Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic and Genomic Differences

Xavier Prudent,[1,2] Genis Parra,[1,2] Peter Schwede,[1,2] Juliana G. Roscito,[1,2] and Michael Hiller*[,1,2]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

[2]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

*Corresponding author: E-mail: hiller@mpi-cbg.de.

Associate editor: Michael Rosenberg

Abstract

The growing number of sequenced genomes allows us now to address a key question in genetics and evolutionary biology: which genomic changes underlie particular phenotypic changes between species? Previously, we developed a computational framework called Forward Genomics that associates phenotypic to genomic differences by focusing on phenotypes that are independently lost in different lineages. However, our previous implementation had three main limitations. Here, we present two new Forward Genomics methods that overcome these limitations by (1) directly controlling for phylogenetic relatedness, (2) controlling for differences in evolutionary rates, and (3) computing a statistical significance. We demonstrate on large-scale simulated data and on real data that both new methods substantially improve the sensitivity to detect associations between phenotypic and genomic differences. We applied these new methods to detect genomic differences involved in the loss of vision in the blind mole rat and the cape golden mole, two independent subterranean mammals. Forward Genomics identified several genes that are enriched in functions related to eye development and the perception of light, as well as genes involved in the circadian rhythm. These new Forward Genomics methods represent a significant advance in our ability to discover the genomic basis underlying phenotypic differences between species.

Source code: https://github.com/hillerlab/ForwardGenomics/

Key words: **phenotype–genotype associations, evolutionary and comparative genomics, gene loss.**

Introduction

Evolution has led to a striking diversity of phenotypes between species. Many phenotypic differences between species are due to differences in their DNA. Today hundreds of animals have sequenced genomes and many more will be sequenced in future by projects like Genome-10K (Haussler et al. 2009), i5K (Robinson et al. 2011), and individual labs. These genomes provide an unprecedented opportunity to uncover the genomic changes that underlie phenotypic changes between species, which is a key question in genetics and evolutionary biology (Wray 2013; Dunn and Ryan 2015).

Some examples of genomic changes that are involved in phenotypic changes between species have already been discovered. For example, amino acid and *cis*-regulatory differences in *Foxp2* are likely involved in the evolution of human speech (Enard et al. 2009; Maricic et al. 2013). Several hearing genes show convergent amino acid changes in echolocating bats and cetaceans (Li et al. 2008; Liu et al. 2010; Shen et al. 2012), and experiments showed that convergent changes in the prestin gene underlie high-frequency hearing in these echolocating lineages (Liu et al. 2014). Furthermore, losses of ancestral genes have been associated with phenotypic changes (Stedman et al. 2004; Cheng and Detrich 3rd 2007; Meredith et al. 2014). Apart from changes affecting genes, comparative genomics uncovered prominent differences in the noncoding portion of the genome. For example, genomic screens detected noncoding regions that are conserved in chimpanzee and other mammals but have accelerated substitution rates in human or are completely deleted in the human lineage (Pollard et al. 2006a, b; Prabhakar et al. 2006; McLean et al. 2011; Hubisz and Pollard 2014). Experiments provided compelling evidence that some of these human-specific genomic changes are involved in human-specific phenotypes related to specific features of the brain and limb or the loss of penile spines (Pollard et al. 2006b; Prabhakar et al. 2008; McLean et al. 2011; Capra et al. 2013). However, despite this progress, we still know little about which genomic differences are associated with particular phenotypic differences between species. To detect such associations without relying on experiments, computational approaches are helpful, but are challenged by the fact that even closely related species exhibit numerous genomic and phenotypic differences.

To predict associations between phenotypic and genomic differences, we previously developed a computational approach called Forward Genomics that focuses on phenotypes that are independently lost in different species (Hiller et al.

Article

2012b). Forward Genomics relies on two main concepts. First, the genetic information for an ancestral phenotype is often conserved in the descendant species in which this phenotype is maintained due to purifying selection. If the phenotype is lost in a descendant species, this genetic information will evolve neutrally. Thus, genomic regions that only contain information for this phenotype will diverge faster and, over time, will be lost in the trait-loss lineages. This principle is exemplified by the loss of olfactory receptors in whales and dolphins (Kishida et al. 2007; McGowen et al. 2008), the loss of egg yolk genes in placental mammals (Brawand et al. 2008), or the loss of enamel-related genes in birds and other toothless species (Meredith et al. 2011, 2014). Second, the repeated loss of a phenotype will result in a specific genomic signature where those genomic regions that only contain information for this phenotype are conserved in the trait-preserving lineages and are diverged or completely lost in the trait-loss lineages. Forward Genomics uses this specific independent divergence signature to obtain specificity in a genome-wide search for genomic regions involved in the loss of a given phenotype. As a proof of concept, by applying Forward Genomics to the trait "synthesis of vitamin C", which is independently lost in primates, guinea pigs, and many bats, we detected higher sequence divergence in the *Gulo* gene, which encodes the vitamin C synthesizing enzyme (Hiller et al. 2012b). Applied to another phenotype, the absence of phospholipids in the bile of guinea pigs and horses, Forward Genomics detected the loss of the *Abcb4* gene, which encodes a phospholipid-secreting transporter (Hiller et al. 2012b). Forward Genomics has also been used together with transcriptional profiling to detect noncoding genes that might play a role in the evolution of folded brains by searching for noncoding RNAs that are conserved among gyrencephalic mammals and diverged in lissencephalic mammals (Johnson et al. 2015). Finally, the power of matching independent phenotypic changes with independent genomic divergence was further demonstrated in a recent study by Marcovitz, et al. (2016) who successfully detected numerous associations between conserved noncoding regions and morphological and physiological phenotypic changes.

Our previous Forward Genomics implementation quantified sequence divergence for each species by reconstructing the likely ancestral DNA sequence of a given genomic region (Hiller et al. 2012b). We defined sequence divergence between an extant species and the common ancestor of all species of interest as the percent of identical bases: $\%id = id/(id + subs + ins + del) * 100$, where *id*, *subs*, *ins*, and *del* are the numbers of identical bases, substitutions, inserted bases, and deleted bases, respectively. Here, we refer to this value as the global %id value, because divergence is measured between the common ancestor and an extant species (fig. 1A). Although sequence divergence in a trait-loss lineage is likely the result of neutral evolution, %id values alone cannot distinguish between divergence caused by neutral or nonneutral processes (supplementary fig. 1, Supplementary Material online). To associate specific genomic regions to the given phenotype, our previous implementation searches for a region where *all* trait-loss species have a lower global %id value (higher sequence divergence) compared with *all* trait-preserving species (fig. 1B). This method is called "perfect-match".

The perfect-match method has three main limitations, however. First, the phylogenetic relatedness between species is not taken into account. Instead, perfect-match simply matches a phenotypic presence/absence profile to a genomic divergence profile. Second, perfect-match does not control for differences in evolutionary rates (proportional to the branch lengths in fig. 1), which influence the global %id values. Third, perfect-match can only rank candidate loci but it cannot compute the significance of this association.

Given that independent losses of phenotypes are frequent (McGhee 2011; Hiller et al. 2012b) and given the growing number of sequenced genomes, Forward Genomics has broad applicability to predict genomic loci involved in phenotypic changes between species. It is therefore worthwhile to increase the sensitivity to detect such phenotype–genotype associations. Here, we present two new Forward Genomics methods that overcome the three above-mentioned limitations of the perfect-match method by (1) directly controlling for phylogenetic relatedness, (2) controlling for differences in evolutionary rates, and (3) computing a statistical significance. We systematically compare these methods on 32 simulated datasets and on real data and show that both new methods provide a significant advance over the previous perfect-match method. We further used these new methods to detect genomic differences involved in the loss of vision in blind subterranean mammals. This genome-wide screen detected many genes with a function in eye development and the perception of light, as well as genes involved in the circadian rhythm. An implementation of the new methods is available at https://github.com/hillerlab/ForwardGenomics/.

Results

Two New Forward Genomics Methods Control for Phylogenetic Relatedness and Differences in Evolutionary Rates

Both new Forward Genomics methods require a phylogenetic tree with branch lengths proportional to the number of substitutions per neutral site. Since Forward Genomics only considers species with sequenced genomes, we can safely assume that this tree is available or can be inferred from genomic sequence data. Our first new Forward Genomics method makes use of the global %id values of a given genomic region, computed between all extant species of interest and their common ancestor (fig. 1A). To address the three limitations described above, we first control for the phylogenetic relatedness between species by converting relatedness into a covariance matrix, which is a well-known method from phylogenetic comparisons (Grafen 1989; Martins and Hansen 1997; Pagel 1997, 1999). The covariance between two species is the sum of the length of all branches that are shared by these two species, as illustrated in figure 1C. Second, since global %id values are influenced by the total branch length (L) from the common ancestor to an extant

