J Mol Med (2012) 90:471–480
DOI 10.1007/s00109-012-0901-4

REVIEW

Functional primate genomics—leveraging the medical potential

Wolfgang Enard

Received: 5 March 2012 / Revised: 4 April 2012 / Accepted: 5 April 2012 / Published online: 4 May 2012

Abstract Within biomedicine, comparative genomics is crucial for interpreting human genetic variants and building proper animal models. As our closest relatives, primates are of particular relevance in this frame work. Here, I review principles and concrete examples of this approach. Since one can expect that generating the necessary genomic DNA sequences will not be the major limiting factor in the near future, I argue that in analogy to human biomedicine, comprehensive phenotyping of different primates will be a crucial next step to tap the full potential of comparative genomics. Especially the possibility to generate pluripotent stem cells from primates should allow extending the comparative approach to many medically relevant questions.

Keywords Evolutionary medicine · Comparative genomics · Primates · APOE · LPA

Introduction

Just as politics does usually not need to care about history, medicine does usually not need to consider evolution. However, for a more complete understanding, history matters and this is increasingly realized in biomedicine [1, 2]. Maybe best established are areas in which evolutionary processes can be observed directly. This includes the evolution of pathogens [3], pathogen resistance [4], and the development of cancer [5]. But also the inference of the genetic past has in recent years become medically relevant. One reason is that genome-wide association studies use statistical tools rooted in population genetic theory, for example to control for population stratification [6]. Another is that genetic variants can reach high frequencies due to selection, and identifying such loci could be medically informative [7]. Genomic comparisons across species have also increasingly become possible, but so far had a relatively limited or at least a not well-recognized impact on medicine. However, the use of model organisms such as the mouse is inherently an evolutionary question, and genetic conservation across species is currently the central tool to interpret human genetic variants associated with diseases. With the prospect that 10,000 vertebrate genomes might be available soon [8], I review areas in which the comparative genomic approach has been directly relevant for medical questions. I focus on primates (Fig. 1) that are of special relevance due to their close relationship with humans and try to give a perspective on how the incorporation of comprehensive phenotypic data across many species could be an important tool for medical research.

Identifying constrained regions in the human genome

Probably the most common use of comparative sequence data is assessing conservation. The increased sequencing and genotyping capacities have led to an explosion of genetic variants that are associated with human diseases. However, genetic information alone is often not sufficient to identify single variants that are causally related to disease. Hence, additional information, i.e., different prior probabilities for causality, is required. Currently, the most informative single source is evolutionary conservation [9]. A sequence can be considered as conserved and hence functional if less nucleotide substitutions are observed among species than

W. Enard (✉)
Max Planck Institute for Evolutionary Anthropology,
Deutscher Platz 6,
04103 Leipzig, Germany
e-mail: enard@eva.mpg.de

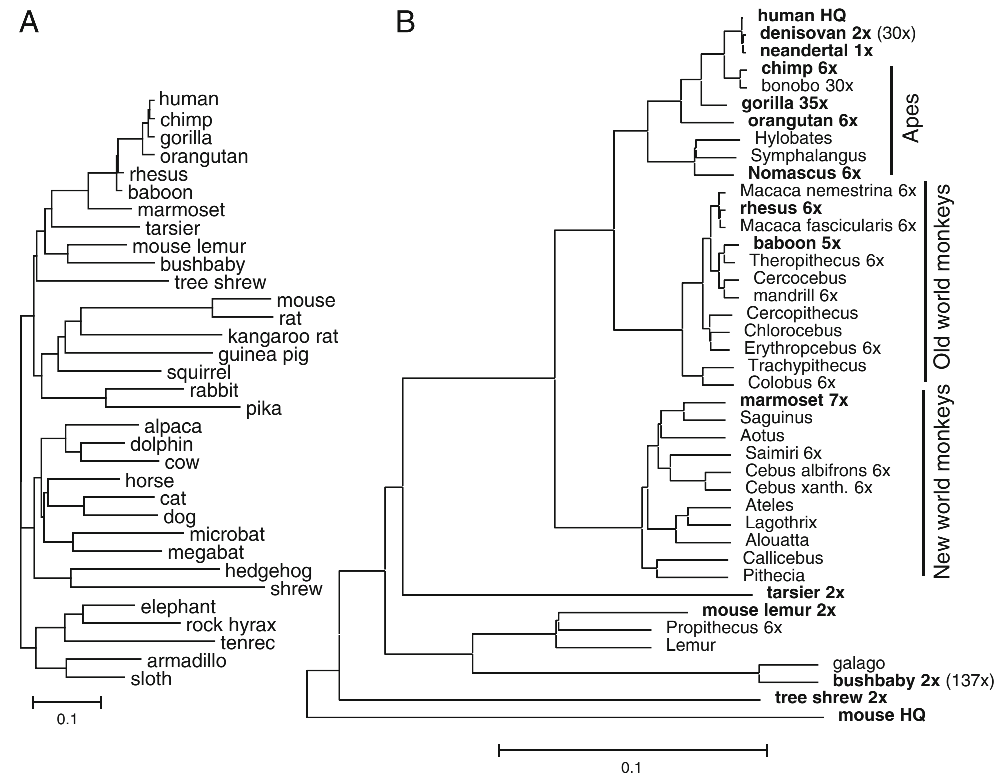

Fig. 1 Mammalian and primate genomes. **a** Phylogeny of mammalian genome sequences that are currently used to estimate conservation (genome.ucsc.edu). Note that most are assemblies built from twofold (2×) sequence coverage and contain many gaps. Scale is 0.1 substitutions at fourfold degenerate sites as given in genome.ucsc.edu. **b** Phylogeny of primates according to Perelman et al. [25]. From the 186 species, I chose those for which genomic sequence is currently available from www.ensembl.org (in *bold* with sequence coverage), are approved sequencing targets (mainly http://www.genome.gov/10002154, *non-bold* with aimed coverage), or illustrate the available diversity in old world and new world monkeys. Approximate relative branch lengths are taken from Perelman et al. 2011 [25] and Reich et al. 2011 [29] for Denisovan and Neanderthal. *Scale bar* is drawn to match branch lengths in (**a**)

