ARTICLE

Widespread Signals of Convergent Adaptation to High Altitude in Asia and America

Matthieu Foll,[1,2,5,*] Oscar E. Gaggiotti,[3,4] Josephine T. Daub,[1,2] Alexandra Vatsiou,[4] and Laurent Excoffier[1,2]

Living at high altitude is one of the most difficult challenges that humans had to cope with during their evolution. Whereas several genomic studies have revealed some of the genetic bases of adaptations in Tibetan, Andean, and Ethiopian populations, relatively little evidence of convergent evolution to altitude in different continents has accumulated. This lack of evidence can be due to truly different evolutionary responses, but it can also be due to the low power of former studies that have mainly focused on populations from a single geographical region or performed separate analyses on multiple pairs of populations to avoid problems linked to shared histories between some populations. We introduce here a hierarchical Bayesian method to detect local adaptation that can deal with complex demographic histories. Our method can identify selection occurring at different scales, as well as convergent adaptation in different regions. We apply our approach to the analysis of a large SNP data set from low- and high-altitude human populations from America and Asia. The simultaneous analysis of these two geographic areas allows us to identify several candidate genome regions for altitudinal selection, and we show that convergent evolution among continents has been quite common. In addition to identifying several genes and biological processes involved in high-altitude adaptation, we identify two specific biological pathways that could have evolved in both continents to counter toxic effects induced by hypoxia.

Introduction

Distinguishing between neutral and selected molecular variation has been a long-standing interest of population geneticists. This interest was fostered by the publication of Kimura's seminal paper[1] on the neutral theory of molecular evolution. Although the controversy rests mainly on the relative importance of genetic drift and selection as explanatory processes for the observed biodiversity patterns, another important question concerns the prevalent form of natural selection. Kimura[1] argued that the main selective mechanism was negative selection against deleterious mutations. However, an alternative point of view emphasizes the prevalence of positive selection, the mechanism that can lead to local adaptation and eventually to speciation.[2,3]

A powerful approach to uncover positive selection is the study of mechanisms underlying convergent evolution. When different populations or evolutionary lineages are exposed to similar environments, positive selection should indeed lead to similar phenotypic features. Convergent evolution can be achieved through similar genetic changes (sometimes called "parallel evolution") at different levels: the same mutation appearing independently in different populations, the same existing mutation being recruited by selection in different populations, or the involvement of different mutations in the same genes or the same biological pathways in separate populations.[4] However, existing statistical genetic methods are not well adapted to the study of convergent evolution when data sets consist in multiple contrasts of populations living in different environments.[5] The current strategy is to carry out independent genome scans in each geographic region and to look for overlaps between loci or pathways that are identified as outliers in different regions.[6] Furthermore, studies are often split into a series of pairwise analyses that consider sets of populations inhabiting different environments. Whereas this strategy has the advantage of not requiring the modeling of complex demographic histories,[7,8] it often ignores the correlation in gene frequencies between geographical regions when correcting for multiple tests.[9] As a consequence, current approaches are restricted to the comparison of lists of candidate SNPs or genomic regions obtained from multiple pairwise comparisons. This suboptimal approach might also result in a global loss of power as compared to a single global analysis and thus to a possible underestimation of the genome-wide prevalence of convergent adaptation.

One particularly important example where this type of problem arises is in the study of local adaptation to high altitude in humans. Human populations living at high altitude need to cope with one of the most stressful environments in the world, to which they are likely to have developed specific adaptations. The harsh conditions associated with high altitude include not only low oxygen partial pressure, referred to as high-altitude hypoxia, but also other factors like low temperatures, arid climate, high solar radiation, and low soil quality. While some of these stresses can be buffered by behavioral and cultural adjustments,