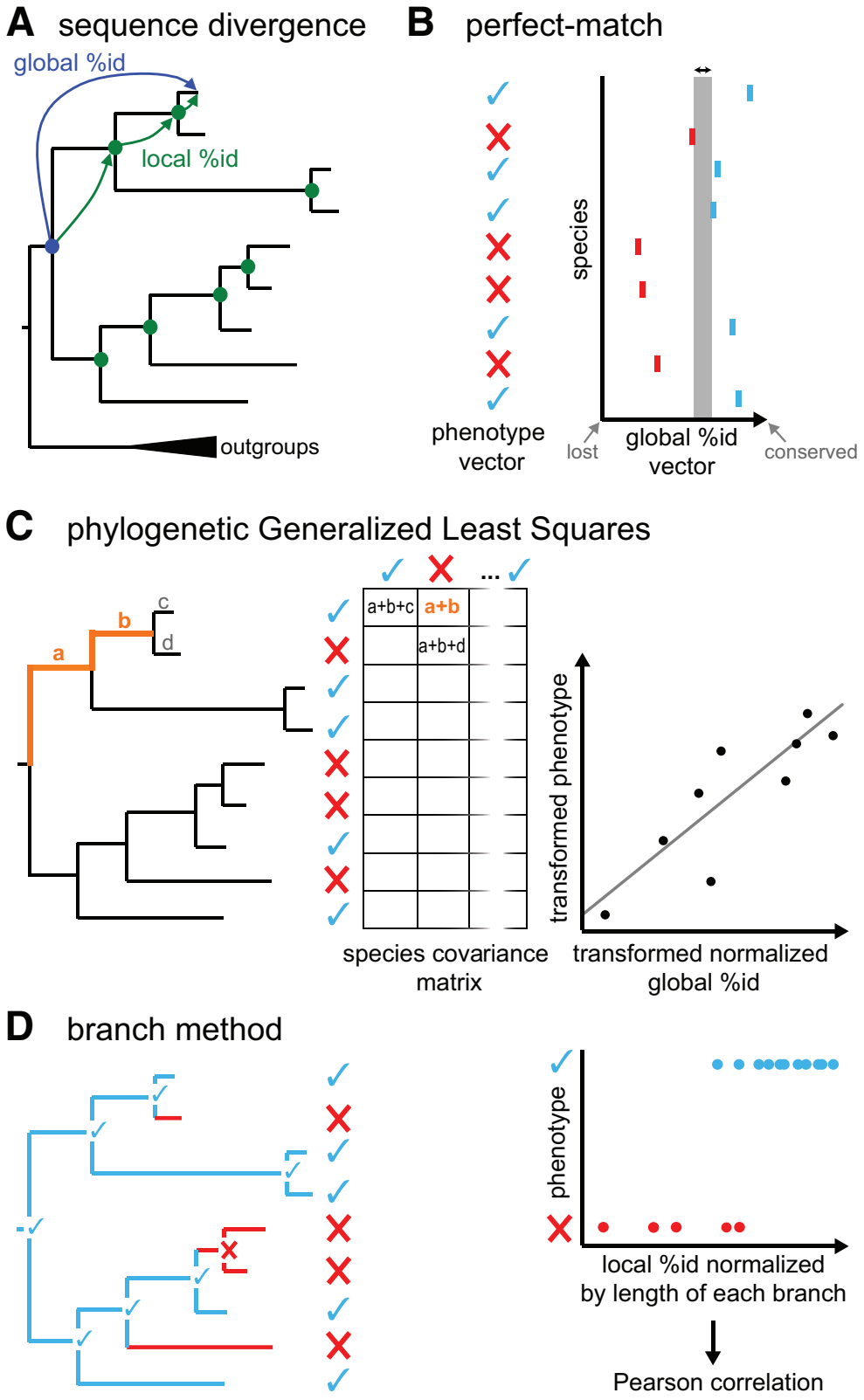

FIG. 1. Overview of three Forward Genomics methods. (*A*) Global %id values are computed by comparing the reconstructed sequence of the common ancestor of the species of interest (blue circle) to the sequence of an extant species. Local %id values are computed between the sequences at the start and end of each branch, which is either a reconstructed ancestral sequence (blue or green circle) or the sequence of an extant species. The branches in the phylogenetic tree are proportional to the number of substitutions per neutral site. Outgroup species are used to reconstruct the common ancestor. (*B*) The perfect-match method (Hiller et al. 2012b) assumes that the given phenotypic presence/absence (checkmark/cross) vector includes trait-losses in independent lineages and conducts a genome-wide search for genomic regions where *all* trait-loss species have a lower global %id value (higher sequence divergence) compared with *all* trait-preserving species. This is illustrated by a positive

species, we control for differences in the evolutionary rates between species by normalizing global %id values as $\%id_{norm} = (\%id - 100)/L$. As shown in supplementary figure 2, Supplementary Material online, this removes the influence of differences in evolutionary rates. Third, to compute the significance of the association between $\%id_{norm}$ and the phenotype, we use a phylogenetic Generalized Least Square (GLS) approach (Grafen 1989; Martins and Hansen 1997; Pagel 1997, 1999) and compute the significance of a positive slope of the linear regression line. This Forward Genomics method is referred to as GLS.

Unlike perfect-match and GLS that use global %id values, the second new method relies on sequence divergence measured for every branch, and is called "branch method". This method first uses parsimony to classify branches into two groups: branches where the trait is lost and branches where the trait is preserved (fig. 1D). Then, for a given genomic element, we test if the trait-loss branches tend to have higher sequence divergence compared with the trait-preserving branches. To this end, we compute a local %id value for every branch in the tree by comparing the sequence at the start and end of a branch (fig. 1A). In contrast to the global %id values, local %id values are independent of each other, as every branch in the phylogenetic tree represents independent evolution. Thus, the branch method does not need to further control for phylogenetic relatedness. However, local %id values are influenced by the length of the respective branch, similar to global %id values that are influenced by the total distance to the common ancestor (supplementary fig. 2, Supplementary Material online). To compare local %id values between the trait-loss and trait-preserving branches, we remove the influence of the branch length by calculating the difference between a given local %id value and the expected value for a branch of the same length evolving under purifying selection. These expected values are pre-computed from simulated data (see Methods). Thus, normalized local %id values >0 or around 0 indicate conservation of ancestral sequence along this branch, while values <0 indicate neutral evolution. To test if trait-loss branches are associated with lower normalized local %id values, we compute the significance of a positive Pearson correlation coefficient (fig. 1D). Like GLS, the branch method controls for phylogenetic relatedness and differences in evolutionary rates, and computes the significance of the association between phenotypic loss and sequence divergence.

Simulating Genome Evolution and Trait Loss Shows That Both New Methods Outperform the Perfect-Match Method

To compare the three Forward Genomics methods (fig. 1), one needs to know which genomic regions are involved in the given phenotypic change (positives) and which genomic regions are not involved (negatives). Since we lack comprehensive knowledge of real phenotypic changes, we created datasets by simulating whole genome evolution and independent trait loss with parameters as realistic as possible. In brief, we evolved an ancestral genome along the placental mammal phylogeny (fig. 2A left, supplementary fig. 3, Supplementary Material online) to obtain 30 simulated genomes of placental mammals. These genomes contain a total of 368,767 functional elements (147,776 coding exons, 220,991 nonexonic elements) that evolve under selection. To assure that simulated data are statistically similar to the real data, we annotated our ancestral genome such that the functional elements in the evolved simulated genomes match the length, nucleotide composition and %id distribution of real genomes (see Methods). In order to simulate independent trait loss, we randomly selected a small subset of 210 functional elements (111 coding exons from 10 genes, and 99 nonexonic elements) that are involved in the loss of this trait (positives) and let them evolve neutrally after trait loss. Thus, 99.94% of all functional elements evolve under selection in all lineages and are considered as negatives. Then, we computed global and local %id values for all functional elements and compared how well the three Forward Genomics methods can distinguish positives from negatives, based solely on sequence divergence.

To comprehensively compare the performance of the three methods, we tested a total of 32 trait-loss scenarios (supplementary table 1, Supplementary Material online) that differ in the age of the trait loss (0.025, 0.05, 0.075, or 0.1 substitutions per neutral site ago), the number of independent trait-loss lineages (between 2 and 4) and the evolutionary rate of the trait-loss lineages (low, medium, and high evolutionary rates, see supplementary fig. 3, Supplementary Material online). Given that the vast majority of functional elements are negatives, receiver operating characteristics plots that compare sensitivity and specificity are not suitable on such highly imbalanced datasets (Saito and Rehmsmeier 2015). The reason is that Forward Genomics achieves a high specificity (a high proportion of genomic regions that are not

grey margin that separates the global %id values of both groups of species. (*C*) The GLS Forward Genomics method derives a covariance matrix that captures the phylogenetic relatedness between species. As illustrated for the first two species, the covariance between two species is the summed length of the branches that are shared between both species (highlighted in orange). The variance of a species is the summed length of all branches from the common ancestor to this species. Lower case letters indicate the length of the branches in the phylogenetic tree. A phylogenetic GLS approach (Grafen 1989) is used to compute a linear regression between the transformed normalized global %id values and the phenotypic pattern. The significance of a positive slope of the regression line is used as the significance of the association between phenotype and genotype. (*D*) The Branch method uses Dollo parsimony to estimate ancestral phenotypic states given the presence/absence pattern of the trait in the extant species. Each branch is then classified as trait-loss (red) or trait-preserving (blue). Local %id values are normalized by the expected value of a branch of the same length. If a genomic region is involved in the trait-loss, we expect that trait-loss branches are associated with lower normalized local %id values. The significance of a positive Pearson correlation coefficient is used as the significance for the association between phenotype and genotype.

