654 nutritional environments on this union metabolic network, and by comparing the biomass flux of $s_b$'s network with the union metabolic network, identify all phenotypes acquirable or enhanceable to $s_b$ if it were to receive any segments from $s_a$; 3) the reaction pathways of these acquirable or enhanceable phenotypes are determined using pFBA (6), with those reactions whose flux magnitude smaller than 1E-6 ignored; 4) if the segment can provide a new reaction that is also part of the reactions of those acquirable or enhanceable phenotypes, then it is nonadaptable, and otherwise non-beneficial. Fig 4 shows the probability distribution for a segment to transfer new reactions (black), transfer new phenotypes (cyan), enhance existing phenotypes (green), be adaptable (pink) and be beneficial (yellow) to a recipient. There 76% of the segments will transfer new reactions to some of the recipients, and 37% of them will always transfer new reactions to all recipients; 42% will be adaptable to some of the recipients, and 13% will always be adaptable; 46% will be beneficial to some of the recipients, 16% will always be beneficial. Fig 5 shows the probability for a segment to be adaptable if it is beneficial (adaptable + non-adaptable) to a strain. Here 8% of the segments are always non-adaptable.
to strains that they are beneficial to, they can never be transferred adaptively and so they cannot be the first step in the transfer of a complex phenotype; 81% of the segments are always adaptable to strains that they are beneficial to, and so can be adaptively transferred at the right environment.

**Path-dependency in the multi-segment transfer of complex phenotypes**

Complex phenotypes require the transfer of multiple segments, some segments can be acquired adaptively and provide phenotypic gains in a host while others cannot; these segments thus can serve as the first step in the acquisition of complex phenotype. To examine this dependency of ordering on multi-segment transfer, we considered the simplest case of path-dependence to understand the roughness of the landscape of segment transfer. For every donor-recipient-pair in the HGT model with real segments, new phenotypes and enhanced existing phenotypes requiring exactly two segments are identified. All possible combination of segment pairs, using the real segments defined in the HGT model, that can give rise to the phenotype, are considered, and each of these segment pairs is classified into one of the three types: 1) both cannot be...
Figure 5: Distribution of probabilities for a segment beneficial to a recipient also be adaptable to it at the same time.
transferred adaptively to the host, 2) either one of them can be transferred adaptively and 3) both of them can be transferred adaptively to the host. Simulation shows that a complex phenotype have chances of 3.1%, 43.1%, 53.8% falling into the three categories respectively. This shows that many paths of multi-segment transfer are not adaptive and the landscape is rough.

**Interaction of transferred segments**

Interaction of transferred segments is a consequence of preadaptation, as the tinkering nature of evolution let the bacterium to recruit existing pathways for new phenotype. We looked for segment interaction in the data by enumerating the number of corresponding segments for the each acquired new phenotype or enhanced existing phenotype in every evolutionary step. Here, pFBA is applied to calculate the flux distribution of each of those acquired or enhanced phenotypes in each evolutionary step, and reactions in the flux distribution of a phenotype of interest that have flux magnitude greater than 1E-6 and not present at the progenitor strain of the evolutionary step are considered new reactions introduced by gene transfer, and we searched over all 30kb segments in the genomes of extant strains for a minimal number of segments that can accommodate these new reactions.

For comparison, we also considered every donor-recipient-pair formed by the functional extant strains. We took the union of the donor and recipient metabolic network and performed FBA over the 654 phenotypic environments, and by comparing the biomass flux of the union metabolic network and recipient network, the phenotypes that the recipient can acquire or enhance by receiving segments from the donor are identified. The flux distribution of the these (acquirable and enhanceable) phenotypes are found by pFBA upon the union metabolic network, and reactions with flux magnitude greater than 1E-6 and not present in the recipient are considered as necessary new reactions. The minimal number of 30kb segments of donors genome that can be transferred from the donor to the recipient to carry the required new re-
actions is then enumerated. Fig 6 shows the number distribution of segments, where the blue (cyan) bars is for the acquired new (enhanced existing) phenotypes in the data, and red (yellow) is for the acquired new (enhanced existing) phenotypes in the HGT model. It is observed that almost all phenotype acquisitions / enhancements in the data can be explained by a single HGT segment, while in the HGT model 20% of the phenotypes requires more than a single transfer. This analysis is repeated using a much longer segment length (500kb, see Fig S6), and the distinction between data and HGT model, i.e., suppression of multiple segment phenotypes in data but not in model, remains. Moreover, segments drawn only from descendent extant strains instead of all extant strains are also considered, and it is still observed that phenotypic gains that can be explained by a single segment appears more than what is found in the model (see Fig S7).

