Hum Genet (2012) 131:423–433
DOI 10.1007/s00439-011-1084-8

ORIGINAL INVESTIGATION

High altitude adaptation in Daghestani populations from the Caucasus

Luca Pagani · Qasim Ayub · Daniel G. MacArthur · Yali Xue · J. Kenneth Baillie · Yuan Chen · Iwanka Kozarewa · Daniel J. Turner · Sergio Tofanelli · Kazima Bulayeva · Kenneth Kidd · Giorgio Paoli · Chris Tyler-Smith

Received: 20 May 2011 / Accepted: 19 August 2011 / Published online: 9 September 2011

Abstract We have surveyed 15 high-altitude adaptation candidate genes for signals of positive selection in North Caucasian highlanders using targeted re-sequencing. A total of 49 unrelated Daghestani from three ethnic groups (Avars, Kubachians, and Laks) living in ancient villages located at around 2,000 m above sea level were chosen as the study population. Caucasian (Adygei living at sea level, $N = 20$) and CEU (CEPH Utah residents with ancestry from northern and western Europe; $N = 20$) were used as controls. Candidate genes were compared with 20 putatively neutral control regions resequenced in the same individuals. The regions of interest were amplified by long-PCR, pooled according to individual, indexed by adding an eight-nucleotide tag, and sequenced using the Illumina GAII platform. 1,066 SNPs were called using false discovery and false negative thresholds of ~6%. The neutral regions provided an empirical null distribution to compare with the candidate genes for signals of selection. Two genes stood out. In Laks, a non-synonymous variant within HIF1A already known to be associated with improvement in oxygen metabolism was rediscovered, and in Kubachians a cluster of 13 SNPs located in a conserved intronic region within EGLN1 showing high population differentiation was found. These variants illustrate both the common pathways of adaptation to high altitude in different populations and features specific to the Daghestani populations, showing how even a mildly hypoxic environment can lead to genetic adaptation.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-011-1084-8) contains supplementary material, which is available to authorized users.

L. Pagani (✉) · Q. Ayub · D. G. MacArthur · Y. Xue · Y. Chen · D. J. Turner · C. Tyler-Smith
The Wellcome Trust Sanger Institute, Hinxton, UK
e-mail: lp8@sanger.ac.uk

L. Pagani
The Leverhulme Centre for Human Evolutionary Studies, Cambridge, UK

D. G. MacArthur
University of Sydney, Sydney, Australia

J. K. Baillie
The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, UK

I. Kozarewa
Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK

S. Tofanelli · G. Paoli
Università di Pisa, Pisa, Italy

K. Bulayeva
Vavilov Institute, Moscow, Russia

K. Kidd
Yale University, New Haven, CT, USA

Introduction

During the past ~100,000 years, humans have expanded from a lowland tropical environment in Africa to occupy an enormous range of environments throughout the world, including extremes of heat, cold and ultraviolet radiation, and have adapted to these by both behavioral and biological responses (Beall 2007). High-altitude environments (regions above 1,500 m according to Ward et al. 2001) are of particular interest in studies of adaptation for two reasons. First, all individuals in a population experience the same stress, which cannot readily be modified by technological or cultural means and, therefore, involves primarily

a physiological response. Second, such environments are present in several different regions, including Africa, Asia and South America, and so provide an opportunity to investigate the extent to which similar selective pressures lead independently to similar adaptations (Beall 2007). At an elevation of 2,000 m, the standard atmospheric equation (West 1996) estimates barometric pressure to be 20% lower than at sea level. Hence in a given inspired volume, the partial pressure of oxygen is reduced accordingly. In the absence of any physiological compensatory response, the alveolar gas equation (Fenn et al. 1946) predicts that arterial blood would have a partial pressure of around 9 kPa, compared to the 13.3 kPa, considered normal in healthy young subjects at sea level (Lumb 2011). Even with normal ventilatory compensation, there is detectable arterial hypoxia at rest in subjects exposed to this altitude (Muhm 2007). This produces a commensurate decline in aerobic exercise capacity of 5–10% (Cerretelli 2008) with consequent reduction of outdoor performances such as running, hunting, or escaping from predators, which can lead to a reduction in fitness if not counterbalanced by acclimatization or genetic adaptation. An organism exposed to such conditions has, therefore, to find a metabolic way to compensate for the reduction of oxygen delivery.

Worldwide, three sets of populations living at high altitudes have received most attention: the Ethiopian highlanders, the Himalayan Sherpa and Tibetans, and the Andean Quechua and Aymaras (Moore 2001; Rupert and Hochachka 2001; Suzuki et al. 2003; Rajput et al. 2006; Zhang et al. 2006; Beall 2007; Erzurum et al. 2007). Physiological studies have shown differences in the mechanisms of adaptation in different regions: in the Andes, e.g., increased oxygen delivery is achieved by an increase in haemoglobin concentration, but in Tibetans the haemoglobin concentration is not increased (Simonson et al. 2010) showing that populations from different parts of the world can adapt to a similar environments following more than one pathway. Recently, the genetic basis of some of these adaptations has been investigated and signs of positive selection found in EPAS1, EGLN1 and PPARA in Tibetan populations (Beall et al. 2010; Bigham et al. 2010; Simonson et al. 2010; Yi et al. 2010; Peng et al. 2011; Xu et al. 2011). Because of the diversity of the hypoxic response, the investigation of additional high-altitude populations would be of considerable interest.

The Caucasian mountains on the border between Asia and Europe reach 5,642 m at their highest point (Mt. Elbrus) and show archeological and genetic evidence of continuous human occupation since >10,000 Ybp (Caciagli et al. 2009), so provide an additional environment where humans have adapted to high altitude. The geomorphological, anthropological, and linguistic landscapes are all very complex. For example, several distinct ethnic groups live at high altitudes in the Republic of Daghestan (Russia) speaking Caucasian, Indo-European or Altaic languages (Caciagli et al. 2009). Anthropological and genetic studies have shown that the majority of these groups have high endogamy, high inbreeding, and small population sizes that have remained stationary for many generations (Bulayeva et al. 2005, 2006). Strict patrilocality and endogamous marriages have led to a reduction of diversity within each population, and the small size has led to genetic drift and differentiation between them, as revealed using Y-chromosomal and mitochondrial DNA markers (Barbujani et al. 1994a, b; Bulayeva et al. 2003; Nasidze 2003; Nasidze et al. 2004; Tofanelli et al. 2009; Balanovsky et al. 2011).