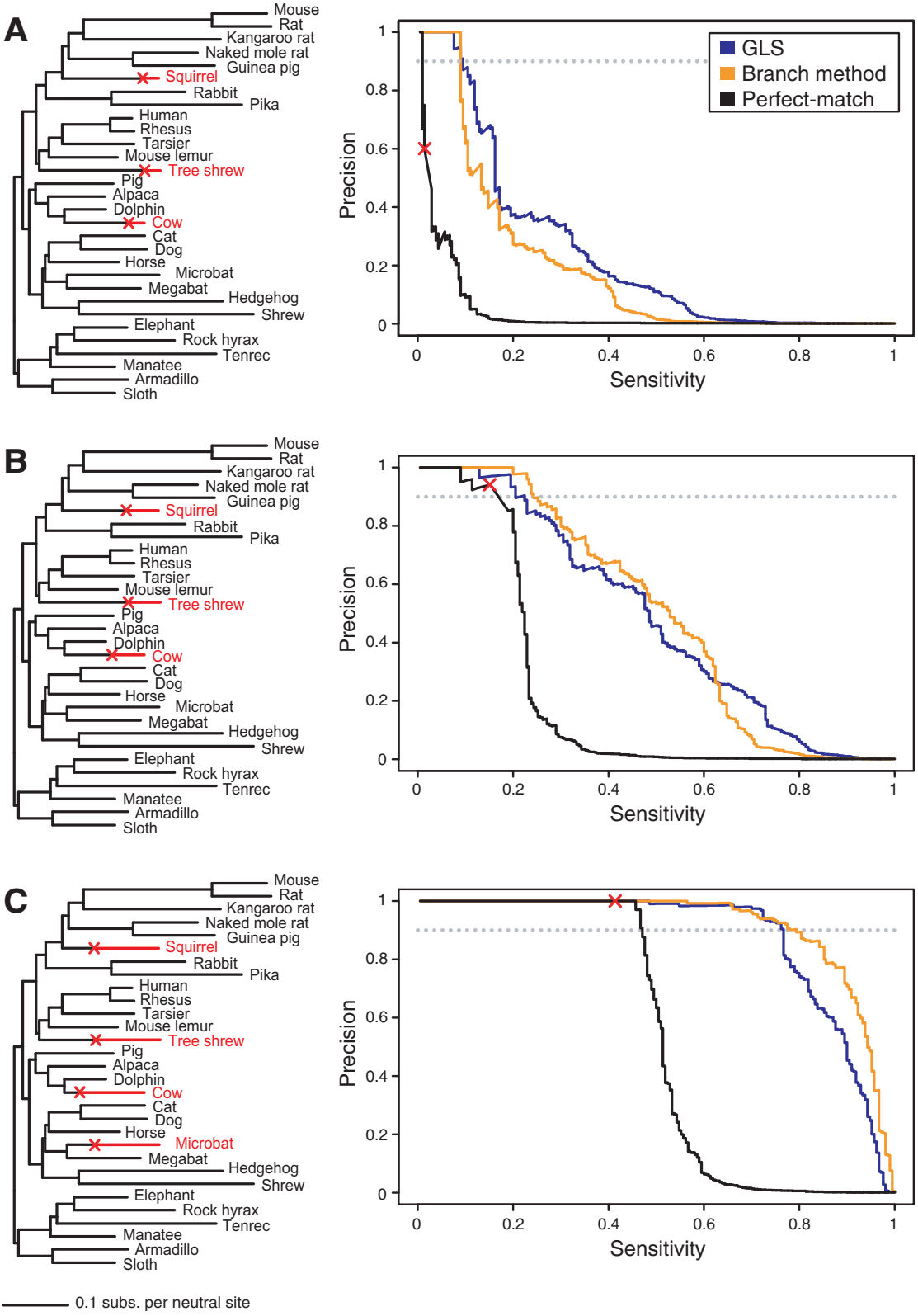

Fig. 2. Performance of the three Forward Genomics methods on simulated data shown by precision-sensitivity plots (right) for three of the 32 trait-loss scenarios (left). Trait losses occurred at the red crosses in the phylogeny, and following trait loss the 210 trait-involved genomic regions evolved neutrally along the parts of the branches shown in red. The red cross in the precision-sensitivity plot for the perfect-match method marks its performance when we consider only genomic regions where *all* trait-loss species have a lower global %id value compared with *all* trait-preserving species. The other 29 trait loss scenarios are shown in supplementary figures 5–33, Supplementary Material online.

involved in the trait loss are correctly identified as such) irrespective of the total number of negatives (supplementary fig. 4, Supplementary Material online). However, the primary goal of Forward Genomics is not to achieve a high specificity; instead we aim at achieving a high precision (also called positive predictive value; Saito and Rehmsmeier 2015), which is defined as the proportion of elements that are involved in the loss of this trait (positives) out of all elements predicted to be involved in the trait loss. For this reason, we compare the sensitivity that is achieved at a certain precision, which depends on the total number of negatives (supplementary fig. 4, Supplementary Material online), and plotted the precision–sensitivity curve for each scenario. The results for three trait loss scenarios are shown in figure 2; the results for the other 29 scenarios are shown in supplementary figures 5–33, Supplementary Material online.

To systematically compare the perfect-match, GLS, and branch method across all 32 scenarios, we used the sensitivity that was achieved for a high precision of 90%. Compared with the perfect-match method, the GLS method achieved the same sensitivity for 3 and a higher sensitivity for 27 scenarios (fig. 3A, supplementary table 1, Supplementary Material online). The branch method achieved a higher sensitivity for all 32 scenarios compared with the perfect-match method. Furthermore, the branch method often outperformed GLS, especially if trait-loss occurred in species with low evolutionary rates.

The GLS and Branch Method Outperform the Perfect-Match Method Regardless of the Properties of the Trait-Loss Scenarios or the Properties of Trait-Involved Genomic Regions

Next, we compared how differences in the trait-loss scenarios affect the performance of the three Forward Genomics methods. To this end, we averaged the sensitivity at a precision of 90% considering only those scenarios (1) where trait-loss happened 0.025, 0.05, 0.075, or 0.1 substitutions per neutral site ago, (2) where 2, 3, or 4 independent trait-losses happened, and (3) where trait-loss happened in lineages with low, medium, or high evolutionary rates (supplementary fig. 3, Supplementary Material online). To assure a fair comparison, we averaged the performance of equivalent trait-loss scenarios (supplementary table 1, Supplementary Material online).

First, all three methods performed substantially better if the trait-loss occurred long ago (fig. 3B). The reason is that neutral evolution over longer evolutionary times will result in higher sequence divergence, which is easier to distinguish from purifying selection. Second, the GLS and branch method performed better for a higher number of independent losses (fig. 3C), which was not the case for the perfect-match method. Third, while the sensitivity of the GLS method varies slightly with differences in the evolutionary rates in the trait loss lineages, the branch and perfect-match method have increased sensitivity if the trait loss includes lineages with low evolutionary rates (fig. 3D).

We further compared how differences in the trait-involved regions affect the performance of each method. For all scenarios where the trait-loss happened 0.05 substitutions per neutral site ago, we averaged the sensitivity at 90% precision separately considering trait-involved elements (1) that differ in length and (2) that are under weak, medium, or strong selection (supplementary table 1, Supplementary Material online). First, we found that the GLS and branch method have increased sensitivity to detect longer elements (fig. 3E). The likely reason is that longer, neutrally evolving elements accumulate mutations more evenly over time. In contrast, shorter, neutrally evolving elements may not accumulate mutations over short timescales simply by chance, which makes them harder to identify. Second, while the sensitivity of the branch method does not depend on the strength of selection, the GLS and perfect-match method have a higher sensitivity to detect trait-involved elements evolving under high constraint (fig. 3F).

Importantly, regardless of the differences in the trait-loss scenarios or differences in the trait-involved elements, the GLS and branch method consistently outperform the perfect-match method. This shows that controlling for phylogenetic relatedness and differences in evolutionary rates improves the sensitivity to detect genomic regions that are involved in phenotypic differences.

The GLS and Branch Method Are Robust to Uncertainties in the Phylogeny

Next, we investigated how robust are the two new methods with respect to precise knowledge of the topology of the phylogenetic tree. While the placental mammalian phylogeny is overall well resolved, there are three conflicting hypothesis for the base of the placental mammals (Morgan et al. 2013; Romiguier et al. 2013). Our simulated genomes evolved along a tree that represents the Atlantogenata hypothesis, where afrotherians and xenarthrans are sister lineages. To test robustness to the tree topology, we used the same simulated data but ran Forward Genomics with trees representing the Epitheria (afrotherians and boreoeutherians are sister lineages) and the Exafroplacentalia (xenarthrans and boreoeutherians are sister lineages) hypothesis (supplementary fig. 34A, Supplementary Material online). We found that the sensitivity that was achieved for a precision of 90% is typically only slightly affected by the uncertainty of the topology at the placental mammal base (fig. 4A).

Next, we tested how uncertainties in the branch lengths affect the results. To this end, we added random noise to each branch length by sampling from a normal distribution with a standard deviation of 0.025. Repeating this three times resulted in three trees where the branch lengths changed on average 0.019 (maximum 0.078) substitutions per neutral site (supplementary fig. 34B, Supplementary Material online). We found that the sensitivity at 90% precision varies depending on the scenario and the tree, however we observed both increases and decreases compared with the sensitivity obtained with the true branch lengths (fig. 4B). Importantly, despite these changes in the topology and the branch lengths, the GLS and the branch method still achieve a higher sensitivity than the perfect-match method for the majority of the 32 trait loss scenarios (fig. 4).