Analysis of segment interaction within one evolutionary step in the data shows that most
phenotypic gains can be explained by a single segment and use fewer segments than expected from model. However, it does not rule out the possibility that a segment acquired in the recent evolutionary step interacts with another segment acquired more than one evolutionary step ago. Therefore, we divided all the metabolic reactions of a strain according to the time they are added, i.e., some reactions are gained in the most recent evolutionary step, some in the second most recent step, and so on, and the remaining reactions that cannot be assigned to any steps are directly inherited from the most recent common ancestor (MRCA). For an acquired new phenotype or enhanced existing phenotype of each evolutionary step, we enumerated the minimal number of possible 30kb segments found in the extant strains to explain its reactions that are not inherited from MRCA, provided that reactions gained in different steps cannot be assigned to the same segment. For comparison, we also enumerated the possible 30kb segments if we were to transfer those reactions directly to the MRCA. Fig 7 shows the number distribution of segments added after MRCA to explain an acquired or enhanced phenotypes (pink) in different evolutionary steps, the number of evolutionary steps involved in segments transfer (green) and the number distribution of possible alternative segments to explain the acquired or enhanced phenotype if segments were to be added directly to the MRCA (cyan). The figure shows that 38% of the phenotypic gains are supported by multiple transferred segments, while only 19% of the phenotypic gains require multiple segments if we were free to choose segments to explain the reactions not inherited from MRCA.

**Discussion**

**Characteristic 30kb HGT segment length**

The presence of a kink in the distance distribution between co-transferred gene pairs in Fig 2 serves as a cutoff for grouping genes into segments. This length scale cannot be explained by transformation, let alone by the fact that transformation is not known to be common in E.
Figure 7: Number distribution of segments added after MRCA that correspond to a phenotypic gain (acquired or enhanced phenotype, pink), number of evolutionary steps involved in its segment transfer (green) and number distribution of possible alternative segments if we were free to choose segments to explain those reactions not inherited from MRCA (cyan).

coli (7). The driving mechanism of HGT in E. coli is transduction and conjugation, which is supported by previous studies (3, 19), and also by the fact that 30kb is on the same order of magnitude of the capacity of bacteriophages or plasmids (2). The length distribution of recombined segments on a genome can be compared with our distance distribution of co-transferred gene-pairs, as recombination is another consequence of DNA sequence acquisition, apart from gene transfer. Ref (5) gives a length distribution of recombined segments in Streptococcus pneumoniae; this distribution has a maximum length 235kb and a median 14kb, and the 30kb cutoff obtained in this study sits right at the third quartile (75%), thus our observation is consistent with other previous studies.
Genomic segments carrying reactions to transfer new phenotypes or enhance existing phenotypes are likely to be adaptively acquired

Coordinated genes have been shown to be driven into cluster on genome for better coordination which later facilitates their transfer (18), and our work investigates the benefits brought by acquisition of those clusters. Here the segments defined based on gene clusters on extant genome sequences are verified to be indeed functional units, as they can broadly match real operons provided in DOOR (11); simulation also showed that transfer of these segments brings more new reactions and new phenotypes than a null model of randomized segments. Nonetheless, real segments are also observed not doing better in enhancing the biomass flux of existing phenotypes when compared with randomized segments, which is because most enhancements require only one additional reaction, and real segments can not do better than random segments on transferring single reaction.

Our analysis showed that many beneficial segments can be transferred adaptively. Simulation on the HGT model with real segments reveals that while 54% of the segments are not beneficial to any recipients, i.e., they do not contain any reactions that are part of new or enhanced phenotypes of the recipient, the remaining 46% are beneficial to some of their recipients (Fig 4). Beneficence does not infer adaptability, as a segment may not contain all the reactions necessary for a new or enhanced phenotype; nonetheless if we ignore the non-beneficial segments to a recipient and consider only the beneficial ones, then there is 81% chance for a beneficial segment to also be adaptable, i.e., immediately provide new phenotypes or enhance existing phenotypes.

Path-dependence nature of segment transfer for complex phenotypes

Despite the high chance for a beneficial segment to also be adaptable, the landscape of segment transfer is rough and highly path-dependent when considering the complex phenotypes.
Analysis on complex phenotypes requiring two segments reveals that, around half of the time any of the segments can be acquired adaptively by the recipient, while in the remaining cases one of the two segments cannot be acquired adaptively, and this shows path-dependency for the segment transfer that lead to complex phenotypes. It is well known that E. coli is a generalist and is preadapted for many phenotypes, and hence it can acquire many segments adaptively. In contrast, a specialist bacterium like Buchnera aphidicola requires more than 70 reactions to transfer most phenotypes, and so only a few segments can be transferred to it adaptively; its landscape of multi-segment transfer will have stronger path-dependency, as few segments can serve as the first step in the multi-segment transfer.