Studies of genetic adaptations to high altitude in Daghestan have been limited, but have revealed some evidence for such adaptations:

- A variant hemoglobin (Hb) alpha subunit sequence was reported in 1987 in some Daghestani individuals (Lacombe et al. 1987). Remarkably, the same position is also variant in some high altitude adaptated deer mice (Storz et al. 2007, 2009, 2010) where in association with other mutations, it contributes to the increase in oxygen affinity of the mouse Hb. However, no data concerning changes in Hb oxygen affinity are available in the literature for the human variant.
- When populations of highlanders moved to the lowlands as a consequence of a Soviet government decision, the mortality rate increased dramatically. Although this increased mortality could be partly explained by novel pathogens encountered in the lowlands, it could not be entirely accounted for in this way (Bulaeva et al. 1995, 1996), so might also reflect a reduction of low-altitude fitness due to genetic adaptation to the high altitude or low genetic diversity in these populations.

Study approach

Three main factors influenced our study design. First, a relatively small number of samples were available (Caciagli et al. 2009) and the physiological parameters of the individual donors were not known. Second, there are likely to be substantial differences between the Daghestani study populations due to genetic drift during their different demographic histories as these have diverged over the last few centuries and millenia, in addition to any adaptive differences linked to high altitude. These two considerations led us to choose an empirical candidate gene approach, in which we would compare neutral regions with genes potentially implicated in high altitude adaptation.

Table 1 Candidate genes analyzed

Gene	Role	Reference	Locus (hg18)	Length (kb)
ACE	The ACE insertion/deletion (I/D) polymorphism has been investigated for association with a phenotypic marker of successful adaptation to hypoxia, pulmonary artery pressure.	(Beall 2007)	17:58,908,166–58,952,935	37.03
EDN1 (ET-1)	Vasoconstrictor produced by endothelial cells in the blood vessels. Of prime importance in high-altitude disorders in sojourners		6: 12,398,599–12,405,399	6.80
EGLN1 (PHD2)	Directly activated by oxygen concentration	(Smith et al. 2008)	1: 229,568,054–229,627,413	59.36
EGLN2 (PHD1)	Directly activated by oxygen concentration	(Smith et al. 2008)	19:45,996,932–46,006,176	9.25
EGLN3 (PHD3)	Directly activated by oxygen concentration	(Smith et al. 2008)	14:33,463,174–33,490,037	26.86
EPO	Stimulates proliferation and differentiation of red blood cell precursor cells and is an antioxidant	(Beall 2007)	7:100,156,359–100,159,257	2.90
EPOR	EPO receptor, limiting factor in EPO physiology		19:11,349,475–11,356,019	6.54
HBA1	60 Lys → Glu variant already found in a Daghestanian population. A mutation in the same position was found in high altitude adapted Deer Mice	(Lacombe et al. 1987; Storz et al. 2007)	16:166,679–167,521	0.84
HBB	Together with alpha and delta is the oxygen carrier in adults		11:5,203,270–5,207,201	1.73
HBD	Together with alpha and beta is the oxygen carrier in adults		11:5,210,484–5,213,176	1.81
HBG1	May play a role in fetal adaptation to high altitude. Possible co-adaptation with the adult hemoglobins in order to maintain the same difference in oxygen affinity		11:5,226,089–5,227,693	1.60
HIF1A	Master gene in oxygen regulation	(Smith et al. 2008)	14:61,231,992–61,284,729	52.74
NOS3	One of three NOS catalyzing the synthesis of nitric oxide. Glu298Asp and eNOS4b/a polymorphisms found in Sherpas populations	(Droma et al. 2006)	7:150,319,080–150,342,608	23.53
VEGFA	Stimulates the development of new blood vessels and increases blood vessel permeability	(Beall 2007)	6: 43,845,899–43,862,202	16.27
VHL	On/off Hif1 activator with potential to alter the expression level of many hypoxia related genes	(Ang et al. 2002a, b; Bushuev et al. 2006)	3: 10,158,319–10,168,744	10.43

A third relevant factor was that our study was initiated in 2008, before the publication of the recent genome-wide surveys, and was, therefore, restricted to candidate genes known in 2008. In particular, at that time there was, in our assessment, not enough evidence to consider PPARA and EPAS1 as strong candidate genes, although EPAS1 has subsequently been associated with high altitude adaptation in multiple populations (Beall et al. 2010; Bigham et al. 2010; Simonson et al. 2010; Yi et al. 2010; Peng et al. 2011; Xu et al. 2011). Thus, we set out to compare 20 autosomal regions accepted as likely to be neutral (Wall et al. 2008) with 15 candidate genes for high altitude adaptation: HIF1A; EGLN1; EGLN2; EGLN3; VHL; EPO; EPOR; VEGFA; EDN1; NOS3; ACE; HBA; HBB; HBD, and HBG1 (Table 1). The three Daghestani highlander populations (Avars $N = 16$, 2,120 m; Laks $N = 21$, 2,100 m; Kubachians $N = 12$, 1,890 m) would be compared with a geographically matched lowlander control population (Adygei, $N = 20$, 17 m). We would search for an excess of differentiation in the candidate genes between the lowland and highland populations (F_{ST}) when compared with the neutral regions as a potential sign of adaptation. To provide calibration standards and quality controls for the resequencing data, we also analyzed 20 CEU individuals from the HapMap3 panel (Frazer et al. 2007), 12 of whom were also included in the 1000 Genomes pilot project (Durbin et al. 2010).