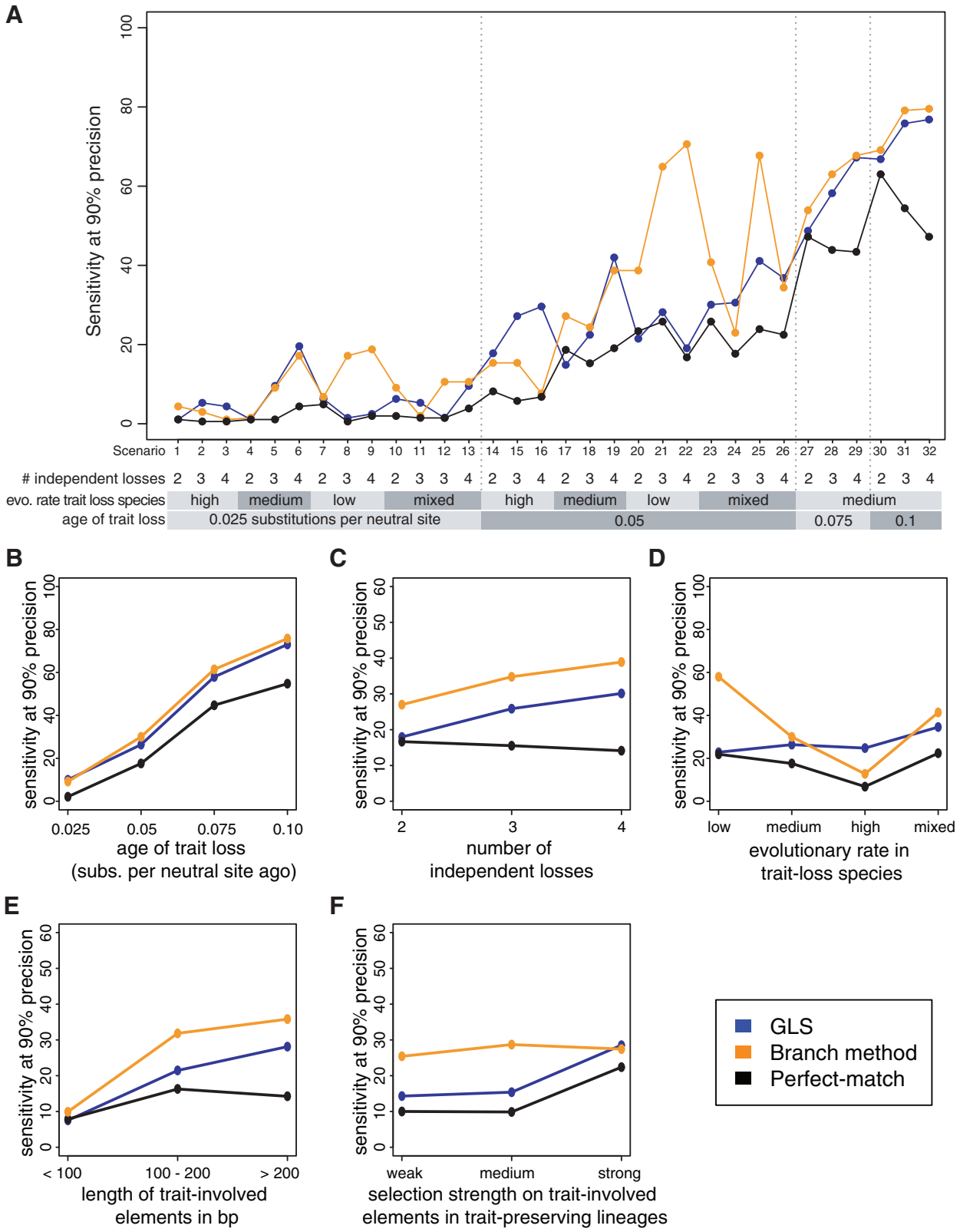

FIG. 3. Performance of the three Forward Genomics methods on 32 trait-loss scenarios. (*A*) The sensitivity at 90% precision is plotted for 32 different trait-loss scenarios. Consistently, the GLS and branch method improve the sensitivity compared with the perfect-match method. (*B*–*F*) Properties of the trait-loss scenarios and properties of the trait-involved genomic regions influence the performance: (*B*) Age of the trait loss, measured by how long the trait-involved elements evolved neutrally; (*C*) number of independent trait-losses; (*D*) evolutionary rate in the trait-loss lineages; (*E*) length of trait-involved elements; and (*F*) strength of selection on trait-involved elements in the branches where they evolve under selection. Weak, medium, or strong refers to genomic regions that accept mutations with an average probability of >0.66, 0.33–0.66, <0.33, respectively.

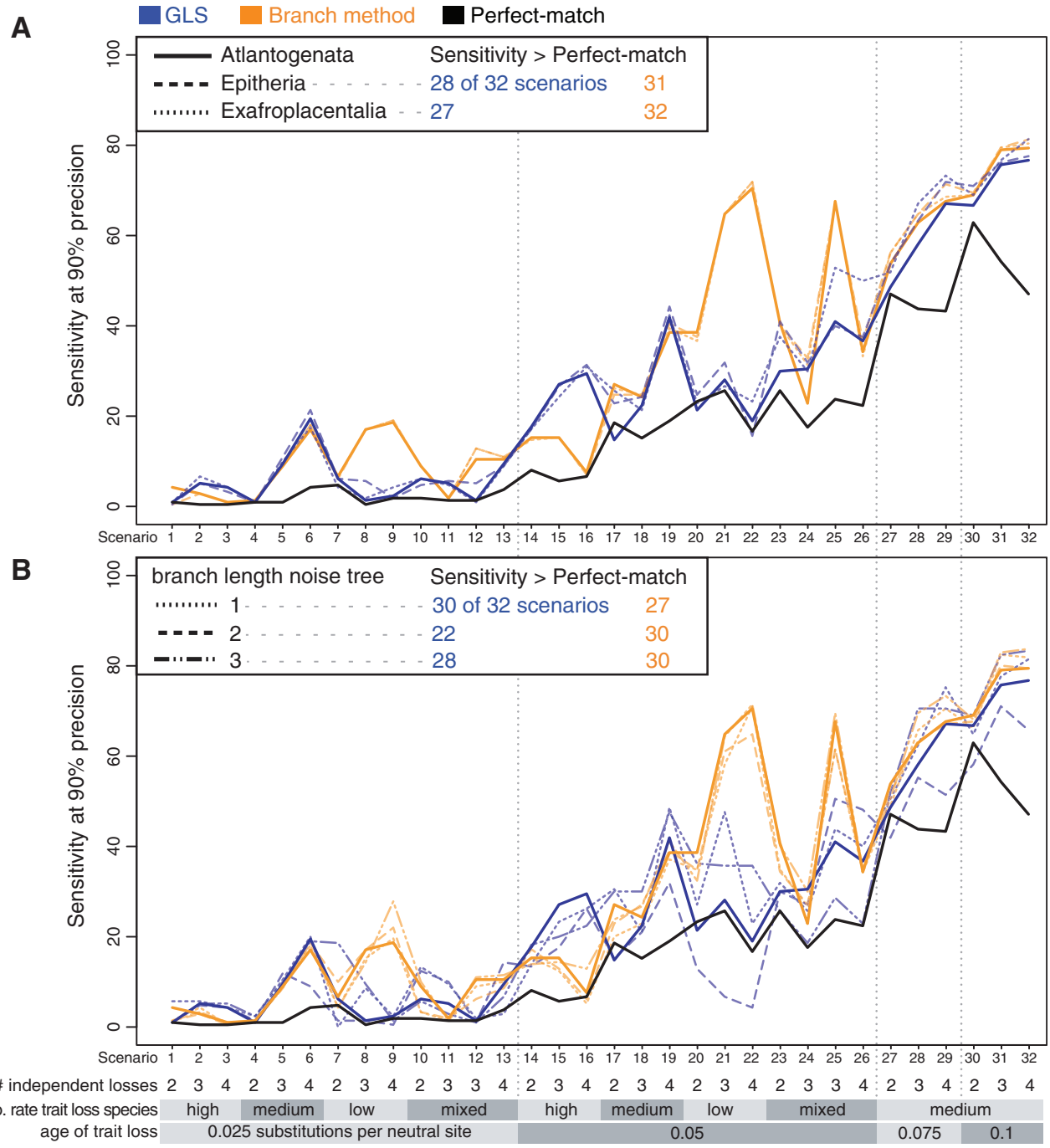

Fig. 4. Robustness of the new Forward Genomics methods to uncertainties in the phylogenetic tree. The sensitivity at a precision of 90% of all 32 trait loss scenarios is shown for (*A*) the Epitheria and the Exafroplacentalia tree topology (supplementary fig. 34*A*, Supplementary Material online) and (*B*) three trees where random noise was added to each branch length (supplementary fig. 34*B*, Supplementary Material online). Solid lines show the results using the phylogeny that was used to produce the simulated data (reproduced from fig. 3*A* for comparison). Please note that the perfect-match method considers neither topology nor branch lengths, thus always gives the same results. The number of scenarios where the achieved sensitivity is higher than the sensitivity of the perfect-match method is shown in the legend.

Both the GLS and Branch Method Outperform Perfect-Match for the Trait "Loss of Vitamin C Synthesis"

To test if the GLS and the branch method also have increased sensitivity on real data, we used the independent loss of Vitamin C synthesis in Haplorrhini primates, in guinea pigs, in many microchiroptera families, and in the megachiroptera *Pteropus vampyrus* as a test case. Since it is well known that the loss of the *Gulo* gene is responsible for this trait loss (Hiller et al. 2012b), we considered all exons of the *Gulo* gene as trait-involved regions and all other 184,723 conserved coding regions as negatives.