[1]CMPG, Institute of Ecology and Evolution, University of Berne, Berne 3012, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; [3]School of Biology, Scottish Oceans Institute, University of St Andrews, St Andrews, Fife KY16 8LB, UK; [4]Laboratoire d'Ecologie Alpine (LECA), UMR 5553 CNRS-Université de Grenoble, Grenoble, France
[5]Present address: School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland
*Correspondence: matthieu.foll@epfl.ch
http://dx.doi.org/10.1016/j.ajhg.2014.09.002. ©2014 by The American Society of Human Genetics. All rights reserved.

important physiological changes have been identified in populations living at high altitude (see below). Recently, genomic advances have unveiled the first genetic bases of these physiological changes in Tibetan, Andean, and Ethiopian populations.[10–19] The study of convergent or independent adaptation to altitude is of primary interest,[11,19,20] but this problem has been superficially addressed so far, because most studies focused on populations from a single geographical region.[10,13,14,16–19]

Several candidate genes for adaptation to altitude have nevertheless been clearly identified,[21,22] the most prominent ones being involved in the hypoxia inducible factor (HIF) pathway, which plays a major role in response to hypoxia.[23] In Andeans, *VEGFA* (vascular endothelial growth factor A, MIM 192240), *PRKAA1* (protein kinase, AMP-activated, alpha 1 catalytic subunit, MIM 602739), and *NOS2A* (nitric oxide synthase 2A, MIM 163730) are the best-supported candidates, as well as *EGLN1* (egl-9 family hypoxia-inducible factor 1, MIM 606425), a downregulator of some HIF targets.[12,24] In Tibetans,[10,11,13,14,16,25] the HIF pathway gene *EPAS1* (endothelial PAS domain protein 1, MIM 603349) and *EGLN1* have been repeatedly identified.[22] Recently, three similar studies that focused on Ethiopian highlanders[17–19] suggested the involvement of HIF genes other than those identified in Tibetans and Andeans, with *BHLHE41* (MIM 606200), *THRB* (MIM 190160), *RORA* (MIM 600825), and *ARNT2* (MIM 606036) being the most prominent candidates.

However, there is little overlap in the list of significant genes in these three regions,[18,19] with perhaps the exception of alcohol dehydrogenase genes identified in two out of the three analyses. Another exception is *EGLN1*: a comparative analysis of Tibetan and Andean populations[12] concluded that "the Tibetan and Andean patterns of genetic adaptation are largely distinct from one another," identifying a single gene (*EGLN1*) under convergent evolution, but with both populations exhibiting a distinct dominant haplotype around this gene. This limited convergence does not contradict available physiological data, as Tibetans exhibit some phenotypic traits that are not found in Andeans.[26] For example, Tibetan populations show lower hemoglobin concentration and oxygen saturation than Andean populations at the same altitude.[27] Andeans and Tibetans also differ in their hypoxic ventilatory response, birth weight, and pulmonary hypertension.[28] Finally, *EGLN1* has also been identified as a candidate gene in Kubachians, a high altitude (~2,000 m above sea level) Daghestani population from the Caucasus,[15] as well as in Indians.[29]

Nevertheless, it is still possible that the small number of genes under convergent evolution is due to a lack of power of genome scan methods done on separate pairs of populations. In order to overcome these difficulties, we introduce here a Bayesian genome scan method that (1) extends the F-model[30,31] to the case of a hierarchically subdivided population consisting of several migrant pools, and (2) explicitly includes a convergent selection model. We apply this approach to find genes, genomic regions, and biological pathways that have responded to convergent selection in the Himalayas and in the Andes.