Results

We successfully re-sequenced 15 candidate genes and 20 neutral control regions (480 kb in all) in 89 individuals from three Daghestani high-altitude populations (Avars, Kubachians and Laks) and two control populations, the Adygei and the CEU, with an estimated concordance rate

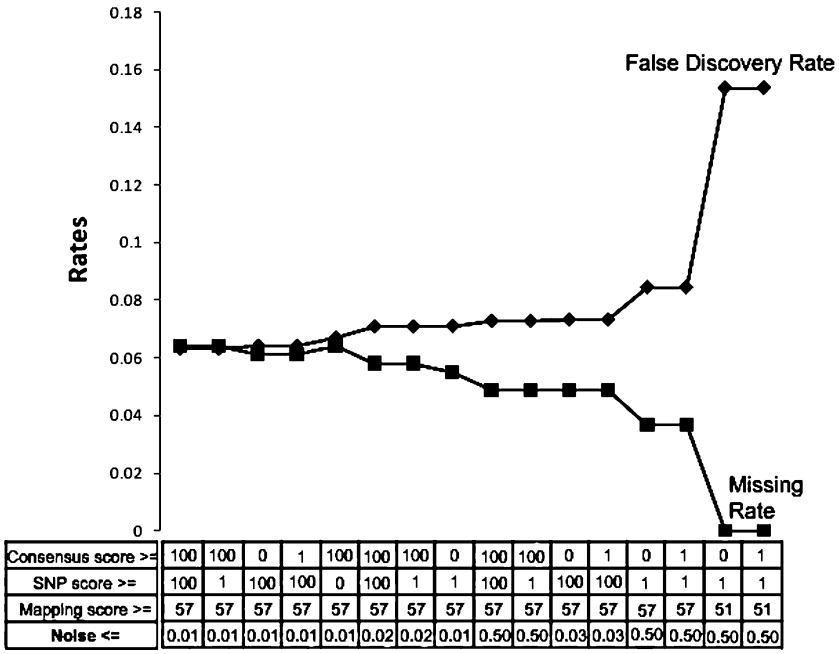

Fig. 1 False discovery and missingness rates using a range of filtering thresholds

with HapMap3 of 99.25%, false discovery rate of 6.6% and false negative rate of 6% (fourth data point in Fig. 1; false discovery and false negative estimates based on comparisons with 1000 Genomes data). These estimates are conservative, since they assume that both the 1000 Genomes and HapMap3 datasets are error-free. We detected 1,066 SNPs (514 in neutral regions and 552 in candidate genes), 316 (207 and 109) novel, which together form the basis for the subsequent analyses. The derived allele frequency spectrum (DAF) shows a large excess of rare alleles (Supporting Fig. 1), as expected in human populations (Durbin et al. 2010) although our stringent filtering process (see "Quality checks" in "Materials and methods") removed 54% of the rare variants when compared with other high depth re-sequencing studies (1000 Genomes exome pilot project, data not shown).

We first compared the Daghestani and control populations with the HapMap samples (20 individuals each), performing a PCA using the HapMap SNPs from the neutral regions. The Daghestani populations cluster together, close to the CEU and distant from the CHB and YRI (Fig. 2). This result is expected from their geographical location, confirming the reliability of the SNP calls and suggesting that drift in the Daghestani populations has not been so high that geographical signatures have been erased.

To investigate the differentiation between highlanders and lowlanders, we calculated pairwise F_{ST} values for each SNP between each Daghestani population and the Adygei (three pairwise comparisons per SNP). The 95% upper boundary calculated from the 20 neutral regions was taken as the empirical significance level for each highlander-control population pair. We considered as outliers those

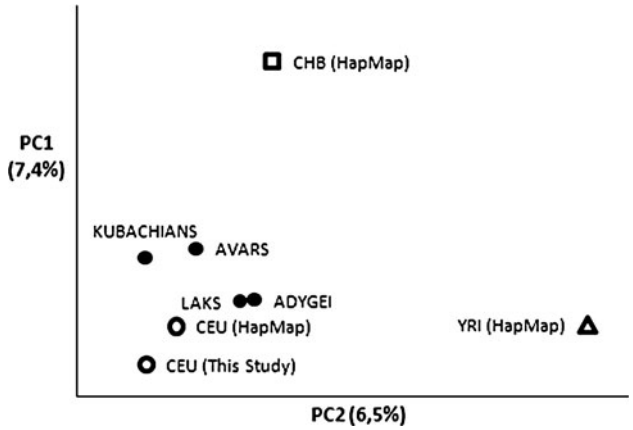

Fig. 2 Principal components analysis (PCA). The Daghestani populations form a sub-cluster within the European/Caucasian cluster

candidate gene SNPs exceeding this value, corrected for multiple tests (three per SNP) according to Bonferroni's formula: significance/number of tests, i.e. SNPs included in the top 1.7% F_{ST} distribution (Supporting Table 2). Averaged over all genes, the proportion of outliers exceeded 1.7% only in Kubachians (Table 2). When looking at each gene individually, we saw an excess of outlier SNPs in the EGLN1, EGLN3 and HIF1A genes in the Kubachians, as well as among the β and δ globin genes in several populations. Due to the high sequence similarity among the globins, the filtering process removed many potential SNPs from these regions. Therefore the high F_{ST} values in the globin genes may be explained by the small numbers of SNPs detected in these genes, and were not considered further. The Daghestani variant of HBA was not

Table 2 Distribution of F_{ST} outlier SNPs in each population compared with Adygei by number (left) or proportion (right)

GENE	SNPs Total	Outlier AVARS	Outlier KUB.	Outlier LAKS	Prop. in AVARS	Prop. in KUB.	Prop. in LAKS
ACE	84	0	0	0	0.00	0.00	0.00
EDN1	25	1	2	0	**0.04**	**0.08**	0.00
EGLN1	121	2	**13**	0	0.02	**0.11**	0.00
EGLN2	11	0	0	0	0.00	0.00	0.00
EGLN3	70	1	6	0	0.01	**0.09**	0.00
EPO	11	0	0	0	0.00	0.00	0.00
EPOR	3	0	0	0	0.00	0.00	0.00
HBA	4	0	0	0	0.00	0.00	0.00
HBB	18	3	0	1	**0.17**	0.00	**0.05**
HBD	9	1	1	0	**0.11**	**0.11**	0.00
HBG1	13	0	0	0	0.00	0.00	0.00
HIF1A	74	3	5	**1**	**0.04**	**0.07**	0.01
NOS3	38	0	0	0	0.00	0.00	0.00
VEGFA	45	0	2	0	0.00	0.04	0.00
VHL	26	0	0	1	0.00	0.00	**0.04**
Total	583	11	29	2	0.02	**0.05**	0.01