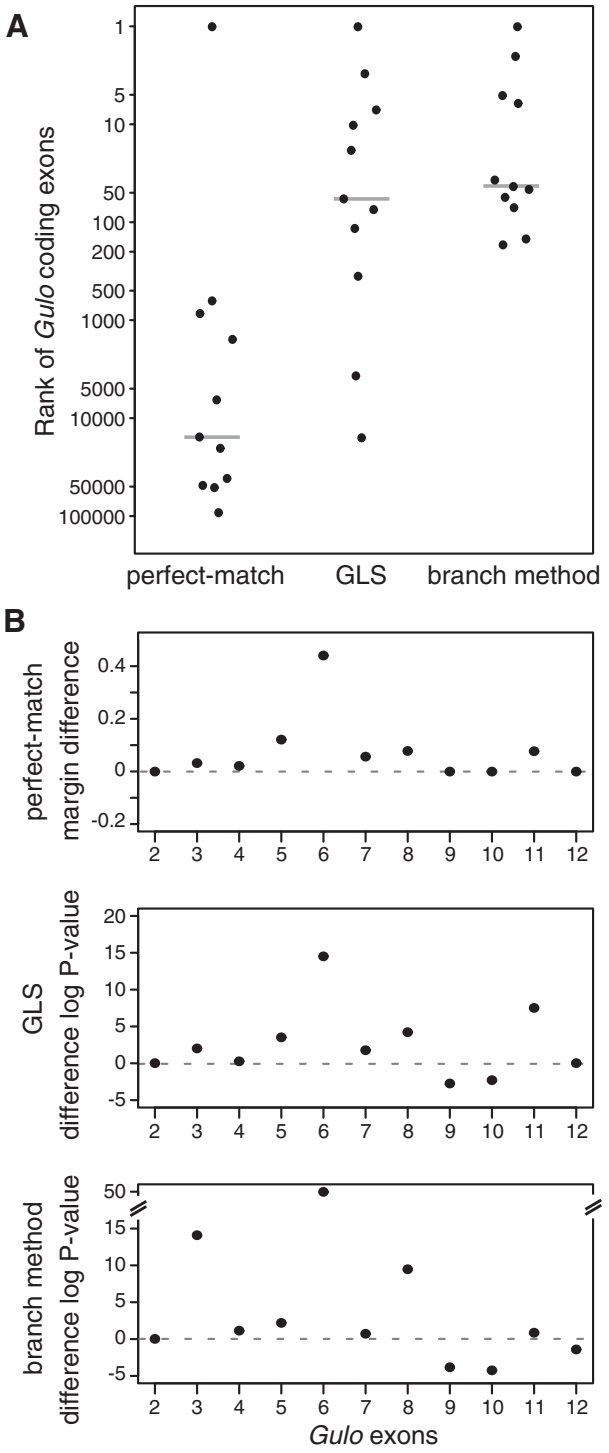

Fig. 5. The GLS and branch method outperform the perfect-match method on the trait "loss of vitamin C synthesis". (*A*) *Gulo* exons are ranked higher with the GLS and the branch method than with perfect-match. For GLS and the branch method, each conserved coding region was ranked by its *P*-value. For perfect-match, we used the size of the margin for ranking, which is the difference between the lowest %id value of a trait-preserving species and the highest %id value of a trait-loss species. *Gulo* exon 2 is ranked first for all three methods. (*B*) The significance of most *Gulo* exons increases if the megabat *P. vampyrus* is excluded from the list of trait loss species. The trait loss in *P. vampyrus* happened more recently than in Haplorrhini primates, guinea pig, and the microbat *M. lucifugus*. We computed the difference between the margin (perfect-match) and the log *P*-value (GLS and branch method) between the screen that used all nonvitamin C synthesizing species and the screen where *P. vampyrus* was excluded. Positive differences indicate a better match to the trait loss. The significance of *Gulo* exons 9 and 10 decreases because both exons are deleted in *P. vampyrus*. *Gulo* exon 1, which only encodes the start codon, is excluded.

Previous results from the perfect-match method using a genome alignment with less species (Hiller et al. 2012b) found only one conserved coding region, which corresponded to *Gulo* exon 2, with lower (not-normalized) global %id values for *all* trait-loss species compared with *all* trait-preserving species (positive margin, as illustrated in fig. 1*B*). Consistent with this, the perfect-match and both new methods detected *Gulo* exon 2 as the top hit (fig. 5*A*) using a newer alignment with more species (Methods). To compare how well the three Forward Genomics methods can identify the conserved coding regions that correspond to the 11 coding exons of *Gulo*, we ranked each coding region. As shown in figure 5*A*, the 11 *Gulo* exons are ranked much higher with the GLS and, in particular, with the branch method, where all 11 exons are within the top 164 hits. This shows that GLS and the branch method outperform perfect-match on real data and that both new methods have increased sensitivity to detect genomic regions involved in the loss of vitamin C synthesis phenotype.

Among the four nonvitamin C synthesizing lineages, the megabat (*P. vampyrus*) has lost this trait most recently, as closely related bats are able to synthesize vitamin C and the *Gulo* gene is intact in other species of the *Pteropus* genus (Cui et al. 2011a, b). Since our simulations showed that there is lower sensitivity to detect genotype–phenotype associations for recent trait losses, we tested if excluding *P. vampyrus* from the list of trait-loss lineages would increase the sensitivity to detect *Gulo* exons. Indeed, the significance of most *Gulo* exons increased for all three Forward Genomics methods (fig. 5*B*), which makes the *Gulo* genomic locus to stand out even more in a genome-wide screen (supplementary figure 35, Supplementary Material online). This supports our findings based on simulated data and suggests a strategy that combines results from two Forward Genomics searches: one search that includes all trait loss species and a second search that may detect additional associations by excluding lineages with a recent trait loss.

The GLS and Branch Method Detect Numerous Genes Associated with the Loss of Vision in Blind Mammals

We applied our new methods to detect genes involved in the loss of vision in two blind mammals, the blind mole rat and the cape golden mole (Fang et al. 2014). Both species live in a subterranean environment, have rudimentary eyes completely covered by skin, and a degenerated visual system (Cooper et al. 1993). We built a genome alignment of these two and 17 other mammals and three nonmammalian outgroup species, and computed local and global %id values of 184,412 conserved coding regions (see Methods). We ranked each gene by the *P*-value and number of exons, which

resulted in a set of 208 genes detected by the GLS and/or branch method.

We found that this set is enriched for genes known to be involved in eye development and perception of light, and genes expressed in the retina and lens (fig. 6, table 1). Furthermore, knockout or mutations in these genes are associated with abnormal eye phenotypes in mouse or eye diseases in human. Among these 208 genes are six crystalline genes that are involved in the formation of the lens and maintenance of its transparency and refractive index. Several other genes code for structural components of the lens, the retina and the cornea, or are involved in the transduction of light stimulus (table 1). The complete list of 36 genes with a function in eye development and vision is given in supplementary table 2, Supplementary Material online. Overall, this shows that the new Forward Genomics methods can detect many genes that are likely involved in the degeneration of the visual system in these two subterranean species.

Apart from genes related to vision, Forward Genomics detected divergence in genes involved in the transduction of light stimulus and regulation of the circadian clock: three gamma-aminobutyric acid (GABA) C receptors (*Gabrr1*, *Gabrr2*, *Gabrr3*) and the melatonin receptor 1b (*Mtnr1b*). These three GABA receptors are primarily found in the retina and are involved in inhibition of light-evoked synaptic currents from retinal neurons (Lukasiewicz et al. 2004). Removal of this inhibition is potentially an adaptation to the low light levels in the burrows, resulting in an increased sensitivity to light which, in turn, might allow for regulating the circadian clock. *Mtnr1b* is also expressed primarily in the retina. This gene is inactivated in the blind mole rat (as shown previously by Fang et al. 2014) as well as in the cape golden mole due to stop codon and frameshift mutations in exon 1 and the deletion of exon 2 (supplementary fig. 36, Supplementary Material online). Furthermore, the naked mole rat, another independent subterranean lineage, has also lost *Mtnr1b* (Kim et al. 2011). Melatonin is a hormone that is produced at night, in response to the absence of light stimulus from the retina, and is the major regulator of the daily biological rhythm of animals. The two blind species analyzed here live in a subterranean, constantly dark environment where minimal light propagates into the burrows (Kott et al. 2014), however, they still have an intrinsic rhythmicity, which can be measured by their thermoregulatory capacity (Haim et al. 1983; Pevet et al. 1984). Melatonin is also known for its role as a potent antioxidant and free radical scavenger (Hardeland et al. 1995). Blind mole rats produce melatonin in the greatly developed Harderian gland, which has high levels of the melatonin-synthesizing enzyme (Balemans et al. 1980) and this was suggested to contribute to protection against higher levels of reactive oxygen species produced by oxidative stress that the animals are subjected to in their hypoxic subterranean environment (Caballero et al. 2006). Given that sustained melatonin levels interfere with the circadian clock (Pevet et al. 1984), the loss of one melatonin receptor might be a way of overcoming potentially deleterious effects on the circadian clock.

Discussion

The growing number of sequenced genomes provides an unprecedented opportunity to use computational approaches to discover which genomic changes underlie

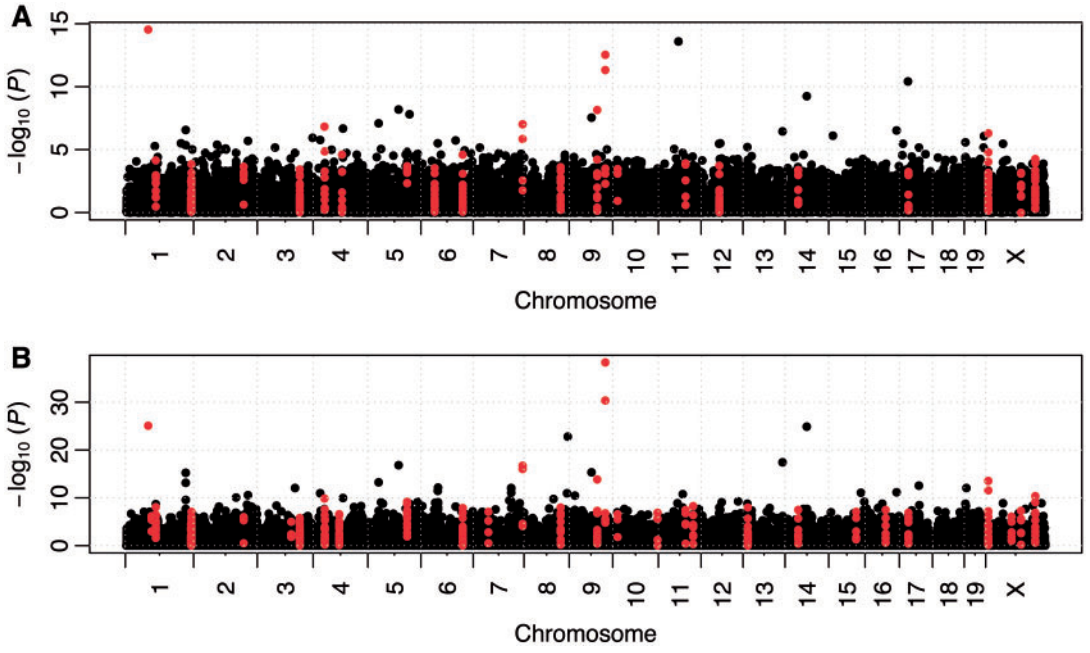

FIG. 6. The GLS and branch method detects several conserved coding regions that are diverged in two blind mammals, the blind mole rat, and the cape golden mole. Manhattan plots show the genomic location of 184,412 conserved coding regions and their associated *P*-values computed by the GLS (*A*) and branch method (*B*). All conserved coding regions that correspond to exons of the genes with a function in eye development and perception of light (supplementary table 2, Supplementary Material online) are shown in red.

particular phenotypic changes between species. While this is challenging due to the large number of genomic and phenotypic differences, computational methods can detect statistical signals for the subset of phenotypic differences that comprise independent losses (Hiller et al. 2012b; Marcovitz et al. 2016). Our previously developed Forward Genomics method searched for matches between a phenotypic presence/absence pattern and a genomic divergence profile, and thus did not account for the phylogenetic relatedness between species. Here, we present two new methods that directly control for phylogenetic relatedness and compute the statistical significance of the associations. Both methods also control for differences in evolutionary rates, which influences sequence divergence. We showed that both methods have substantially improved sensitivity in detecting associations between phenotypic and genomic differences.