expected because genetic variants that result in fewer offspring do not reach high frequencies in populations and are therefore less likely to be observed as differences among species (see e.g., Hurst for a general introduction [10]). A recent landmark study [11] has analyzed genome sequences of 29 mammals and estimated that at least 5.5 % of all 12-bp windows in the human genome have acquired significantly less nucleotide differences among mammals than expected, i.e., can be considered functional and 76 % of these windows could also be reliably located in the human genome. Especially because the majority of disease associated variants are located outside of protein-coding transcripts, this is crucial information to identify putative causal variants for functional follow-up studies. Disease-associated variants are clearly enriched within conserved regions [11, 12], but it is of course an open question how often this information will be really decisive to understand and eventually treat diseases, especially those that affect humans later in life or depend on human-specific gene environment interactions.

The power to detect conservation depends on the expected number of neutral, i.e., nonfunctional, substitutions across the included species [13]. In the analyzed 29 mammals, one expects 4.5 substitutions per base pair and an increase to 15–25 substitutions per base pair with 100–200 eutherian mammals would allow single-nucleotide resolution [11]. Importantly, these estimates assume that the sequence is constrained across all mammals, but especially functional non-coding sequence might have a high turnover rate [14]. Hence, a proportion of the functional elements would be restricted to particular lineages, which could add another 5 % of the genome that would be functional in any particular species and not be detected currently [15]. Hence, comparative primate genomics will be of particular importance for annotating the human genome. If genomes of most primates were available (see Box 1 and Fig. 1), this would sum up to an expected 1.5 substitutions per base pair, which will have a lower resolution of detecting constrained elements, but is without alternative for primate-specific functional elements. That functional elements can be identified just from primate sequences have already been shown [16–19], and the need to do this is most obvious for genes that only exist in humans and primates.

Box 1 Primate genomics

Currently the genome sequences of 13 non-human primates are available and at least 11 are approved sequencing targets (Figure 1). These have and further will reveal basic insights into evolutionary processes of mutation, selection and recombination [20], will be essential tools for primate model organisms [21], allow new types of studies in primatology [22] and – as pointed out in this review – will also be directly informative for medically relevant questions. It is crucial to realize that the quality of these genomes range from a finished state - currently only available for human and mouse -, over draft assemblies that have usually a 6-7x coverage (i.e. each base of the genome is covered 6-7 fold on average) to low quality 2x assemblies. Although this will certainly improve considerably in the nearer future when sequencing costs drop further, analyses need to be aware of different qualities of genome data as well as genome assemblies. This is especially critical for regions of the genome that duplicated recently [23] and when interpreting outliers such as positively selected genes [24]. Primates have been classified into 261–377 species and Perelman et al. have recently published the most comprehensive analysis of primate phylogeny to date, including 186 species and ~90% of primate diversity [25]. Eventually genome sequences will be available for most of these species whereas targeted approaches using exome sequencing might be a cost-effective mid-term solution [26]. When closely related species, e.g. the chimpanzee and bonobo, are included in the analysis it will be important to develop models that take the influence of the ancestral population size into account to make full use of the available information [27]. Exciting and special cases are genome sequences generated from fossils such as the 1.3x coverage from three Neandertal individuals [28] and the 1.9x coverage from a small finger bone found in the Denisova cave in Siberia [29]. These genomes are on average slightly more related to each other than to modern human genomes, but most genomic regions still fall within the variation of modern humans [29]. Interestingly, those regions where this is not the case, i.e. where all modern humans are closer related to each other than to Denisovans or Neandertals, are enriched for regions that have been positively selected after the population split some 270,000–440,000 years ago [28]. So in addition to insights in human population history [30], these ancient genomes provide a unique source of information for inferring selection in humans, which can have medical relevance (see below). For example, it could be shown that a particular HLA allele (HLA-B*73) introgressed from Denisovans and spread to high frequency in Eurasia [31]. Further data and the identification of additional fossils will lead to considerably better assemblies of these ancient genomes and 30x coverage data for Denisovans was recently made available (http://www.eva.mpg.de/denisova). Although it is unlikely that endogenous DNA sequences can be obtained from much older hominin fossils, the unexpected finding of Denisovans allows optimism that genomes from more hominins can be discovered and will improve our understanding of human evolution and even some aspects of human disease.

LPA as an example for a primate-specific gene

Although it is clear that the gene repertoire is fairly similar across mammals, there is still a significant proportion of genes that are specific to particular lineages [32]. It has been estimated that ~9 % of all human genes arose after the split from mouse [33]. Evolutionary conservation of many of those genes clearly suggests that they do have a function, and many seem to be involved in testis development [32, 33] and primate brain development [34]. One medically relevant example of a primate-specific gene is *LPA*, the gene encoding the defining component of the lipoprotein a (Lp(a)) [35]. Genetic variants in *LPA* that increase Lp(a) levels increase the risk of coronary disease in humans [36]. Lp(a) is restricted to old world monkeys, apes, and humans [37, 38] because *LPA* arose as a gene duplication of plasminogen in the common ancestor of old world monkeys and new world monkeys [37, 39, 40]. Curiously, a gene analogous to *LPA* evolved independently by a gene duplication from plasminogen in hedgehogs [37, 39, 40]. By comparing 18 sequences from old world monkeys, Boffelli et al. [16] identified and verified regulatory elements in the *LPA* promoter, serving as a proof-of-principle that identifying conserved regulatory elements in a restricted set of primates is possible.