Numbers in bold highlight the EGLN1 13-SNP cluster in Kubachians, the HIF1A non-synonymous SNP in Laks, and the genes showing >0.017 proportion of outlier SNPs

detected in this study, but is rare and has only been reported from a few families (Lacombe et al. 1987). However, the EGLN1 region, where Kubachians display a cluster of 13 intronic SNPs with high F_{ST} values (Fig. 3; Supporting Tables 2, 3), cannot be accounted for by any known artefact, and points to an unusually highly differentiated region. We validated all of the SNPs tested and 85% of the genotype calls through capillary sequencing with the remaining 15% of genotypes being incorrect heterozygous calls, in line with the validation rates (70–90%) obtained by the 1000 Genomes Project (Durbin et al. 2010). This intronic region appears to be quite conserved among mammals and is a site of histone H3K4 methylation (Fig. 3) indicative of transcriptional activation (Lupien and Brown 2009). In addition, a single highly differentiated non-synonymous SNP (rs11549465) in HIF1A exon 12 in the Laks also stood out because of its known functional importance, influencing the transactivation response of the HIF dimers and previously associated with oxygen metabolism (Prior et al. 2003; Tanimoto et al. 2003; Doring et al. 2010), and the complete absence of the derived allele (Fig. 4; Supporting Table 3).

To further investigate the EGLN1 and HIF1A regions identified by the F_{ST} results, we calculated six statistics

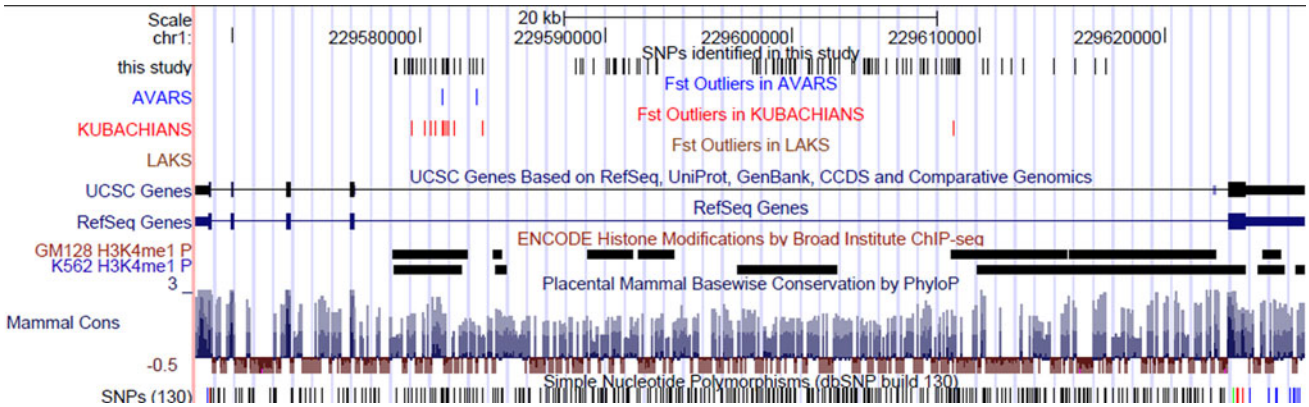

Fig. 3 F_{ST} outlier SNPs in EGLN1 from the three high-altitude populations. The customized UCSC track shows the genomic location of all the SNPs called in this study from the EGLN1 genic region (*black*) and the outlier SNPs for the F_{ST} statistic in Avars (*blue*), Kubachians (*red*) and Laks (*brown*). Below the SNPs is the gene structure, some relevant histone modifications, the conservation of the region and all the SNPs reported in the dbSNP(130) database. The Kubachian 13-SNP cluster is seen in the *left half* of the plot, spanning the 229580000 coordinate

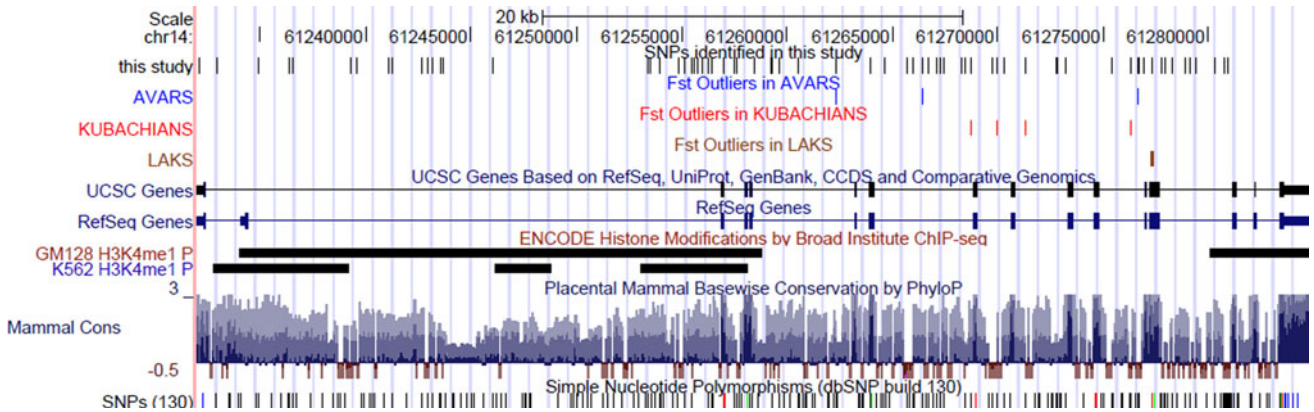

Fig. 4 F_{ST} outlier SNPs in HIF1A from the three high-altitude populations. The customized UCSC track shows the genomic location of all the SNPs called in this study in the HIF1A genic region (*black*) and the outlier SNPs for the F_{ST} statistic in Avars (*blue*), Kubachians (*red*) and Laks (*brown*). Below the SNP position is the gene structure, some relevant histone modifications, the conservation of the region and all the SNPs reported in the dbSNP(130) database. The non-synonymous SNP rs11549465 is identifiable as the only outlier in the Laks track

(Tajima's D, Fu and Li's F and D, Fu's Fs, Fay and Wu's H and π) to compare the evolutionary histories of these genes and the neutral regions. To obtain an empirical threshold for the significance of each test, we used the maximum and minimum scores obtained for the neutral regions in each population. EGLN1 stood out again, since most of the populations gave outlier results at each test although with different trends between the tests. The values, expressed as fold above or below the empirical threshold, are plotted in Fig. 5a. The results reveal an excess of intermediate-frequency variants in these regions, possibly due to the positive selection on an allele that has reached about 50% in the population, or balancing selection. This pattern is confirmed by the network (Bandelt et al. 1999) shown in Fig. 5b, which reveals two distinct clusters of high-frequency haplotypes separated by several steps. Although these neutrality test results only show nominal significance (not corrected for multiple testing), we interpret them as further support for non-neutral evolution of regions identified by the previous F_{ST} results.