Forward Genomics is more effective the older the independent trait losses are, because old losses give enough time for neutral evolution to leave a clearly detectable sequence divergence signal in those genomic regions that are involved in the trait. If trait-loss is known to be recent in some lineages, but older in at least two other independent lineages, it can be advantageous to run a second Forward Genomics search that excludes those lineages where the trait was lost more recently. This second search may detect additional associations that were missed in a first search where all trait loss species were included (fig. 5B). Two sources can be used to estimate when a trait was lost. First, dated fossils can provide information when a trait was still present and when it was lost along one lineage. Second, species divergence times give upper and lower boundaries for when a trait was lost. If a sister lineage to a trait-loss species has also lost the trait, we would assume by parsimony that the loss is as old as the common ancestor. In contrast, if sister species possess this trait, the loss is younger than the split of these species. The latter is illustrated for vitamin C synthesis: it is known that *R. leschenaultia*, a sister species to the nonvitamin C synthesizing bat *P. vampyrus* can synthesize this vitamin (Cui et al. 2011a, b). Both species split around 23 Ma (Hedges et al. 2006), which gives an upper bound for the age of this trait loss. Ideally, one would like

Table 1. Functional Enrichments of the 208 Genes for Which the GLS and Branch Method Detected Increased Divergence in Blind Mammals.

Ontology	Adjusted P-value	Genes
GO biological process		
Sensory perception of light stimulus (GO:0050953)	2.3E−09	CRYBB1;ABCA4;CRYBB3;CRYBA1;KRT12;CRYBB2;CRYBA4;CACNA1F;ARR3;GUCY2F;USH2A;GABRR2;BFSP2;OPN1MW;RDH5;GJA8;RGR;IMPG1
Visual perception (GO:0007601)	2.3E−09	
Sensory perception (GO:0007600)	8.8E−05	CRYBB1;ABCA4;CRYBB3;CRYBA1;CRYBB2;KRT12;CRYBA4;TAAR3;ARR3;CACNA1F;GUCY2F;USH2A;GABRR2;BFSP2;OPN1MW;RDH5;GJA8;RGR;IMPG1
Lens development in camera-type eye (GO:0002088)	0.008	LIM2;CRYBA2;CRYBA1;GJA8;GJE1
Detection of light stimulus (GO:0009583)	0.011	OPN1MW;GJA10;ABCA4;RDH5;CACNA2D4;CACNA1F;RGR;GUCY2F
Detection of visible light (GO:0009584)	0.026	OPN1MW;GJA10;ABCA4;RDH5;CACNA2D4;CACNA1F;GUCY2F
GO molecular function		
Structural constituent of eye lens (GO:0005212)	1.5E−09	LIM2;BFSP2;BFSP1;CRYBB1;CRYBB3;CRYBA2;CRYBB2;CRYBA1;CRYBA4
MGI mammalian phenotype		
MP0005551_abnormal_eye_electrophysiology	7.8E−08	ABCA4;CACNA2D4;CACNA1F;ARR3;USH2A;GUCY2F;GABRR1;GJA10;RDH5;SLC16A8;GJA8;RGS11;RGR
MP0002697_abnormal_eye_size	0.020	LIM2;HECTD1;HSF4;CRYBA1;CRYBB2;GJA8;GJE1
MP0005193_abnormal_anterior_eye	0.018	LIM2;BFSP2;BFSP1;HSF4;KRT12;CRYBB2;CRYBA1;GJA8;LYST;GJE1
MP0003787_abnormal_imprinting	0.032	SNRPN;ARID4A;ARID4B
MP0008877_abnormal_DNA_methylation	0.040	
MP0005253_abnormal_eye_physiology	0.040	BFSP2;ABCA4;RDH5;GJA8;RGR
Human phenotype ontology		
Zonular cataract (HP:0010920)	2.6E−07	BFSP1;CRYBB1;HSF4;CRYBB3;CRYBA1;CRYBB2;CRYBA4;GJA8
Corneal dystrophy (HP:0001131)	0.007	OPN1MW;CRYBB1;CRYBB2;KRT12;CRYBA4;GJA8
Nuclear cataract (HP:0100018)	0.004	CRYBB1;HSF4;CRYBB3;GJA8
OMIM		
Cataract	4.6E−09	LIM2;BFSP2;BFSP1;CRYBB1;CRYBB3;HSF4;CRYBA1;CRYBB2;CRYBA4;GJA8
Human gene atlas		
Retina	0.009	CH25H;ABCA4;RDH5;CRYBB2;SLC16A8;RGR;ARR3;IMPG1;TNS1
Mouse gene atlas		
Lens	0.044	TKTL1;LIM2;UHRF2;CRYBB1;CYP3A44;CRYBB3;CRYBA2;CRYBA1;CRYBB2;CRYBA4;GJE1;BFSP2;BFSP1;EWSR1;GJA10;CAPRIN2;HSF4;PCNX;GJA8;BIRC7;AGFG1
Retina	0.044	PPM1N;ABCA4;BRAF;CACNA2D4;ARR3;USH2A;GABRR2;GABRR1;PIK3CA;OPN1MW;BC030499;UBN2;FBXL5;IMPG1;DRD4;SERINC4
Pfam InterPro domains		
Crystallin	5.8E−05	CRYBB1;CRYBA2;CRYBA1;CRYBB2;CRYBA4

Enrichments were computed by Enrichr (Chen et al. 2013).

to estimate how much sequence divergence can be expected in a lineage after trait-loss. This requires a phylogenetic tree where branch lengths correspond to substitutions per neutral site. The increasing availability of species with sequenced genomes will facilitate reliably estimating such molecular phylogenies.

While more and more genomes are sequenced, there is still a substantial mismatch between the number of species with sequenced genomes and the number of species for which phenotypic data are available. For example, out of 46 mammals included in a very large phenotype dataset (O'Leary et al. 2013), only 20 have sequenced genomes (Marcovitz et al. 2016). To effectively utilize the present and future genomes to associate phenotypic differences and genomic differences, comprehensive phenotypic knowledge of the sequenced species must be accessible to computers (Deans et al. 2012, 2015), as opposed to free text descriptions in which the majority of our phenotypic knowledge is still described. Phenotype databases such as Morphobank (O'Leary and Kaufman 2011) that provide phenotypic character matrices have been successfully used to associate phenotypic and genomic differences (Marcovitz et al. 2016). However, in order to reuse, search, and compare phenotypic data, it is necessary to use ontologies that provide a defined vocabulary of anatomical features and relationships between them. The Phenoscape Knowledgebase uses ontologies to describe phenotypes of natural species and phenotypes of model organism mutants (Dahdul et al. 2010), which allows deriving hypotheses which genes may be involved in phenotypic differences between natural species (Manda et al. 2015). For example, candidate genes for the loss of the tongue and scales in the catfish lineage were detected by comparing natural and mutant phenotypes, and indeed, these genes have different expression patterns in the channel catfish that are consistent with their involvement in these phenotypic losses (Edmunds et al. 2016). Another advantage of computer-interpretable phenotype data is that missing data, which is common in phenotypic character matrices, can be drastically reduced by using machine reasoning to infer presence/absence states (Dececchi et al. 2015), which in turn will broaden the applicability of approaches like Forward Genomics.

The two new methods presented here represent a significant advance in our ability to discover the genomic basis underlying phenotypic differences between species. With the increasing number of sequenced genomes and with an increasing accessibility of phenotypic knowledge, these Forward Genomics methods will contribute to our understanding of how nature's phenotypic diversity has evolved.

Methods

GLS

In the GLS method, we control for the phylogenetic relatedness between species by computing the covariance matrix based on a phylogenetic tree. Given a tree with n species, the elements in the $n \times n$ covariance matrix R are defined as $R_{ii} = L_i$ and $R_{ij} = L_{ij}$ where L_i is the total branch length from the common ancestor to species i and L_{ij} is the total branch length shared by species i and j (fig. 1C). Then, we use a phylogenetic GLSs approach (Grafen 1989), implemented in the R package caper (https://cran.r-project.org/web/packages/caper/), to compute a linear regression between the normalized %id values and the phenotype. The P-value of a positive slope of the regression line is used as the significance of the association between the genomic and phenotypic difference.

