Title

Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders

Permalink
https://escholarship.org/uc/item/4r06227h

Journal
European Journal of Human Genetics, 27(1)

ISSN
1018-4813

Authors
Iranmehr, Arya
Stobdan, Tsering
Zhou, Dan
et al.

Publication Date
2019

DOI
10.1038/s41431-018-0270-8

Peer reviewed

European Journal of Human Genetics (2019) 27:150–159
https://doi.org/10.1038/s41431-018-0270-8

ARTICLE

Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders

Arya Iranmehr[1] · Tsering Stobdan[2] · Dan Zhou[2] · Orit Poulsen[2] · Kingman P. Strohl[3] · Almaz Aldashev[4] · Amalio Telenti[5] · Emily H. M. Wong[6] · Ewen F. Kirkness[6] · J. Craig Venter[6,7] · Vineet Bafna[8] · Gabriel G. Haddad[2,9,10]

Received: 5 April 2018 / Revised: 9 August 2018 / Accepted: 30 August 2018 / Published online: 25 September 2018

Abstract

The Central Asian Kyrgyz highland population provides a unique opportunity to address genetic diversity and understand the genetic mechanisms underlying high-altitude pulmonary hypertension (HAPH). Although a significant fraction of the population is unaffected, there are susceptible individuals who display HAPH in the absence of any lung, cardiac or hematologic disease. We report herein the analysis of the whole-genome sequencing of healthy individuals compared with HAPH patients and other controls (total $n = 33$). Genome scans reveal selection signals in various regions, encompassing multiple genes from the first whole-genome sequences focusing on HAPH. We show here evidence of three candidate genes *MTMR4*, *TMOD3* and *VCAM1* that are functionally associated with well-known molecular and pathophysiological processes and which likely lead to HAPH in this population. These processes are (a) dysfunctional BMP signaling, (b) disrupted tissue repair processes and (c) abnormal endothelial cell function. Whole-genome sequence of well-characterized patients and controls and using multiple statistical tools uncovered novel candidate genes that belong to pathways central to the pathogenesis of HAPH. These studies on high-altitude human populations are pertinent to the understanding of sea level diseases involving hypoxia as a main element of their pathophysiology.

Introduction

Pulmonary hypertension (PH) is a condition with an abnormally high blood pressure in the pulmonary arteries due to arterial resistance to the pulmonary blood flow. This may be due to a variety of causes and combination of factors such as endothelial dysfunction, vasoconstriction of small pulmonary arteries and endothelial and smooth muscle cells proliferation [1]. Clinically, PH is categorized into five groups ranging from idiopathic, familial/heritable, to presenting as a secondary disease, e.g., congenital heart disease with left-to-right shunt, chronic obstructive pulmonary disease, chronic thromboembolic pulmonary disease and chronic exposure to high-altitude (HA) hypoxia or high-altitude PH (HAPH) [1]. Ernst von Romberg, a German physician, described pulmonary vascular sclerosis, as far back as 1891. Despite the recognition of this disorder more than a century ago, some form of PH, e.g., pulmonary arterial hypertension (PAH), has no known cure. The available treatments are only to relieve the symptoms and slow the progress of the disease. Fortunately, through these efforts, the 3-year mortality rate has decreased over the past two decades to < 30% [1]. Additionally, a recent study using

These authors contributed equally: Arya Iranmehr, Tsering Stobdan

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-018-0270-8) contains supplementary material, which is available to authorized users.

✉ Gabriel G. Haddad
ghaddad@ucsd.edu

[1] Department of Electrical & Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA

[2] Division of Respiratory Medicine, Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

[3] Department of Medicine, University Hospitals Cleveland Medical Center, Cleveland, OH, USA

[4] National Academy of Sciences, Bishkek 720071, Kyrgyz Republic

[5] Department of Integrative Structural and Computational Biology, Scripps Research Institute, La Jolla, CA 92037, USA

[6] Human Longevity Inc., San Diego, CA 92121, USA

[7] J. Craig Venter Institute, La Jolla, CA 92037, USA

[8] Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA

[9] Department of Pediatrics, Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA

[10] Rady Children's Hospital, San Diego, CA 92123, USA

whole-genome sequencing has established the rate of gene mutation to be about 24% of PAH cases [2] providing additional therapeutic targets for the treatment of this disorder.

In countries like the Kyrgyz Republic where 90% of the area is at altitude > 3000 m (Figs. 1a, b), HAPH is a public health issue [3]. Previous studies on Kyrgyz highlanders have reported 14–20% of the population to have signs of HAPH [3]. This was based on signs of cor pulmonale in HA dwellers with HAPH or mean pulmonary pressures (mPAP) of > 25 mmHg [3]. Remarkably, an additional 21% manifest a > 2-fold increase in mPAP on exposure to acute hypoxia, i.e., 30 min of 11% oxygen breathing [3] even though they had normal resting mPAP of < 25 mmHg (Hyper-responder, HR). This high prevalence rate of HAPH in Kyrgyz highlanders provides an opportunity to understand the biology of susceptibility to HAPH or mal-adaptation to HA.