Discussion

We have developed a robust approach to resequencing target regions in populations of interest using next-gen sequencing technology involving PCR enrichment and sample multiplexing, and applied it to an investigation of high-altitude adaptation in a little-studied area, Daghestan in the Eastern Great Caucasus. The next-gen sequencing approach proved to be well suited for our study design, since only 154 of the 1,066 SNPs we identified would have been available on standard genotyping arrays (e.g., Illumina 1 M Omni SNPchip). In particular, the highly-differentiated cluster of intronic SNPs in the EGLN1 gene is not represented at all on this array, so would have been overlooked in an array-based approach. In addition to this, the sequencing data allowed us to support the F_{ST} results by calculating a set of statistics that make use of the full allele frequency spectrum, hence reducing the risk of ascertainment bias. The data from the 20 neutral regions show the expected patterns in the PCA plot (Fig. 2) and derived allele frequency spectrum (Supporting Fig. 1), testifying to the reliability of the results from next-gen sequencing technology. A key aspect of our work was the inclusion of these putatively neutral control regions, which allowed us to identify likely selected genes and SNPs using an empirical approach in populations with complex and poorly understood demographies. Our results suggested unusual evolutionary events at two genes, and we now discuss these findings and their biological consequences, along with their implications for a more general understanding of high altitude adaptation.

The derived allele frequency spectrum of the combined candidate gene regions (Supporting Fig. 1) and the neutrality tests applied to the phased haplotypes identified an excess of intermediate-frequency SNPs in several genes. EGLN1 exemplifies these characteristics (Fig. 5). Furthermore, a substantial intronic region of this gene is highly differentiated between the highlander Kubachians and the Adygei living at sea level (Fig. 3). This segment shows high conservation across mammals, comparable to an exonic region, and annotation for the presence of histone methylation (H3K4me1) which indicates transcription enhancer activity (Lupien and Brown 2009). We speculate that one or more of the SNPs in the cluster may alter the

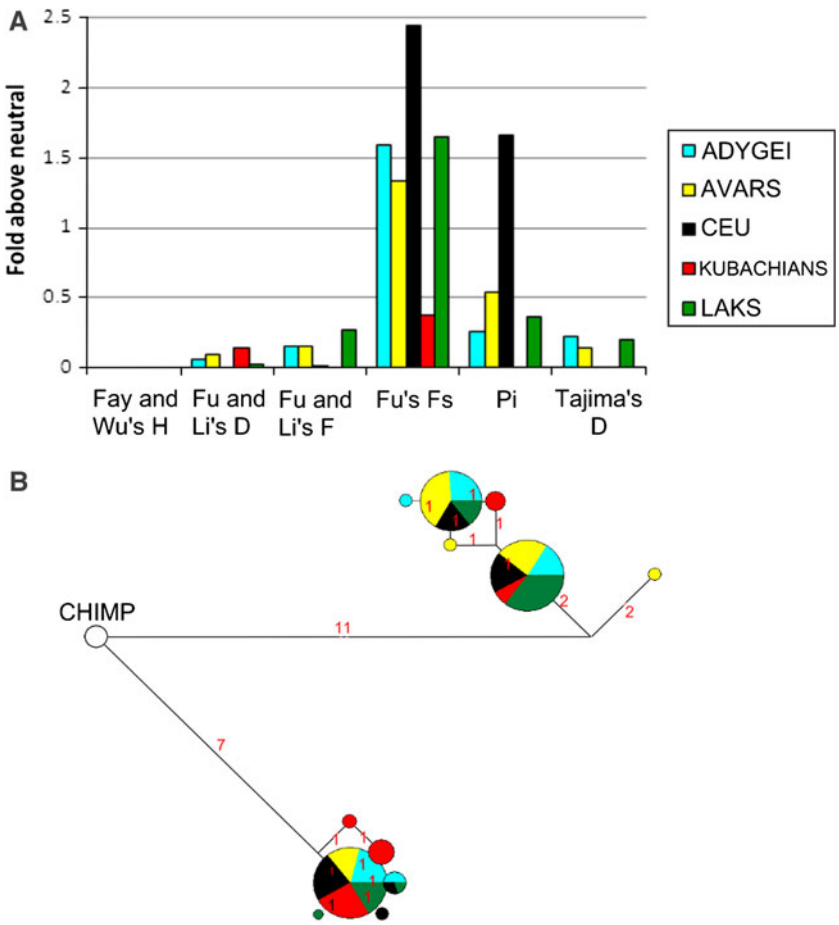

Fig. 5 **a** Neutrality tests for EGLN1. The *y* axis shows the fold above the neutral region maximum, chosen as the empirical threshold for deviation from neutrality. **b** Network of EGLN1 haplotypes. *Circles* represent haplotypes, with area proportional to frequency (smallest = 1) and are coloured according to the population of origin as in A. *Lines* represent mutational steps separating the haplotypes, with the number of steps indicated in *red*

EGLN1 transcription level with a consequent change in the protein production. A decreased availability of EGLN1 (Bernhardt et al. 2010) as well as other modifications of the EGLN-VHL-HIF axis (Smith et al. 2006; Formenti et al. 2011) have been shown to induce upregulation of some genes involved in the HIF-dependent hypoxia response, such as EPO, EGLN1 and VEGF. Future functional studies will be needed to investigate the role of the SNP cluster in regulating gene expression. Interestingly, EGLN1 was picked out as a recently selected gene in Tibetans showing a pattern of high altitude adaptation in three studies (Simonson et al. 2010; Peng et al. 2011; Xu et al. 2011). However, the signature of selection in Tibetans was a long haplotype inferred from SNP array data, centered 20 kb downstream from the transcribed region, and thus distinct from the intronic signature in Kubachians. Another study (Aggarwal et al. 2010) reported two SNPs from the same intron in EGLN1 (rs480902 and rs479200) as markers of putative adaptation to high altitude; however, rs480902 was not detected as polymorphic and rs479200 did not stand out as an outlier in the present study. Although it remains unclear whether or not there might be a common underlying biological mechanism such as a modification of expression, it is indeed striking that the same gene stands out in three independent studies investigating geographically distinct populations, indicating a common path of adaptation to a hypoxic environment.