Material and methods

Hierarchical Bayesian Model

One of the most widely used statistics for the comparison of allele frequencies among populations is F_{ST},[32,33] and most studies cited in the introduction used it to compare low- and high-altitude populations within a given geographical region (Tibet, the Andes, or Ethiopia). Several methods have been proposed to detect loci under selection from F_{ST}, and one of the most powerful approaches is based on the F-model (reviewed by Gaggiotti and Foll[34]). However, this approach assumes a simple island model where populations exchange individuals through a unique pool of migrants. This assumption is strongly violated when dealing with replicated pairs of populations across different regions, which can lead to a high rate of false positives.[35]

In order to relax the rather unrealistic assumption of a unique and common pool of migrants for all sampled populations, we extended the genome scan method first introduced by Beaumont and Balding[30] and later improved by Foll and Gaggiotti.[31] More precisely, we posit that our data come from G groups (migrant pools or geographic regions), each group g containing J_g populations. We then describe the genetic structure by a F-model that assumes that allele frequencies at locus i in population j from group g, $\boldsymbol{p_{ijg}} = \{p_{ijg1}, p_{ijg2}, \ldots, p_{ijgK_i}\}$ (where K_i is the number of distinct alleles at locus i), follow a Dirichlet distribution parameterized with group-specific allele frequencies $\boldsymbol{p_{ig}} = \{p_{ig1}, p_{ig2}, \ldots, p_{igK_i}\}$ and with F_{SC}^{ijg} coefficients measuring the extent of genetic differentiation of population j relative to group g at locus i. Similarly, at a higher group level, we consider an additional F-model where allele frequencies $\boldsymbol{p_{ig}}$ follow a Dirichlet distribution parameterized with meta-population allele frequencies $\boldsymbol{p_i} = \{p_{i1}, p_{i2}, \ldots, p_{iK_i}\}$ and with F_{CT}^{ig} coefficients measuring the extent of genetic differentiation of group g relative to the meta-population as a whole at locus i. Figure S1 (available online) shows the hierarchical structure of our model in the case of three groups ($G = 3$) and four populations per group ($J_1 = J_2 = J_3 = 4$) and Figure S2 shows the corresponding nonhierarchical F-model for the same number of populations. All the parameters of the hierarchical model can be estimated by further assuming that alleles in each population j are sampled from a multinomial distribution.[36] These assumptions lead to an expression for the probability of observed allele counts $\boldsymbol{a_{ijg}} = \{a_{ijg1}, a_{ijg2}, \ldots, a_{ijgK_i}\}$:

$$\Pr\left(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ijg}}, \boldsymbol{p_{ig}}, \boldsymbol{p_i}, \theta_{ijg}, \phi_{ig}\right) = \Pr\left(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ijg}}\right) \Pr\left(\boldsymbol{p_{ijg}} \mid \boldsymbol{p_{ig}}, \theta_{ijg}\right) \times \Pr\left(\boldsymbol{p_{ig}} \mid \boldsymbol{p_i}, \phi_{ig}\right)$$

where $\Pr(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ijg}})$ is the multinomial likelihood, $\Pr(\boldsymbol{p_{ijg}} \mid \boldsymbol{p_{ig}}, \theta_{ijg})$ and $\Pr(\boldsymbol{p_{ig}} \mid \boldsymbol{p_i}, \phi_{ig})$ are Dirichlet prior distributions, $\theta_{ijg} = 1/F_{SC}^{ijg} - 1$, and $\phi_{ig} = 1/F_{CT}^{ig} - 1$. This expression can be simplified by integrating over $\boldsymbol{p_{ijg}}$ so as to obtain:

$$\Pr\left(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ig}}, \boldsymbol{p_i}, \theta_{ijg}, \phi_{ig}\right) = \Pr\left(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ig}}, \theta_{ijg}\right) \Pr\left(\boldsymbol{p_{ig}} \mid \boldsymbol{p_i}, \phi_{ig}\right)$$

where $\Pr(\boldsymbol{a_{ijg}} \mid \boldsymbol{p_{ig}}, \theta_{ijg})$ is the multinomial-Dirichlet distribution.[34] The likelihood is obtained by multiplying across loci, regions, and population

Download English Version:

https://daneshyari.com/en/article/2811384

Download Persian Version:

https://daneshyari.com/article/2811384

Daneshyari.com