**Prevalence of preadaptation in the evolution of E. coli metabolic network**

When a bacterium acquires a segment and develops a new phenotype, it also recruits existing genes to support the new phenotype; hence preadaptation in metabolic network can be detected by examining the interaction between segments, i.e., occasions where a phenotype which involves metabolic reactions brought by different transferred segments. Analysis shows that segment interaction is fewer than that expected from the HGT model in short evolutionary time scale (within one evolutionary step), but is more than expected from directly adding possible segments to MRCA in a longer evolutionary time scale (span multiple steps). This is consistent with the use-it-or-lose-it principle of evolutionary, because gene loss is far more likely than gene gain (13), and hence a bacterium can either acquire one segment adaptively and benefit from the phenotypic gains, or just lose the segment without acquisition, but cannot add multiple segments at the same time since it involves non-adaptive transfer. As an example, the lineage of UM146 has acquired the uridine transporter and the uracil transporter in different HGT events to develop anaerobic metabolism that uses UDP-N-acetyl-D-glucosamine as a carbon source (Fig 8).
Figure 8: The added transporters (red arrows) for uridine and uracil in the pathway of anaerobic metabolism of UDP-N-acetyl-D-glucosamine (red cell) in UM146.

Relation to other works

Recruiting old pathways for new innovations is common to many complex systems (17), but how is our result where roughly 40% of new phenotypes supported by multiple segment when compared with other models? Ref (12) and (16) use a toolbox model to relate the growth of metabolic pathway with the growth of metabolic phenotypes across many species, it can also calculate the probability for a newly added pathway (segment in our case) to interact with a previously added pathway. This model assumes a newly acquired pathway connects with only one existing gene in the recipient network, and so the chance of segment interaction is equal to the ratio of gene added after the MRCA. Given that there exists 743 universal metabolic genes and the average number of non-universal metabolic genes in one strain is 653, the probability to observe a new segment to depend on a previously added segment is \( \frac{653}{653 + 743} = 47\% \), which means that segments are observed to be somewhat less likely to connect to previously added segments that expected from random connection.

Recent study has also revisited the importance of preadaptation as a non-adaptive strategy of evolution. In Ref (1), the authors explored a space of possible metabolic networks, and pointed out that the development of one nutritional phenotype also lead to the development related nutritional phenotypes, and this non-adaptive evolution serves as the basis for the de-
velopment of new metabolic phenotypes. In our work, we mined for past transfer events in the metabolic network through phylogenetic analysis, and modelled a phenotypic space by transferring real genome segments to real metabolic networks; by comparing data with model, we have demonstrated the prevalence of preadaptation, i.e., HGT segment added for one phenotype is later recruited for another phenotype, throughout the evolutionary history of E. coli metabolic network.

Acknowledgements

This work is supported by SFB 680.

Materials and Methods

Reconstruction of phylogenetic tree, ancestral genomes and metabolic networks of E. coli

Give the genbank files of 53 E. coli (plus Shigella) strains and 17 sequence of similar species, the amino acid sequences of the putative genes are extracted and the gene orthologues are mapped using Proteinortho (10) with the option on synteny applied, which identified 13,138 gene families with more than one gene. The amino acid sequence of the 1,334 genes universal to 70 genome sequences are aligned using MAFFT (8) using the default option, and their phylogeny is inferred from the alignment using RAxML (20) with 200 fast bootstrap along with the PROT-CATAUTO model. This generated a phylogenetic tree that has at least 60% bootstrap support at each internal node. The 70 strains phylogenetic tree is then rerooted to group all 53 E. coli strains into the same subtree and each of the 52 internal nodes of the tree is considered as an ancestral strain. A maximum parsimony algorithm (9) is applied to each orthologous gene to determine its presence or absence in the ancestral genome from its states in the 53 extant genomes. Starting from the tip nodes towards the root, the presence or absence of a gene in a
node is equal to those of its descendants if they have uniform configurations; if the number of gains is greater than losses by a threshold \( \text{GAIN} \) in its subtree, then the gene is present; if the number of losses is greater than gains by a threshold \( \text{LOSS} \), then the gene is absent; otherwise, it is left unassigned. Finally, starting from the root, unassigned nodes are set to be absent. We determined the thresholds \( \text{GAIN} \) and \( \text{LOSS} \) to be both 0 instead of both 2 mentioned in Ref \( (9) \) to avoid continuously shrinking genomes in the ancestral strains.

We collected the gene rules of each reaction from the xml files of the published metabolic models of the 53 strains. The rules in these models are not always consistent, as for example a gene catalyzing a reaction in one model does not always carry out the same reactions in another model. Thus we compiled a universal set of gene-reaction rules based on the following criteria:

1. if a reaction is catalyzed by a gene (or a group of genes), then in the universal set it is also catalyzed by the same gene (the same group of genes);

2. if a reaction is not catalyzed by any genes in any models, then in the universal set it is not catalyzed by genes.