Despite its medical relevance and research on Lp(a) for almost 50 years, relatively little is known about its physiological function [35], also because neither humans nor baboons without Lp(a) have any apparent phenotype [41].

How can one be certain that there is any function? An easy and powerful approach is to estimate evolutionary constraints on reading frame disruptions. Especially insertions and deletions lead easily to reading frame disruptions and are observed genome-wide at a rate of 1 in 0.5 billion base pairs per generation or approximately at a tenth of the nucleotide substitution rate [42]. The open reading frame of *LPA* is 13,644 bp, and one would expect that on average ~16 indels would accumulate between a human and a chimpanzee that are separated by 12 million years or ~0.6 million generations. Hence, the chance to observe no indel can be calculated using a binominal distribution and is 8×10^{-8}. The chance to see no frameshift between a human and a baboon that are separated by over 50 million years [25] is essentially zero. Hence, for the case of *LPA*, it is clear that some physiological function that conserved the open reading frame must exist. For genes that are more restricted to particular lineages, it will be important to develop more precise models that especially take into account context-dependent indel rates among primates. It is important to keep in mind that a physiological function just needs to ensure that individuals with LPA have on average more offspring than individuals without LPA. How many more offspring are needed to overcome chance effects depends on the effective number of individuals (Ne), respectively chromosomes (2Ne). Chance (also called genetic drift) and selection are equally strong when the selective advantage is 1/2Ne. So selection dominates when the advantage is >>1/2Ne (one can think of it as how often one needs to toss a coin to measure a bias towards one side). Effective population size estimates from current variation ranges e.g. between ~10,000 and 30,000 in human and chimpanzee populations [43]. So as a very rough estimate, physiological functions are conserved in primates when they ensure on average considerably more than 0.0017–0.005 % more offspring (see e.g., Hurst [10] for an introduction).

Hence, it is no discrepancy that humans or baboons without Lp(a) [41] have no apparent phenotype since the evolutionary advantage of possessing Lp(a) could be generally small or just matter under particular environmental conditions. With this evolutionary background in mind, it might be worth to reinvestigate human null alleles. It would also be informative whether *LPA* is present in all old world monkeys and whether one could find correlations with its expression levels with environmental variables in different species, such as pathogen load or diet. In this respect, it is remarkable that a gene analogous to LPA evolved independently by a gene duplication from plasminogen in hedgehogs [37, 39, 40]. A priori a good hypothesis for a physiological function in genes that change relatively rapidly across mammals is an involvement in the immune system. Genes annotated in the immune system show evidence for positive selection more often than most other categories [11], and pathogenic environment had a larger impact on genetic differences among human populations than diet regimes or climatic conditions [44]. Since evidence is accumulating that lipoproteins in general are important components of the immune system [45] and apo(a) regulates neutrophil recruitment [46], a physiological function of Lp(a) in this context is certainly well compatible with the evolution of Lp(a). Alternative explanations, such as a role of Lp(a) in wound healing [47], would need to explain why Lp(a) evolved independently in old world monkeys and hedgehogs and is absent in other mammals and primates.

In summary, comparative primate genomics is essential to annotate primate-specific genes, and LPA is a medically relevant example of such a gene. As also argued below, additional information on genotype–phenotype correlations across primates might be informative for understanding the physiological functions of such genes.

APOE as an example for compensatory mutations

Whether a particular genetic variant causes a disease can depend strongly on other sites in the genome, a phenomenon called genetic interaction or epistasis (see e.g., [48] for a recent review on molecular mechanisms). This can be medically very relevant if for example a disease mutation in humans is not causing disease phenotypes in a model organism such as the mouse. Remarkably, this seems to occur frequently: Analyzing vertebrate orthologues of 32 proteins with well-known disease mutations in humans, Kondrashov et al. [49] estimated that 10 % of all amino acid substitutions observed in vertebrates would be pathogenic in humans. Another way of describing this is that nucleotide substitutions that are known to be pathogenic in humans are just five times less likely to be observed in other species than substitutions for which no pathogenic association is known. Genome-wide studies, e.g., using the chimpanzee genome [50], the rhesus genome [51], or the Neanderthal genome [52], have confirmed such a high rate. The reason for the vast majority of the cases is probably that one or several other substitutions—often in the same protein—compensate the effect [53]. Apolipoprotein E (APOE) is a medically relevant example of this phenomenon

Apolipoprotein E is a ligand for lipoprotein receptors and important for lipid metabolism and transport (see e.g., [54] for a recent review). Three isoforms (APOE2, APOE3, and APOE4) are frequent in humans and have been associated with a risk for cardiovascular disease and especially late-onset Alzheimer's disease, for which APOE4 is a stronger risk predictor than any other common variant [55]. The isoforms derive from two polymorphisms that change the amino acids of the processed APOE at position 112 and 158 (see Fig. 2). Interestingly, the two cysteines defining APOE4 at these two positions are the ancestral state, present