The other remarkable finding was an outlier SNP in Laks within exon 12 of HIF1A. This SNP, rs11549465, shows the derived allele at a frequency of ~10% in the Adygei, consistent with the Human Genome Diversity Project data (Rosenberg 2006), but is detected only in its ancestral status in Laks, Avars and Kubachians, although partial sequencing failures make conclusions about the latter two populations preliminary. The variant causes a proline to be replaced with a serine in its derived state. As a result, the transactivation of the HIF1A protein dramatically increases during hypoxia (Tanimoto et al. 2003), triggering a downstream hypoxic response. Conversely, the ancestral proline allele was found at significantly higher frequencies in elite endurance athletes when compared with non-athletes (Doring et al. 2010) and increase in VO_{2max} after 1-month training in elderly people (Prior et al. 2003). Hence the absence of the derived allele in the highlander populations can be seen as a twofold consequence of selection: against the damaging excessive

activity of a hypoxia master regulator gene, favouring instead the ancestral variant capable of conferring better endurance and VO_{2max} plasticity in a mildly hypoxic environment.

In conclusion, although a study based on known candidate genes for high altitude adaptation can only reveal the involvement or lack of involvement of these genes, it is striking that in the intronic EGLN1 cluster in Kubachians and the rs11549465 SNP in Laks, we find two distinct patterns of high altitude adaptation previously unknown in Daghestani highlanders. Furthermore, our findings show how even a mildly hypoxic environment (2,000 m) can induce genetic adaptation. These results will benefit from further functional follow-up, but already illustrate both shared aspects between high altitude adaptation in Daghestan and other areas, and features that differentiate Daghestan from other studied regions.

Materials and methods

Samples and ethics statement

The Daghestani DNA samples used in this study were extracted from blood samples collected from healthy adult male donors after obtaining individual written informed consent, and were transported under a research agreement between G. Paoli's group at the Department of Biology of the University of Pisa and K. Bulayeva's group at Vavilov Institute of General Genetics of Russian Academy of Sciences, and have been used in previous studies (Caciagli et al. 2009). The CEU samples were obtained from the Coriell Institute (www.coriell.org), while the Adygei DNAs (extracted from lymphoblastoid cell cultures) were from the Kidd Lab, Department of Genetics, Yale University School of Medicine. Approval for the study was provided by the Cambridgeshire 2 Research Ethics Committee (09/H0308/1).

Candidate genes and their functions

The genes chosen as candidates for high altitude adaptation, together with their main features are reported in Table 1.

Control regions

Twenty genomic regions with no known association with high altitude adaptation and considered to be neutral on the basis of lack of documented function and long distance from known coding regions (Wall et al. 2008) were used as controls. Three sections (1–3 kb each) spanning each of the 20 genomic regions were resequenced. The coordinates and sequences of each trio region (referred to "neutral regions") were kindly provided by Dr. Michael Hammer (http://hammerlab.biosci.arizona.edu) and the primers were designed independently (Supporting Table 1).

Long template PCRs, primer design and standardization

Primers were designed using Primer3 (http://primer3.sourceforge.net/) and BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The design was automated using the software Pfetch and custom Perl scripts available at lp8@sanger.ac.uk . BLAST was used to check for primer specificity. Primer pairs (www.sigma-aldrich.com) were tested and long-PCR (1–6 kb) conditions standardized using a HapMap DNA (NA07029) as a test sample. The 11,904 PCR amplifications were performed in 96-well plates using *Platinum Taq* DNA polymerase high fidelity (Invitrogen), using one plate for each of the 124 primer pairs (protocols available on request). Amplicons were visualized by agarose gel electrophoresis in the presence of ethidium bromide and those from the same individual pooled using volumes inversely proportional to reaction yield, in order to establish approximate equimolarity. The pooled PCR products were subsequently purified through post-PCR columns (Qiagen QIAquick PCR Purification Kit).

Illumina GAII resequencing

All samples were sequenced on an Illumina GAII Genome Analyzer. To exploit the high yield of an Illumina run, an indexed re-sequencing strategy was adopted (Kozarewa and Turner 2011). A unique eight-nucleotide sequence "tag" specific to each individual was added to one of the linkers ligated to each genomic DNA fragment. Tagged DNAs from eight individuals were then pooled, and the resulting mix sequenced using 76 bp paired-end reads. Data from each lane were split according to the individual-specific tag, mapped to the reference sequence (build 36) and SNPs called using Burrows–Wheeler Aligner (BWA) and SAMtools (Li and Durbin 2009; Li et al. 2009).

Validation

Six of the 13 EGLN1 SNPs showing outlier F_{ST} values in Kubachians were validated by a standard PCR/Sanger sequencing reaction on an ABI Genetic Analyzer 3730xl using genomic DNA.. Two pairs of primers were used for sequencing: Pair1 FWD: GCTCTGGTGACAGGAATACT GAA; Pair1 REV: CTGTAGTCCTAGCACTTTGGGAG; pair2 FWD: AAACAGGGATACAAAGCTTAGAGAA; Pair2 REV: AAGTTTCCAAGAACCTATCGAGG.