Branch Method

The branch method computes a local %id value, which corresponds to the divergence between the sequence at the start and end of a given branch (fig. 1C). To remove the influence of the branch length on the local %id value, we pre-computed the expected local %id value of a branch of length b that evolves under selection. To this end, we simulated genome evolution for b substitutions per neutral site (see below) and averaged the local %id value of all functional elements. Supplementary table 3, Supplementary Material online, shows the expected local %id values for branch lengths varying from 0.01 to 1.0 substitutions per neutral site in steps of 0.01. Then, we obtained a normalized local %id value by calculating the difference between the given local %id value and the local %id value that is expected for a branch of the same length evolving under selection.

To classify branches as trait-loss or trait-preserving, we used Dollo parsimony, which allows for an unambiguous reconstruction of ancestral character states by assuming that lost traits cannot be regained. Alternatively, one could use maximum likelihood to reconstruct ancestral character states, which allows lost traits to be regained. Branches where a trait was likely regained should then be excluded, as they might not fully preserve the ancestral trait information, which would confound the analysis. It should be noted that for all our simulated scenarios, for the loss of vitamin C synthesis, and for the loss of vision, both maximum likelihood and Dollo parsimony lead to the same branch classification.

Individual branches in a tree can be very short. On such short branches, it is more likely that not enough random mutations occur such that the sequence divergence of a neutrally evolving element can be distinguished from an element that evolves under selection. Therefore, we assigned a branch-length-dependent weight to each branch that is proportional to the power to detect neutral evolution along this branch. To compute the weight for a branch of length b, we first obtained the local %id distribution for elements evolving neutrally and elements evolving under selection by simulating genome evolution for b substitutions per site. Then, we define $f(v,\ \text{neutral})$ and $f(v,\ \text{selection})$ as the fraction of the neutral and selection distribution below the %id value v. The weight of a branch of length b is then $\max_v\ (f(v,\ \text{neutral}) - f(v,\ \text{selection}))$. Thus, if the two distributions are nonoverlapping, the weight will be $\sim$1. If the two distributions are similar as it is expected for short branches, the weight will be $<$1. Because the average constraint differs between coding regions and nonexonic elements, weights and expected local %id values were computed separately for these two groups (supplementary

fig. 37 and supplementary table 4, Supplementary Material online). These weights are used to calculate a weighted Pearson correlation between the normalized local %id values of the trait-loss and the trait-preserving branches.

Annotating the Ancestral Genome for Simulating Genome Evolution

We used Evolver (http://www.drive5.com/evolver/) to simulate the evolution of an entire ancestral genome along a phylogeny. Evolver models genome evolution including substitutions, insertions, and deletions, transposon insertions, and tandem repeat expansion and contraction. Important for our purpose is that Evolver uses an ancestral genome, where genes (untranslated and coding regions, start and stop codons, and splice sites) and nonexonic functional elements are explicitly annotated and the bases in functional regions evolve under a specified level of constraint. Evolver has an explicit model of protein evolution and maintains the gene structure for genes under selection.

Evolver requires an ancestral genome, where the type, position, and constraint of functional elements are annotated, and a phylogeny along which the genome will evolve. To assure that we compare the different Forward Genomics methods on realistic data, we created an ancestral genome annotation such that the functional elements in the simulated evolved genomes match the length, nucleotide composition and %id distribution of real genomes. We chose the mouse genome as our ancestral genome because it is well assembled and well annotated. First, we replaced *N*'s by random bases in chromosome 1–19 of the mouse mm10 assembly using Evolver's "evo –findns". Second, we assigned the position of functional elements to this genome. To annotate the position of 5′ and 3′ untranslated regions and coding regions of genes, we used the longest isoform of knownGenes from the UCSC genome browser (Rosenbloom et al. 2015). To annotate the position of conserved nonexonic elements in the ancestral genome, we used PhastCons elements (Siepel et al. 2005) from the UCSC mouse mm10 60way alignment (Rosenbloom et al. 2015) that are longer than 70 bp and that do not overlap exons. This results in a length and nucleotide composition distribution that is comparable to real data. Third, we needed to adjust the evolutionary constraint in the ancestral genome such that simulating genome evolution produces a distribution of global %id values that matches the real %id values. While we use the global %id value as a proxy for constraint, Evolver uses an "acceptance probability" that specifies the probability of a base to accept a mutation. Therefore, we created a map between global %id values and mean acceptance probability. To this end, in a first pass, we evolved a genome with randomly assigned mean acceptance probabilities for 0.19 substitutions per neutral site, which is the distance of human to the placental mammal ancestor, and measured the global %id value of all functional elements. Then we grouped all elements according to their %id value into bins of width 2%. Thus, for each %id value bin, we get the distribution of mean acceptance probabilities that resulted in evolved elements with this %id value. Second, we used this map to assign mean acceptance probabilities to each functional element in the ancestral genome based on the real human %id value. Specifically, we iterated over each functional element, obtained the real human %id value and then sampled from the mean acceptance probability distribution of the respective bin. Given the mean acceptance probability of an element, we used Evolver's "evo –assprobs" to assign base-wise acceptance probabilities to each base in this functional element. We excluded functional elements that are shorter than 70 bp, which results 147,776 coding exons and 220,991 nonexonic elements (368,767 elements in total) in the final ancestral genome.

Simulating Genome Evolution and 32 Different Trait-Loss Scenarios

Simulating the entire phylogeny for all 19 chromosomes requires 1.1 TB disk space and over 65 CPU days, which is not feasible for testing many different trait-loss scenarios. To reduce runtime and disk space, we selected the 210 trait-involved elements only from chromosome 1. Then we simulated the evolution of chromosome 1 for each trait-loss scenario and extracted the %id values of the 210 trait-involved elements (positives). To obtain the %id values of all negative elements, we simulated the evolution of chromosome 1–19 only once, evolving all 368,767 functional elements under selection in all lineages. The %id values of these negatives were used in all trait-loss scenarios.

To assure comparability between scenarios, we used a fixed set of trait-involved elements for all scenarios. This set comprises a total of 210 randomly selected elements (111 coding exons, 99 nonexonic elements) that are a representative sample as they closely match the length and mean acceptance probability distribution of all functional elements (supplementary fig. 38, Supplementary Material online).

For each scenario, we evolved chromosome 1 of the ancestral genome along the placental mammal phylogeny. The phylogenetic tree of placental mammals with branch length values corresponding to substitutions per neutral site was downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/mm10/multiz60way/. To simulate independent trait-loss, we stopped the simulation 0.025, 0.05, 0.075, or 0.1 substitutions per site before reaching the endpoint of a branch leading to a trait-loss species. Then, we removed any constraint on the 210 trait-involved regions by setting the probability of accepting a mutation to 1 and restarted the simulation. Global and local %id values were computed for all elements and used for Forward Genomics.

Applying Forward Genomics to the Loss of Vitamin C Synthesis

We used the UCSC 60way genome alignment where 59 vertebrates are aligned to the mouse mm10 assembly (Rosenbloom et al. 2015). To detect conserved regions in this alignment, we used PhastCons (Siepel et al. 2005) with the parameters expected-length = 45, target-coverage = 0.3 and rho = 0.3, and GERP (Davydov et al. 2010) with default parameters. Then, we extracted conserved coding regions by intersecting conserved elements with coding exons from the

Ensembl, RefSeq, and UCSC knownGenes annotation. We required that each conserved coding region is at least 30 bp. This resulted in 184,723 conserved coding regions covering 27.1 Mb (1.03% of the mm10 assembly). Percent identity values were computed by reconstructing ancestral sequences as described in Hiller et al. (2012b). Bases with a low-quality score (>1% error rate) for the assemblies where quality scores are available were ignored in the %id calculation. Similarly, unaligning regions that map to assembly gaps were ignored and not counted as lost ancestral sequence (Hiller et al. 2012a). To screen for regions associated with the loss of vitamin C, we applied our R implementation of the three Forward Genomics methods (https://github.com/hillerlab/ForwardGenomics/) using the following species as trait-loss species: microbat (myoLuc2), megabat (pteVam1), tarsier (tarSyr1), squirrel monkey (saiBol1), marmoset (calJac3), baboon (papHam1), rhesus macaque (rheMac3), gibbon (nomLeu2), orangutan (ponAbe2), gorilla (gorGor3), chimpanzee (panTro4), human (hg19), and guinea pig (cavPor3).

Applying Forward Genomics to the Loss of Vision in Blind Mammals

To analyze the loss of vision in the blind mole rat and cape golden mole, we first build a genome alignment with mouse as the reference species that included both blind species. Specifically, we used the UCSC lastz/chain/net pipeline (Kent et al. 2003) to build pairwise genome alignments between mouse (mm10 assembly) and the following species: rat (rn5), guinea pig (cavPor3), pika (ochPri3), rabbit (oryCun2), prairie vole (micOch1), blind mole rat (nanGal1), squirrel (speTri2), human (hg19), crab-eating macaque (macFas5), bushbaby (otoGar3), cow (bosTau7), dog (canFam3), horse (equCab2), cat (felCat5), elephant (loxAfr3), manatee (triMan1), cape golden mole (chrAsi1), opossum (monDom5), Anolis lizard (anoCar2), chicken (galGal4), and frog (xenTro7). For all species, we used lastz (Schwartz et al. 2003) version 1.03.54 with the parameters $H = 2{,}000$ $Y = 3{,}000$ $L = 3{,}000$ $K = 2{,}400$, and the HoxD55 scoring matrix, and kept all local alignment that have at least one ≥30 bp region with ≥60% sequence identity and ≥1.8 bits entropy as described in Hiller et al. (2013). For all nonmammalian species, we additionally used highly-sensitive local alignments (Hiller et al. 2013) with lastz parameters $W = 5$, $L = 2{,}700$, and $K = 2{,}000$. For mammals, we kept only alignment chains with a score of ≥70,000 that span ≥9,000 bp in both genomes. In order to keep also chains with very strong alignments spanning only a shorter region, we also kept chains with a score of ≥150,000 that span ≥6,000 bp in both genomes. For nonmammals, we kept only alignment chains with a score of ≥15,000. All other chains are discarded as they typically do not represent strong syntenic alignments. Chains were 'netted' using chainNet (Kent et al. 2003). The pairwise syntenic alignment nets are the input to MULTIZ (Blanchette et al. 2004) to build a multiple alignment. The neutral distances between all species were determined using phyloFit (Siepel et al. 2005) and 4-fold degenerate sites. The tree with branch lengths measuring substitutions per neutral site is given in supplementary figure 39, Supplementary Material online. As above, we used PhastCons and GERP to obtain 184,412 conserved coding regions covering (27.4 Mb, 1.04% of the mm10 assembly). After applying the GLS and branch method to all conserved coding regions, we selected those multi-exon genes where at least two exons are in the top 1,000 of the most significant hits and selected those single exon genes that are in the same top 1,000 hits. This resulted in a list of 141 (124 multi-exon and 17 single exon) genes for the GLS method and 164 (132 multi-exon and 32 single exon) genes for the branch method. The union of both lists comprises 208 genes. We used Enrichr (Chen et al. 2013) to detect functional enrichments of these 208 genes (table 1). Similar enrichments related to eye and vision were also found for the individual sets of 141 and 164 genes, however the 164 genes detected by the branch method have additional functional enrichments (supplementary table 5, Supplementary Material online).