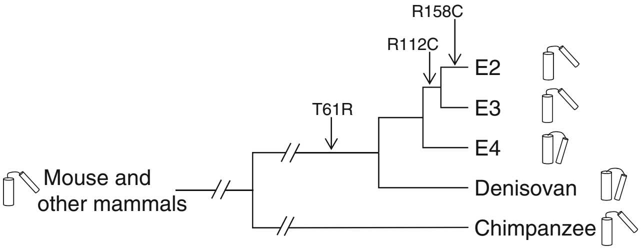

Fig. 2 Evolution of domain interaction in APOE. The substitution of threonine to arginine at position 61 occurred after the split of humans and chimpanzees some 6 million years ago and before the split of modern humans and Denisovans some 800,000 years ago [29]. This enables an interaction of the N-terminal and C-terminal domain of APOE which gets disrupted by the change from arginine to cysteine at position 112. Note that there are additional amino acid substitutions on the tree (e.g., four more on the human lineage and three on the chimpanzee lineage). The Denisovan sequence at the three depicted positions was inferred from the available sequence at genome.ucsc.edu

in chimpanzees, all other primates, and most mammals (http://genome.ucsc.edu/). Without any further information, this would imply that most mammals would have an APOE with properties similar to human APOE4. However, it turns out that at a structural and functional level most mammals including the mouse are more like APOE3 (reviewed in [56]). The reason is that human APOE has an arginine residue at position 61, whereas chimpanzees, all other primates, and most mammals have a threonine at this position. Without this arginine residue, there is no interaction of the N-terminal and C-terminal domain of APOE4, and knock-in mice in which residue 61 is “humanized,” i.e., changed to arginine, suggest that this domain interaction might be responsible for most of the APOE4-associated neuropathology [55, 57, 58]. Disrupting this domain interaction pharmaceutically is a promising route to treat Alzheimer’s disease [59]. It would be interesting to investigate more systematically whether the different mammalian APOEs indeed show no domain interaction, i.e., if this is a conserved feature of APOE structure. More generally, the example shows that a more systematic evolutionary approach might lead faster to appropriate mouse models and structure–function relationships also for other disease proteins.

This functional insight has also consequences for the evolutionary interpretations of the existing APOE isoforms (e.g., [60]). These have mainly tried to explain what selective advantage could have driven the spread of APOE3 and APOE2. But if the scenario described above is correct, the central question is what drove the Arg61 variant to fixation on the human lineage, i.e., why did the APOE4 allele arise. Given the strong conservation at this position and the variety of mostly negative effects associated with the APOE4 allele, it seems likely that this change had some negative consequences. Such slightly deleterious mutations can nevertheless get fixed by chance, especially when selection is weaker in small populations (see e.g., [61]). Fixation could also occur due to positive selection on sites in linkage disequilibrium [62]. Just 20 kbp upstream of APOE is the start of the gene poliovirus receptor-related 2 (PVRL2) that experienced strong positive selection throughout mammalian evolution [63], potentially related to its role as viral receptor. The two amino acid changes leading to APOE3 and APOE2 alleles could then be viewed as compensatory mutations, and sequence data [64] as well as simulation data [65] are compatible with such a scenario. Since compensatory mutations are frequent (see above), this scenario can be regarded as an appropriate null hypothesis for explaining the existence of APOE alleles. The alternative is that the negative impact of the T61R change was outweighed by some advantage that could be related to the immune system, reproduction, or cognitive functions (see Trotter et al. [66] for a well-balanced recent review). Plausible explanations would need to take into account that the threonine at position 61 is well conserved across mammals and has apparently not been selected in other lineages. It has been claimed that APOE has a higher rate of protein evolution on the human lineage due to positive selection using APOE sequences from human, mouse, rat, chimpanzee, and dog [67]. However, when using the same method with the now available sequences from human, chimpanzee, orangutan, macaque, baboon, marmoset, rat, mouse, dog, and cow, this result does not hold up (W. Enard, unpublished observation). So I think that one cannot currently reject the null hypothesis that APOE4 got initially fixed in humans by chance or due to linkage to a selected variant in PVRL2 and that APOE3 and then APOE2 rose to high frequency to compensate for this slightly deleterious change. Obviously, it will be important to use mice models and human data to further explore functional differences among these alleles.

Identifying positively selected regions in the human genome

The power of comparative genomics lies in detecting constraints because it uses information from multiple lineages in which the function of the analyzed genomic element is conserved. However, a genomic element could also acquire a new or altered function on one or a few lineages, either due to chance or because the new function was adaptive on these lineages. Understanding these processes is of course highly relevant to understand evolution and human evolution in particular. One possible example is the transcription factor FOXP2, in which two amino acid changes occurred during human evolution that could have been relevant for adapting particular brain circuits to speech and language (reviewed in [68]). Recent adaptations that are caused by genetic variants that are still polymorphic in humans are medically even more relevant. Variants in hemoglobin causing sickle cell anemia and malaria resistance are a classical example, others are skin pigmentation, lactase tolerance, or adaptations to low oxygen at high altitudes [69]. A positively selected variant leaves more offspring than a neutral variant, and this can lead to a “selective sweep” signature in the linked genomic region. This signature is erased over time and in humans gets difficult to detect after roughly 10,000 generations (~250,000 years) [70]. If for a particular gene or genomic element such adaptive events occurred often enough, one can detect positive selection also by comparing different species, in the case of protein-coding genes, by an elevated rate of non-synonymous substitutions that change the encoded amino acid (often called Ka or dN) versus the rate of synonymous substitutions that do not change it (Ks or dS). These two principal ways to detect positive selection (reviewed e.g. in [71–73]) have been applied genome-wide and have e.g. identified more than 2,000 genes as potentially selected during recent human evolution [74]. Across species, the recent analysis of 29 mammals [11] is the most comprehensive analysis to date and finds for 84 % of the 6.05 million codons in 12,871 gene trees evidence of strong purifying selection (dN/dS<0.5) and for 2.4 % of codons evidence for positive selection (dN/dS>1.5).