The metabolic networks of these 53+52 strains are constructed based on their gene content and the universal set of reaction rules; each reaction is present in the metabolic network if any of its corresponding genes (groups of genes) is present in the genome, and all reactions not catalyzed by any genes are always present. As a reference, table S2 shows the number of all genes / metabolic genes / reactions in each of the ancestral+extant strains, tables S3 shows the mapping of ortholog ID used in our study to various gene annotations such as gene locus tag / gene name / gene ID / protein ID in different strains, table S4 shows the mapping of our ortholog ID reaction names, table S5 shows the orthologs present in each of the ancestral+extant strains (ancestral strains are named by the ID of their extant descendants); the metabolic models of the ancestral+extant strains are also included in the Supplementary Materials.
References and Notes

1. Aditya Barve and Andreas Wagner. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461):203–206, August 2013.

2. Louis-Marie Bobay, Marie Touchon, and Eduardo P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences*, 111(33):12127–12132, August 2014.

3. Alexandrine Froger and James E. Hall. Transformation of plasmid DNA into E. coli using the heat shock method. *Journal of visualized experiments : JoVE*, (6), 2007.

4. Gabriel Gelius-Dietrich, Abdelmoneim A. Desouki, Claus J. Fritzemeier, and Martin J. Lercher. sybil ¨a Efficient constraint-based modelling in R. *BMC Systems Biology*, 7(1):125–8, November 2013.

5. N. Luisa Hiller, Azad Ahmed, Evan Powell, Darren P. Martin, Rory Eutsey, Josh Earl, Benjamin Janto, Robert J. Boissy, Justin Hogg, Karen Barbadora, Rangarajan Sampath, Shaun Lonergan, J. Christopher Post, Fen Z. Hu, and Garth D. Ehrlich. Generation of Genic Diversity among Streptococcus pneumoniae Strains via Horizontal Gene Transfer during a Chronic Polyclonal Pediatric Infection. *PLoS Pathog*, 6(9):e1001108+, September 2010.

6. Hermann-Georg G. Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European journal of biochemistry/FEBS*, 271(14):2905–2922, July 2004.

7. Ola Johnsborg, Vegard Eldholm, and Leiv Sigve. Natural genetic transformation: prevalence, mechanisms and function. *Research in microbiology*, 158(10):767–778, December 2007.
8. Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

9. Victor Kunin and Christos A. Ouzounis. GeneTRACEreconstruction of gene content of ancestral species. *Bioinformatics*, 19(11):1412–1416, July 2003.

10. Marcus Lechner, Sven FindeiÁ, Lydia Steiner, Manja Marz, Peter F. Stadler, and Sonja J. Prohaska. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1):124+, April 2011.

11. Xizeng Mao, Qin Ma, Chuan Zhou, Xin Chen, Hanyuan Zhang, Jincai Yang, Fenglou Mao, Wei Lai, and Ying Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic acids research*, 42(Database issue), January 2014.

12. Sergei Maslov, Sandeep Krishna, Tin Y. Pang, and Kim Sneppen. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proceedings of the National Academy of Sciences*, 106(24):9743–9748, June 2009.

13. A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in genetics : TIG*, 17(10):589–596, October 2001.

14. Jonathan M. Monk, Pep Charusanti, Ramy K. Aziz, Joshua A. Lerman, Ned Premyodhin, Jeffrey D. Orth, Adam M. Feist, and Bernhard Ø. Palsson. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences*, 110(50):20338–20343, December 2013.
15. Csaba Pal, Balazs Papp, and Martin J. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12):1372–1375, November 2005.

16. Tin Y. Pang and Sergei Maslov. A Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology. *PLoS Comput Biol*, 7(5):e1001137+, May 2011.

17. Tin Y. Pang and Sergei Maslov. Universal distribution of component frequencies in biological and technological systems. *Proceedings of the National Academy of Sciences*, 110(15):6235–6239, April 2013.

18. Morgan N. Price, Katherine H. Huang, Adam P. Arkin, and Eric J. Alm. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Research*, 15(6):809–819, June 2005.

19. Francisco Rodriguez-Valera, Ana-Belen Martin-Cuadrado, Beltran Rodriguez-Brito, Lejla Pasic, T. Frede Thingstad, Forest Rohwer, and Alex Mira. Explaining microbial population genomics through phage predation. *Nat Rev Micro*, 7(11):828–836, November 2009.

20. Alexandros Stamatakis. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 30(9):btu033–1313, January 2014.