Data Availability

The following data are available at http://bds.mpi-cbg.de/hillerlab/ForwardGenomics/: data of all 32 simulated trait loss scenarios (%id values, output of the Forward Genomics implementation, all scripts to reproduce the results), phastCons and GERP conservation scores (bigWig format) and the conserved elements (bed format), the conserved coding regions (bed format) and their local and global %id values that we used for the loss of vitamin C and loss of vision phenotype, and the genome alignment (maf format, 9.5 GB) that includes the blind mammals and the associated phylogenetic tree. An R implementation of the three Forward Genomics methods is available at https://github.com/hillerlab/ForwardGenomics/.

Supplementary Material

Supplementary tables 1–5 and figures 1–39 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Virag Sharma for help with *Mtnr1b* mutations, Peter Steinbach and Ulf Markwardt for a fast transalign implementation, David Orme for help with the Caper R package, and the Computer Service Facilities of the MPI-CBG and MPI-PKS for their support. This work was supported by the Max Planck Society and fellowship 2012/01319-8 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) to JGR.

References

Balemans MG, Pevet P, Legerstee WC, Nevo E. 1980. Preliminary investigations of melatonin and 5-methoxy-tryptophol synthesis in the pineal, retina, and harderian gland of the mole rat and in the pineal of the mouse "eyeless". *J Neural Transm*. 49:247–255.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 14:708–715.

Brawand D, Wahli W, Kaessmann H. 2008. Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol*. 6:e63.

Caballero B, Tomas-Zapico C, Vega-Naredo I, Sierra V, Tolivia D, Hardeland R, Rodriguez-Colunga MJ, Joel A, Nevo E, Avivi A, et al. 2006. Antioxidant activity in *Spalax ehrenbergi*: a possible adaptation to underground stress. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*. 192:753–759.

Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci*. 368:20130025.

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128.

Cheng CH, Detrich HW. 3rd. 2007. Molecular ecophysiology of Antarctic notothenioid fishes. *Philos Trans R Soc Lond B Biol Sci*. 362:2215–2232.

Cooper HM, Herbin M, Nevo E. 1993. Visual system of a naturally microphthalmic mammal: the blind mole rat, *Spalax ehrenbergi*. *J Comp Neurol*. 328:313–350.

Cui J, Pan YH, Zhang Y, Jones G, Zhang S. 2011a. Progressive pseudogenization: vitamin C synthesis and its loss in bats. *Mol Biol Evol*. 28:1025–1031.

Cui J, Yuan X, Wang L, Jones G, Zhang S. 2011b. Recent loss of vitamin C biosynthesis ability in bats. *PLoS One* 6:e27114.

Dahdul WM, Balhoff JP, Engeman J, Grande T, Hilton EJ, Kothari C, Lapp H, Lundberg JG, Midford PE, Vision TJ, et al. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS One* 5:e10708.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP ++. *PLoS Comput Biol*. 6:e1001025.

Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, et al. 2015. Finding our way through phenotypes. *PLoS Biol*. 13:e1002033.

Deans AR, Yoder MJ, Balhoff JP. 2012. Time to change how we describe biodiversity. *Trends Ecol Evol*. 27:78–84.

Dececchi TA, Balhoff JP, Lapp H, Mabee PM. 2015. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Syst Biol*. 64:936–952.

Dunn CW, Ryan JF. 2015. The evolution of animal genomes. *Curr Opin Genet Dev*. 35:25–32.

Edmunds RC, Su B, Balhoff JP, Eames BF, Dahdul WM, Lapp H, Lundberg JG, Vision TJ, Dunham RA, Mabee PM, et al. 2016. Phenoscape: identifying candidate genes for evolutionary phenotypes. *Mol Biol Evol*. 33:13–24.

Enard W, Gehre S, Hammerschmidt K, Holter SM, Blass T, Somel M, Bruckner MK, Schreiweis C, Winter C, Sohr R, et al. 2009. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* 137:961–971.

Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. *Nat Commun*. 5:3966.

Grafen A. 1989. The phylogenetic regression. *Philos Trans R Soc Lond B Biol Sci*. 326:119–157.

Haim A, Heth G, Pratt H, Nevo E. 1983. Photoperiodic effects on thermoregulation in a 'blind' subterranean mammal. *J Exp Biol*. 107:59–64.

Hardeland R, Balzer I, Poeggeler B, Fuhrberg B, Uria H, Behrmann G, Wolf R, Meyer TJ, Reiter RJ. 1995. On the primary functions of melatonin in evolution: mediation of photoperiodic signals in a unicell, photooxidation, and scavenging of free radicals. *J Pineal Res*. 18:104–111.

Haussler D, O'Brien S, Ryder O, Barker F, Clamp M, Crawford A, Hanner R, Hanotte O, Johnson W, McGuire J, et al. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*. 100:659–674.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.

Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, Bejerano G. 2013. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. *Nucleic Acids Res*. 41:e151.

Hiller M, Schaar BT, Bejerano G. 2012a. Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Res*. 40:11463–11476.

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G. 2012b. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep*. 2:817–823.

Hubisz MJ, Pollard KS. 2014. Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr Opin Genet Dev* 29:15–21.

Johnson MB, Wang PP, Atabay KD, Murphy EA, Doan RN, Hecht JL, Walsh CA. 2015. Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat Neurosci*. 18:637–646.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*. 100:11484–11489.

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479:223–227.

Kishida T, Kubota S, Shirayama Y, Fukami H. 2007. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: evidence for reduction of the functional proportions in cetaceans. *Biol Lett*. 3:428–430.

Kott O, Moritz R, Šumbera R, Burda H, Nemec P. 2014. Light propagation in burrows of subterranean rodents: tunnel system architecture but not photoreceptor sensitivity limits light sensation range. *J Zool*. 294:68–76.

Li G, Wang J, Rossiter SJ, Jones G, Cotton JA, Zhang S. 2008. The hearing gene Prestin reunites echolocating bats. *Proc Natl Acad Sci U S A*. 105:13959–13964.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol*. 20:R53–R54.

Liu Z, Qi FY, Zhou X, Ren HQ, Shi P. 2014. Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals. *Mol Biol Evol*. 31:2415–2424.

Lukasiewicz PD, Eggers ED, Sagdullaev BT, McCall MA. 2004. GABAC receptor-mediated inhibition in the retina. *Vision Res*. 44:3289–3296.

Manda P, Balhoff JP, Lapp H, Mabee P, Vision TJ. 2015. Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. *Genesis* 53:561–571.

Marcovitz A, Jia R, Bejerano G. 2016. "Reverse genomics" predicts function of human conserved non-coding elements. *Mol Biol Evol*. 33:1358–1369.

Maricic T, Gunther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, Naumann R, Burbano HA, Meyer M, Lalueza-Fox C, et al. 2013. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Mol Biol Evol*. 30:844–852.

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat*. 149:646–667.

McGhee G. 2011. Convergent evolution: limited forms most beautiful. Cambridge, MA: The MIT Press.

McGowen MR, Clark C, Gatesy J. 2008. The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. *Syst Biol*. 57:574–590.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471:216–219.

Meredith RW, Gatesy J, Cheng J, Springer MS. 2011. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. *Proc Biol Sci*. 278:993–1002.

Meredith RW, Zhang G, Gilbert MT, Jarvis ED, Springer MS. 2014. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* 346:1254390.

Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. 2013. Heterogeneous models place the root of the placental mammal phylogeny. *Mol Biol Evol*. 30:2145–2156.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, et al. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* 339:662–667.

O'Leary MA, Kaufman S. 2011. MorphoBank: phylophenomics in the "cloud". *Cladistics* 27:529–537.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. *Zool Scripta* 26:331–348.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Pevet P, Heth G, Hiam A, Nevo E. 1984. Photoperiod perception in the blind mole rat (*Spalax ehrenbergi*, Nehring): involvement of the Harderian gland, atrophied eyes, and melatonin. *J Exp Zool*. 232:41–50.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* 2:e168.

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786.

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, et al. 2008. Human-specific gain of function in a developmental enhancer. *Science* 321:1346–1350.

Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ. 2011. Creating a buzz about insect genomes. *Science* 331:1386.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol Biol Evol* 30:2134–2144.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*. 43:D670–D681.

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10:e0118432.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res*. 13:103–107.

Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genetics* 8:e1002788.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034–1050.

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428:415–418.

Wray GA. 2013. Genomics and the evolution of phenotypic traits. *Annu Rev Ecol Evol Syst*. 44:51–72.