It is beyond the scope of this article to review these approaches in detail, especially since its impact on medical genomics has just recently been discussed [7]. However, I would like to make a few, rather cautionary, remarks:

Firstly, the false positive rate and false negative rate of scans for selective sweeps are probably high [75]. This is mainly due to the inherently stochastic nature of how individuals are related for a particular genomic region (reviewed e.g. by [76]). Hence, sequencing more human genomes will help, but will not help much. However, what does help is obtaining genomic information from extinct humans such as Denisovans and Neanderthals (see Box 1). Secondly, the signature of a selective sweep and background selection, i.e., the removal of haplotypes from the population because they are linked to deleterious variants, are in many respects similar and difficult to disentangle [77–79]. Hence, negative selection rather than positive selection could be responsible for many cases of selective sweep candidates. Thirdly, for comparisons across species, i.e., when detecting repeated positive selection in a genomic element, the power is good if positive selection occurs in many species. Many immune related genes that interact directly with pathogens fall in this category, but also unexpected categories show a strong signal such as meiotic chromosome segregation [11]. An example of medical relevance is the antiretroviral factor TRIM5α, which shows a strong signature of selection on several primate lineages including humans [80]. Whereas the human ortholog restricts replication of an extinct retrovirus [81], the rhesus ortholog restricts HIV-1 [82]. In contrast, identifying positive selection that is specific to a particular lineage and could be linked to species-specific adaptations is much more difficult. It would be very helpful for interpreting differences in selection across lineages, if one could correlate them with phenotypes (Fig. 3), as pointed out in the next section. Finally, it is important to keep in mind that adaptations can lead to signatures of positive selection, but not all or maybe even only a small minority of signatures of positive selection are adaptations to an ecological niche [83]. As laid out in the previous section, differences between species that compensate slightly deleterious variants seem rather frequent. Although this is less likely for strong selective sweeps, which are probably rare [84], it could make up a substantial proportion of substitutions fixed by positive selection, which in the case of humans could be for example 10 % of all amino acid substitutions [85].

The main consequence from the issues pointed out above is that additional functional information needs to be added since genetic information alone is not sufficient to reliably exclude false positives except in the most extreme cases [69, 86]. Furthermore, biological information is required to identify affected functions and potentially selected traits. If the selected variants under question are still segregating in the human population, then many possibilities exist to test selective scenarios for example in large human cohorts (see e.g. [87] for such an approach). However, if the variants are fixed in humans, it is much more difficult to investigate the phenotypic consequences, although the case of a mouse model for studying the human-specific effects of FOXP2 might allow for careful optimism [68, 88]. Another way of putting it is that patterns of positive selection can be medically informative, especially if combined with functional assays.

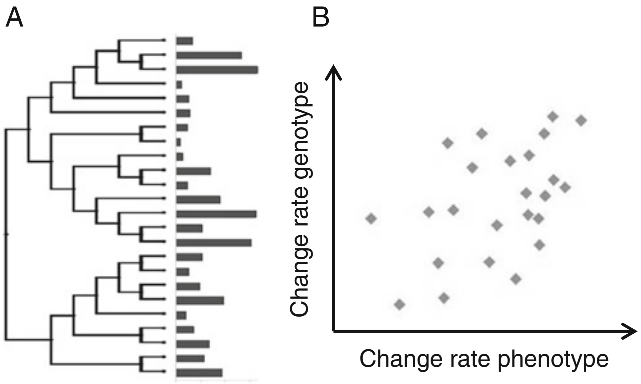

Fig. 3 Correlating changes in phenotypes and changes in genotype across species. **a** A primate phylogeny and a putative trait such as relative testis size. **b** A putative correlation of a measure of phenotype change and genotype change (e.g., Ka/Ks as a measure of protein evolution). Note that such a correlation has to take into account that measures are correlated due to the phylogeny as well as other deadly sins of comparative analysis [108]

The perspective of evolutionary systems biology in primates

On the one hand, sequencing and genotyping technologies have already or will soon improve to an extent that they are not longer the major limiting factor. On the other hand, the link between genetic variation and human disease is much more complex than initially hoped. Hence, the next hope and next challenge in biomedicine is to collect and integrate phenotypic data, in particular molecular phenotypes such as gene expression, which can be measured with high throughput [89, 90]. This approach, whether called systems biology, functional genomics, or biology, should profit from comparative data in a similar way as has the analysis on the DNA level. This approach works in yeasts (see e.g., [91] for a recent review or [92] for a recent analysis in fission yeasts) and starts to be applied in mammals. One example is the modeling of sequence differences, expression, and transcription factor binding in preimplantation development using human, mouse, and cow stem cells that allowed the identification of conserved and species-specific regulatory networks in these species [93]. Extending this approach to primates and to a variety of phenotypes, especially those of medical relevance, should be a worthwhile endeavor:

At a relatively simple level, it will be interesting to see how measures of protein evolution or positive selection on primate lineages correlate with phenotypic changes on these lineages (Fig. 3). Such correlations have so far been shown only for individual genes. For example, the rate of protein evolution for *CDK5RAP2* and *ASPM*, two genes associated with primary microcephaly, was found to correlate with neonatal brain size in primates [94]. Other examples include a faster evolutionary rate of SEMG2 [95, 96] or immunity genes [97] in more promiscuous primate species. It will be interesting to see how often such correlations are found genome-wide, also because it is a unique way of obtaining functional information for genes. For many species, especially for the well-studied primates, a lot of such phenotypic information is already available. In the light of the coming genomic data, it will be valuable to collect additional, well comparable data across as many primate species as possible. In addition to ecologically relevant parameters such as mating systems, pathogens, or diet, it would be worthwhile to collect phenotypes of more direct medical relevance. Imagine, for example, one could measure Lp(a) levels across a range of primates and correlate this with environmental and genetic changes across the phylogeny. This might reveal crucial information about the still unknown physiological functions of this lipoprotein (see above). A good entry point for such comparative data might be human cohort studies that measure e.g. a large range of blood parameters for medical reasons (e.g., [98]).