Quality checks

Raw SNP calls were calibrated by comparing genotype calls made using a range of coverage and allelic ratio thresholds with known genotypes at the 763 HapMap3 sites present in the sequenced regions in the 20 CEU individuals. The maximum concordance of the two dataset was 99.25%, achieved by limiting genotype calls to positions with at least 25× read depth and defining heterozygotes when the allelic ratio lay between 0.25 and 0.75. The reliability of these calibrated calls was then tested by comparing the power to call SNPs in the 12 CEU individuals in the 1000 Genomes Project. Assuming the 1000 Genomes calls to be correct, we applied four additional filters to provide a false discovery rate of 6.6% with a false negative rate to 6%. The filters (thresholds) were goodness of the mapping to the reference sequence (≥57), SAMtools SNP score (≥100), consensus score (≥100) and noise of the signal (≤0.01) (Fig. 1).

Data analyses

Principal components analysis (PCA) was performed using the software Eigensoft (Price et al. 2006) on the 41 neutral region SNPs overlapping with the HapMap3 dataset. The following neutrality tests were applied to identify possible signs of positive selection using an in-house script on PHASEd (Stephens and Donnelly 2003) haplotypes: Fay and Wu's H; Fu and Li's D; Fu and Li's F; Fu's Fs; π; Tajima's D (Tajima 1989; Fu and Li 1993; Fu 1997; Fay and Wu 2000). Some haplotypes were displayed in a median-joining network using Network software (Bandelt et al. 1999). The pairwise F_{ST} between Adygei and each highland population was calculated using the R package Hierfstat (de Meeus and Goudet 2007). The ancestral/derived alleles and functional consequences of the variants were identified from the Ensembl database of genomic annotations (www.ensembl.org).

Acknowledgments We thank all the participants for taking part in this study, Prof. Peter Robbins from the University of Oxford for his advice on high altitude physiology and the Sanger capillary sequencing team for data production. This work was supported by Scuola Normale Superiore of Pisa, Italy; European Union Lifelong Learning Project; University of Pisa, Italy, 60% and BIOCLIMA Project grants to GP and ST respectively and The Wellcome Trust, UK.

References

Aggarwal S, Negi S, Jha P, Singh PK, Stobdan T, Pasha MA, Ghosh S, Agrawal A, Prasher B, Mukerji M (2010) EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. Proc Natl Acad Sci USA 107(44):18961–18966

Ang SO, Chen H, Gordeuk VR, Sergueeva AI, Polyakova LA, Miasnikova GY, Kralovics R, Stockton DW, Prchal JT (2002a) Endemic polycythemia in Russia: mutation in the VHL gene. Blood Cells Mol Dis 28(1):57–62

Ang SO, Chen H, Hirota K, Gordeuk VR, Jelinek J, Guan Y, Liu E, Sergueeva AI, Miasnikova GY, Mole D, Maxwell PH, Stockton DW, Semenza GL, Prchal JT (2002b) Disruption of oxygen homeostasis underlies congenital Chuvash polycythemia. Nat Genet 32(4):614–621

Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E, Haber M, Platt D, Schurr T, Haak W et al (2011) Parallel evolution of genes and languages in the caucasus region. Mol Biol Evo. doi: 10.1093/molbev/msr126 [Epub ahead of print]

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37–48

Barbujani G, Nasidze IS, Whitehead GN (1994a) Genetic diversity in the Caucasus. Hum Biol 66(4):639–668

Barbujani G, Whitehead GN, Bertorelle G, Nasidze IS (1994b) Testing hypotheses on processes of genetic and linguistic change in the Caucasus. Hum Biol 66(5):843–864

Beall C (2007) Detecting natural selection in high-altitude human populations. Respir Physiol Neurobiol 158(2–3):161–171

Beall CM et al (2010) Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci USA 107(25):11459–11464

Bernhardt WM, Wiesener MS, Scigalla P, Chou J, Schmieder RE, Gunzler V, Eckardt KU (2010) Inhibition of prolyl hydroxylases increases erythropoietin production in ESRD. J Am Soc Nephrol 21(12):2151–2156

Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, Lopez Herraez D, Brutsaert T, Parra EJ, Moore LG and Shriver MD (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6(9):e1001116

Bulaeva KB, Kurbatova OL, Pavlova TA, Guseinov GG, Bodia IE, Charukhilova SM, Akhkuev S (1995) Genetic-demographic study of mountain populations from Dagestan and their migrants to the lowlands. Comparison of basic parameters of fitness. Genetika 31(9):1300–1307

Bulaeva KB, Pavlova TA, Charukhilova SM, Bodia IE, Guseinov GG, Akhkuev S (1996) A genetic and demographic study of Dagestan highland populations and migrants to the lowlands. The relationship between levels of consanguinity, homozygosity and physiologic sensitivity. Genetika 32(1):93–102

Bulayeva K, Jorde LB, Ostler C, Watkins S, Bulayev O, Harpending H (2003) Genetics and population history of Caucasus populations. Hum Biol 75(6):837–853

Bulayeva KB, Leal SM, Pavlova TA, Kurbanov RM, Glatt SJ, Bulayev OA, Tsuang MT (2005) Mapping genes of complex psychiatric diseases in Daghestan genetic isolates. Am J Med Genet B Neuropsychiatr Genet 132B(1):76–84

Bulayeva KB, Jorde L, Watkins S, Ostler C, Pavlova TA, Bulayev OA, Tofanelli S, Paoli G, Harpending H (2006) Ethnogenomic diversity of Caucasus, Daghestan. Am J Hum Biol 18(5):610–620

Bushuev VI, Miasnikova GY, Sergueeva AI, Polyakova LA, Okhotin D, Gaskin PR, Debebe Z, Nekhai S, Castro OL, Prchal JT, Gordeuk VR (2006) Endothelin-1, vascular endothelial growth factor and systolic pulmonary artery pressure in patients with Chuvash polycythemia. Haematologica 91(6):744–749