A very powerful phenotyping method is assessing genome-wide expression patterns, especially since high-throughput sequencing allows to simultaneously assess transcript structure and expression levels as recently applied for six organs across nine mammals [99]. For primates, especially humans, chimpanzees and rhesus macaques have been compared, and the field has matured from a few samples 10 years ago [100] to studies that integrate metabolic, miRNA, and proteomic data across postnatal development in dozens of samples [101, 102]. For example, a recent analysis has revealed that synaptic development is extended in human childhood compared to chimpanzees and macaques, specifically in the prefrontal cortex [103]. If one could extend such approaches to more species, more tissues, and more developmental periods, one could expect a tremendous insight into human biology and disease by identifying constraints and flexibility in such developmental systems. Unfortunately, the availability of suitable tissues is a huge limiting factor, in particular for developmental stages. This limitation is akin to the limited access to tissue samples of human patients. Overcoming this limitation and modeling relevant phenotypes of human diseases in vitro is a major promise of induced pluripotent stem (iPS) cells [104]. It is likely that human protocols can be readily applied to generate iPS cells of many primates [105, 106]. One could imagine generating a panel of human, primate, and mammalian iPS cells to which disease-relevant assays can be applied and variable and conserved phenotypes can be distinguished to interpret disease-related variation (Fig. 4). The prospect

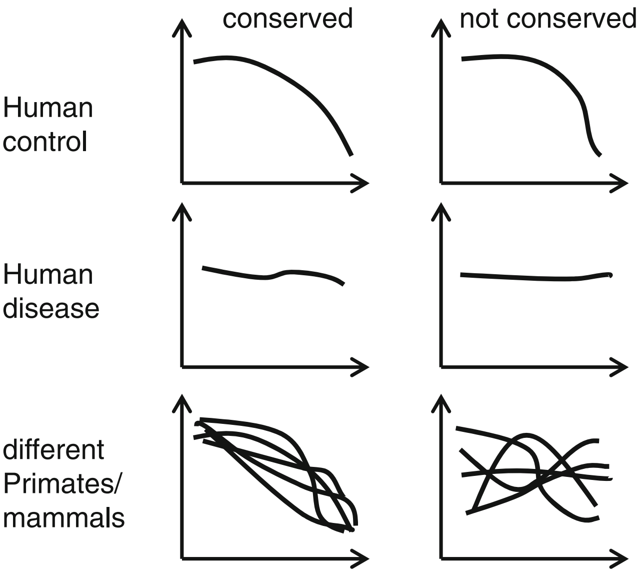

Fig. 4 Illustration how comparative data could help to interpret disease-related phenotypes. Imagine one would measure gene expression levels (*y*-axis) across time (*x*-axis) in differentiating iPS cells from patients and controls and identify genes or groups of genes that differ. To interpret disease-associated changes, one could collect the same data from primate and mammalian iPS cells and distinguish among disease-associated patterns that are conserved and that are more variable, similar to the approach of interpreting disease-associated variants on the DNA level

that this can be combined with targeted genetic modifications in these cell lines using engineered nucleases such as zinc fingers or TALENs [107] could make this a decisive tool in leveraging the potential of comparative data for medical questions.

Conclusions

Biomedicine cannot afford ignoring the unique information that can be obtained from comparative genomic data, especially those from humans' closest relatives, the primates. Identifying constraints, including primate-specific constraints and epistatic constraints, is crucial in order to interpret disease-associated variants and to improve animal models for diseases. Just as functional studies are needed to interpret human genetic variation, functional studies are crucial to interpret evolutionary changes for particular genes. Hence, collecting comprehensive and comparable phenotypic data across many species is a necessary next step. High-throughput methods for molecular phenotypes will be particularly valuable, and iPS cell technology should allow measuring such phenotypes in a comparable way across a large number of species.

Acknowledgments This work was supported by the Max Planck Society.