Caciagli L, Bulayeva K, Bulayev O, Bertoncini S, Taglioli L, Pagani L, Paoli G, Tofanelli S (2009) The key role of patrilineal

inheritance in shaping the genetic variation of Dagestan highlanders. J Hum Genet 54(12):689–694
Cerretelli P (2008) Energy sources for muscular exercise. Int J Sports Med 13:S106–S110
de Meeus T, Goudet J (2007) A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. Infect Genet Evol 7(6):731–735
Doring F, Onur S, Fischer A, Boulay MR, Perusse L, Rankinen T, Rauramaa R, Wolfarth B, Bouchard C (2010) A common haplotype and the Pro582Ser polymorphism of the hypoxia-inducible factor-1alpha (HIF1A) gene in elite endurance athletes. J Appl Physiol 108(6):1497–1500
Droma Y, Hanaoka M, Basnyat B, Arjyal A, Neupane P, Pandit A, Sharma D, Miwa N, Ito M, Katsuyama Y, Ota M, Kubo K (2006) Genetic contribution of the endothelial nitric oxide synthase gene to high altitude adaptation in sherpas. High Alt Med Biol 7(3):209–220
Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
Erzurum SC, Ghosh S, Janocha AJ, Xu W, Bauer S, Bryan NS, Tejero J, Hemann C, Hille R, Stuehr DJ, Feelisch M, Beall CM (2007) Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans. Proc Natl Acad Sci 104(45):17593–17598
Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413
Fenn WO, Rahn H, Otis AB (1946) A theoretical study of the composition of the alveolar air at altitude. Am J Physiol 146:637–653
Formenti F et al (2011) Cardiopulmonary function in two human disorders of the hypoxia- inducible factor (HIF) pathway: von Hippel–Lindau disease and HIF-2alpha gain-of- function mutation. FASEB J 25:2001–2011
Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861
Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2):915–925
Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133(3):693–709
Kozarewa I, Turner DJ (2011) 96-plex molecular barcoding for the illumina genome analyzer. Methods Mol Biol 733:279–298
Lacombe C, Soria J, Arous N, Blouquit Y, Bardakdjian J, Riou J, Galacteros F (1987) A new case of Hb Dagestan [alpha 60(E9)Lys–Glu]. Hemoglobin 11(1):39–41
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079
Lumb AB (2011) Nunn's applied respiratory physiology. Butterworth–Heinemann, Philadelphia
Lupien M, Brown M (2009) Cistromics of hormone-dependent cancer. Endocr Relat Cancer 16(2):381–389
Moore LG (2001) Human genetic adaptation to high altitude. High Alt Med Biol 2(2):257–279
Muhm JM (2007) Effect of aircraft-cabin altitude on passenger discomfort. N Engl J Med 357:18–27
Nasidze I (2003) Haplotypes from the Caucasus, Turkey and Iran for nine Y-STR loci. Forensic Sci Int 137(1):85–93
Nasidze I, Ling EYS, Quinque D, Dupanloup I, Cordaux R, Rychkov S, Naumova O, Zhukova O, Sarraf-Zadegan N, Naderi GA, Asgary S, Sardas S, Farhud DD, Sarkisian T, Asadov C, Kerimov A, Stoneking M (2004) Mitochondrial DNA and Y-chromosome Variation in the Caucasus. Ann Hum Genet 68(3):205–221
Peng Y, Yang Z, Zhang H, Cui C, Qi X, Luo X, Tao X, Wu T, Ouzhuluobu, Basang, Ciwangsangbu, Danzengduojie, Chen H, Shi H, Su B (2011) Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Mol Biol Evol 28(2):1075–1081
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909
Prior SJ, Hagberg JM, Phares DA, Brown MD, Fairfull L, Ferrell RE, Roth SM (2003) Sequence variation in hypoxia-inducible factor 1alpha (HIF1A): association with maximal oxygen consumption. Physiol Genomics 15(1):20–26
Rajput C, Najib S, Norboo T, Afrin F, Qadarpasha M (2006) Endothelin-1 gene variants and levels associate with adaptation to hypobaric hypoxia in high-altitude natives? Biochem Biophys Res Commun 341(4):1218–1224
Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70(6):841–847
Rupert JL, Hochachka PW (2001) Genetic approaches to understanding human adaptation to altitude in the Andes. J Exp Biol 204(Pt 18):3151–3160
Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB, Prchal JT, Ge R (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329(5987):72–75
Smith TG et al (2006) Mutation of von Hippel–Lindau tumour suppressor and human cardiopulmonary physiology. PLoS Med 3:e290
Smith TG, Robbins PA, Ratcliffe PJ (2008) The human side of hypoxia-inducible factor. Br J Haematol 141(3):325–334
Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73(5):1162–1169
Storz JF, Sabatino SJ, Hoffmann FG, Gering EJ, Moriyama H, Ferrand N, Monteiro B, Nachman MW (2007) The molecular basis of high-altitude adaptation in deer mice. PLoS Genet 3(3):e45
Storz JF, Runck AM, Sabatino SJ, Kelly JK, Ferrand N, Moriyama H, Weber RE, Fago A (2009) Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. Proc. Natl. Acad. Sci. USA 106(34):14450–14455
Storz JF, Runck AM, Moriyama H, Weber RE, Fago A (2010) Genetic differences in hemoglobin function between highland and lowland deer mice. J Exp Biol 213(Pt 15):2565–2574
Suzuki K, Kizaki T, Hitomi Y, Nukita M, Kimoto K, Miyazawa N, Kobayashi K, Ohnuki Y, Ohno H (2003) Genetic variation in hypoxia-inducible factor 1alpha and its possible association with high altitude adaptation in Sherpas. Med Hypotheses 61(3):385–389
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595
Tanimoto K, Yoshiga K, Eguchi H, Kaneyasu M, Ukon K, Kumazaki T, Oue N, Yasui W, Imai K, Nakachi K, Poellinger L, Nishiyama M (2003) Hypoxia-inducible factor-1alpha polymorphisms associated with enhanced transactivation capacity, implying clinical significance. Carcinogenesis 24(11):1779–1783
Tofanelli S et al (2009) J1-M267 Y lineage marks climate-driven prehistorical human displacements. Eur J Hum Genet 17(11):1520–1524
Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF (2008) A novel DNA sequence database for analyzing human demographic history. Genome Res 18(8):1354–1361

Ward MP, Milledge JS, West JB (2001) High altitude medicine and physiology, Hodder Arnold

West JB (1996) Prediction of barometric pressures at high altitude with the use of model atmospheres. J Appl Physiol 81:1850–1854

Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, Yang L, Pan X, Wang J, Shen Y, Wu B, Wang H, Jin L (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28(2):1003–1011

Yi X et al (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329(5987):75–78

Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC (2006) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22(17):2122–2128