References

1. Nesse RM, Bergstrom CT, Ellison PT, Flier JS, Gluckman P, Govindaraju DR, Niethammer D, Omenn GS, Perlman RL, Schwartz MD et al (2010) Evolution in health and medicine Sackler colloquium: making evolutionary biology a basic science for medicine. Proc Natl Acad Sci U S A 107(Suppl 1):1800–1807
2. Pennisi E (2011) Evolution. Darwinian medicine's drawn-out dawn. Science 334:1486–1487
3. Little TJ, Allen JE, Babayan SA, Matthews KR, Colegrave N (2012) Harnessing evolutionary biology to combat infectious disease. Nat Med 18:217–220
4. Althouse BM, Bergstrom TC, Bergstrom CT (2010) Evolution in health and medicine Sackler colloquium: a public choice framework for controlling transmissible and evolving diseases. Proc Natl Acad Sci U S A 107(Suppl 1):1696–1701
5. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481:306–313
6. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463
7. Crespi BJ (2011) The emergence of human-evolutionary medical genomics. Evol Appl 4:292–314
8. Genome_10K_Community (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100:659–674
9. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640
10. Hurst LD (2009) Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet 10:83–93
11. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E et al (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482
12. Dudley JT, Chen R, Sanderford M, Butte AJ, Kumar S (2012) Evolutionary meta-analysis of association studies reveals ancient constraints affecting disease marker discovery. Mol Biol Evol. doi:10.1093/molbev/mss079
13. Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3:e10
14. Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. Genome Res 20:1335–1343
15. Ponting CP, Nellaker C, Meader S (2011) Rapid turnover of functional sequence in human and other genomes. Annu Rev Genomics Hum Genet 12:275–299
16. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394
17. Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res 16:855–863
18. Wang QF, Prabhakar S, Wang Q, Moses AM, Chanan S, Brown M, Eisen MB, Cheng JF, Rubin EM, Boffelli D (2006) Primate-specific evolution of an LDLR enhancer. Genome Biol 7:R68
19. Wang QF, Prabhakar S, Chanan S, Cheng JF, Rubin EM, Boffelli D (2007) Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. Genome Biol 8:R1
20. Marques-Bonet T, Ryder OA, Eichler EE (2009) Sequencing primate genomes: what have we learned? Annu Rev Genom Hum Genet 10:355–386
21. Sasaki E, Suemizu H, Shimada A, Hanazawa K, Oiwa R, Kamioka M, Tomioka I, Sotomaru Y, Hirakawa R, Eto T et al (2009) Generation of transgenic non-human primates with germline transmission. Nature 459:523–527
22. Bradley BJ, Lawler RR (2011) Linking genotypes, phenotypes, and fitness in wild primate populations. Evol Anthropol 20:104–119
23. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65
24. Mallick S, Gnerre S, Muller P, Reich D (2009) The difficulty of avoiding false positives in genome scans for natural selection. Genome Res 19:922–933
25. Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y et al (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342
26. George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH (2011) Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res 21:1686–1694
27. Siepel A (2009) Phylogenomics of primates and their ancestral populations. Genome Res 19:1929–1941
28. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH et al (2010) A draft sequence of the Neanderthal genome. Science 328:710–722
29. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060
30. Stoneking M, Krause J (2011) Learning about human population history from ancient and modern genomes. Nat Rev Genet 12:603–614

31. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA et al (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334:89–94
32. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–1326
33. Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M (2010) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol 8
34. Zhang YE, Landback P, Vibranovski MD, Long M (2011) Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 9:e1001179
35. Kronenberg F, Utermann G (2012) Lipoprotein(a): reloaded. Curr Cardiovasc Risk Rep 6:12–20
36. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, Parish S, Barlera S, Franzosi MG, Rust S et al (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N Engl J Med 361:2518–2528
37. Lawn RM, Schwartz K, Patthy L (1997) Convergent evolution of apolipoprotein(a) in primates and hedgehog. Proc Natl Acad Sci U S A 94:11992–11997
38. Makino K, Scanu AM (1991) Lipoprotein(a): nonhuman primate models. Lipids 26:679–683
39. Boffelli D, Cheng JF, Rubin EM (2004) Convergent evolution in primates and an insectivore. Genomics 83:19–23
40. Lawn RM, Boonmark NW, Schwartz K, Lindahl GE, Wade DP, Byrne CD, Fong KJ, Meer K, Patthy L (1995) The recurring evolution of lipoprotein(a). Insights from cloning of hedgehog apolipoprotein(a). J Biol Chem 270:24004–24009
41. Williamsblangero S, Rainwater DL (1991) Variation in Lp(a) levels and Apo(a) isoform frequencies in 5 baboon subspecies. Hum Biol 63:65–76
42. Dupanloup I, Kaessmann H (2006) Evolutionary simulations to detect functional lineage-specific genes. Bioinformatics 22:1815–1822
43. Fischer A, Prufer K, Good JM, Halbwax M, Wiebe V, Andre C, Atencia R, Mugisha L, Ptak SE, Paabo S (2011) Bonobos fall within the genomic variation of chimpanzees. PLoS One 6:e21605
44. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7:e1002355
45. Han R (2010) Plasma lipoproteins are important components of the immune system. Microbiol Immunol 54:246–253
46. Hoover-Plow J, Hart E, Gong Y, Shchurin A, Schneeman T (2009) A physiological function for apolipoprotein(a): a natural regulator of the inflammatory response. Exp Biol Med (Maywood) 234:28–34
47. Brown MS, Goldstein JL (1987) Plasma lipoproteins: teaching old dogmas new tricks. Nature 330:113–114
48. Lehner B (2011) Molecular mechanisms of epistasis within and between genes. Trends Genet 27:323–331
49. Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99:14878–14883
50. Mikkelsen T, Hillier L, Eichler E, Zody M, Jaffe D, Yang S, Enard W, Hellmann I, Lindblad-Toh K, Altheide T et al (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
51. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234
52. Zhang G, Pei Z, Krawczak M, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN (2010) Triangulation of the human, chimpanzee, and Neanderthal genome sequences identifies potentially compensated mutations. Hum Mutat 31:1286–1293
53. Gao L, Zhang J (2003) Why are some human disease-associated mutations fixed in mice? Trends Genet 19:678–681
54. Hauser PS, Narayanaswami V, Ryan RO (2011) Apolipoprotein E: from lipid transport to neurobiology. Prog Lipid Res 50:62–74
55. Bu G (2009) Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci 10:333–344
56. Mahley RW, Weisgraber KH, Huang Y (2009) Apolipoprotein E: structure determines function, from atherosclerosis to Alzheimer's disease to AIDS. J Lipid Res 50(Suppl):S183–188
57. Zhong N, Ramaswamy G, Weisgraber KH (2009) Apolipoprotein E4 domain interaction induces endoplasmic reticulum stress and impairs astrocyte function. J Biol Chem 284:27273–27280
58. Zhong N, Scearce-Levie K, Ramaswamy G, Weisgraber KH (2008) Apolipoprotein E4 domain interaction: synaptic and cognitive deficits in mice. Alzheimers Dement 4:179–192
59. Chen HK, Ji ZS, Dodson SE, Miranda RD, Rosenblum CI, Reynolds IJ, Freedman SB, Weisgraber KH, Huang Y, Mahley RW (2011) Apolipoprotein E4 domain interaction mediates detrimental effects on mitochondria and is a potential therapeutic target for Alzheimer disease. J Biol Chem 286:5215–5221
60. Finch CE (2010) Evolution in health and medicine Sackler colloquium: evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. Proc Natl Acad Sci U S A 107(Suppl 1):1718–1724
61. Charlesworth J, Eyre-Walker A (2007) The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci U S A 104:16992–16997
62. Chun S, Fay JC (2011) Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet 7:e1002240
63. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4:e1000144
64. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E et al (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881–900
65. Drenos F, Kirkwood TB (2010) Selection on alleles affecting human longevity and late-life disease: the example of apolipoprotein E. PLoS One 5:e10022
66. Trotter JH, Liebl AL, Weeber EJ, Martin LB (2011) Linking ecological immunology and evolutionary medicine: the case for apolipoprotein E. Funct Ecol 25:40–47
67. Vamathevan JJ, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, Kumar V, Word M, Simmons MD, Foord SM et al (2008) The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol Biol 8:273
68. Enard W (2011) FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. Curr Opin Neurobiol 21(3):415–424
69. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208–215
70. Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189
71. Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. Trends Genet 23:568–577

72. Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genome Res 20:291–300
73. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8:857–868
74. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19:711–722
75. Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16:702–712
76. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390
77. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5:e1000471
78. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303
79. Casto AM, Feldman MW (2011) Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? PLoS Genet 7:e1001266
80. Johnson WE, Sawyer SL (2009) Molecular evolution of the antiretroviral TRIM5 gene. Immunogenetics 61:163–176
81. Kaiser SM, Malik HS, Emerman M (2007) Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science 316:1756–1758
82. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J (2004) The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature 427:848–853
83. Mustonen V, Lassig M (2009) From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet 25:111–119
84. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331:920–924
85. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083
86. Enard W, Paabo S (2004) Comparative primate genomics. Annu Rev Genom Hum Genet 5:351–378
87. Timpson N, Heron J, Smith GD, Enard W (2007) Comment on papers by Evans et al. and Mekel-Bobrov et al. on Evidence for positive selection of MCPH1 and ASPM. Science 317:1036, author reply 1036
88. 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
89. Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11:855–866
90. Schadt EE, Bjorkegren JL (2012) NEW: network-enabled wisdom in biology, medicine, and health care. Sci Transl Med 4:115rv111
91. Tirosh I, Barkai N (2011) Inferring regulatory mechanisms from patterns of evolutionary divergence. Mol Syst Biol 7:530
92. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI et al (2011) Comparative functional genomics of the fission yeasts. Science 332:930–936
93. Xie D, Chen CC, He X, Cao X, Zhong S (2011) Towards an evolutionary model of transcription networks. PLoS Comput Biol 7:e1002064
94. Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI (2011) Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. Mol Biol Evol 28:625–638
95. O'Connor TD, Mundy NI (2009) Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics 25:i94–100
96. Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet 36:1326–1329
97. Wlasiuk G, Nachman MW (2010) Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution 64:2204–2220
98. Beutner F, Teupser D, Gielen S, Holdt LM, Scholz M, Boudriot E, Schuler G, Thiery J (2011) Rationale and design of the Leipzig (LIFE) Heart Study: phenotyping and cardiovascular characteristics of patients with coronary artery disease. PLoS One 6: e29070
99. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M et al (2011) The evolution of gene expression levels in mammalian organs. Nature 478:343–348
100. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R et al (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296:340–343
101. Fu X, Giavalisco P, Liu X, Catchpole G, Fu N, Ning ZB, Guo S, Yan Z, Somel M, Paabo S et al (2011) Rapid metabolic evolution in human prefrontal cortex. Proc Natl Acad Sci U S A 108:6181–6186
102. Somel M, Liu X, Tang L, Yan Z, Hu H, Guo S, Jiang X, Zhang X, Xu G, Xie G et al (2011) MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. PLoS Biol 9:e1001214
103. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y et al (2012) Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res 22(4):611–622. doi:10.1101/gr.127324.111
104. Robinton DA, Daley GQ (2012) The promise of induced pluripotent stem cells in research and therapy. Nature 481:295–305
105. Ben-Nun IF, Montague SC, Houck ML, Tran HT, Garitaonandia I, Leonardo TR, Wang YC, Charter SJ, Laurent LC, Ryder OA et al (2011) Induced pluripotent stem cells from highly endangered species. Nat Methods 8:829–831
106. Zhong B, Trobridge GD, Zhang X, Watts KL, Ramakrishnan A, Wohlfahrt M, Adair JE, Kiem HP (2011) Efficient generation of nonhuman primate induced pluripotent stem cells. Stem Cells Dev 20:795–807
107. McMahon MA, Rahdar M, Porteus M (2012) Gene editing: not just for translation anymore. Nat Methods 9:28–31
108. Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22:1367–1375