In general, HA studies provide one of the best natural experiments in humans to study gene vs. environment interactions [4]. As hypoxia is involved in the pathophysiology of many diseases including PH, it also provides a unique opportunity to identify the genes involved in hypoxia regulation, which can be explored for therapeutic purposes. With this insight, numerous studies, including some from our group, were conducted to study human adaptation/mal-adaptation to HA in Ethiopians, Tibetans and Andeans, the three major highland populations [5–9]. Using whole-genome sequencing, we were able to identify and functionally validate novel genes involved in altitude adaptation in these populations [5–7, 10, 11]. When describing 'adaptation/mal-adaptation to HA' in Kyrgyz highlanders, it is important to distinguish this unique population from the rest of the HA populations, i.e., Ethiopians, Tibetans and Andeans. Studies conducted on the other three HA populations successfully identified different physiological and genetic modes of adaptation, some common across populations [12] and others exclusive to one population [6]. However, very few studies have been conducted on Kyrgyz highlanders where the studies are either focused on single gene/single variant, i.e., *Angiotensin-Converting Enzyme Insertion/Deletion* polymorphism [3], or on using whole-exome sequencing [13]. It is important to note that the Kyrgyz population offers a unique advantage for the study of HAPH as compared with other HA populations. For example, the Kyrgyz subjects do not present with any other feature of chronic hypoxia besides HAPH, unlike the Andean population. Indeed, the Andean population, where HAPH was first described [14], does show other potentially confounding characteristics such as polycythemia, which, in turn, can lead to PH.

In the current study, we wanted to test the hypothesis that, due to genetic selection, healthy Kyrgyz highlanders are biologically protected from developing HAPH. Towards that end, we sequenced and analyzed the whole genome of Kyrgyz highlanders, in a case–control study design, and identified novel genes involved in HAPH. In addition, the candidate genes identified may also be related to PH in general as there is no study ever attempted to identify novel genes for HAPH using whole-genome sequence analysis.

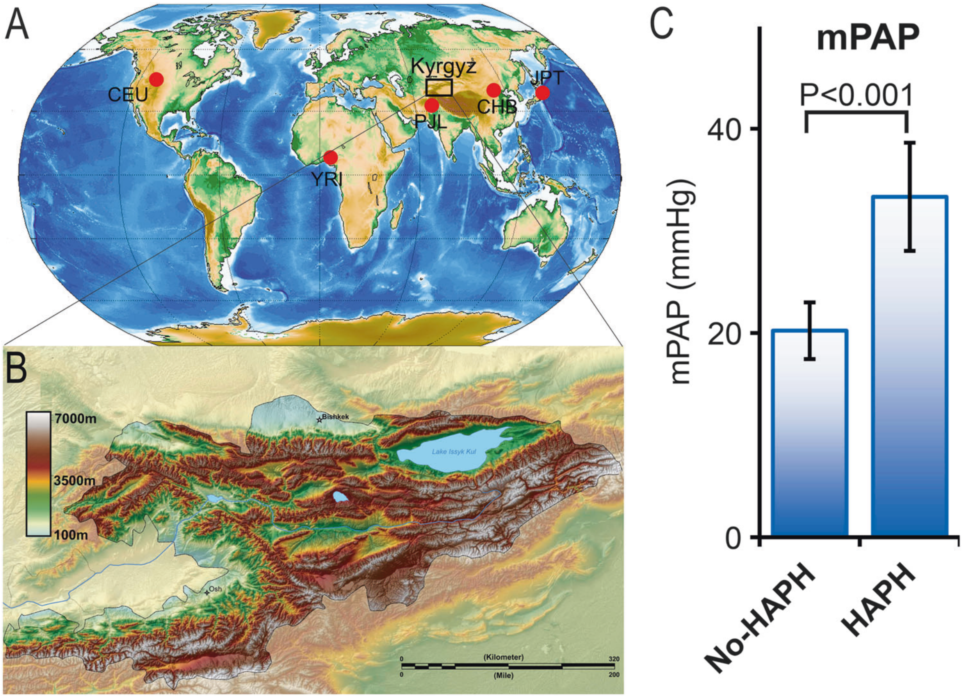

Fig. 1 Geographic location of Kyrgyz population (box) relative to other major populations from the 1000 Genome Project used in the Admixture analysis of current study (**a**). YRI Yoruba in Ibadan, Nigeria, CEU Utah Residents with Northern and Western European Ancestry, JPT Japanese in Tokyo, Japan, CHB Han Chinese in Beijing, China, SAS (PJL), Punjabi from Lahore, Pakistan. **b** Topography of Kyrgyz republic with > 90% of the area at altitude > 3000 m above sea level. **c** Mean pulmonary artery pressure (mPAP) in healthy Kyrgyz highlanders (No-HAPH) is significantly lower than the age-matched HAPH patients. Error bar represents ± standard error

Materials and methods

Study population

We included volunteers from a Kyrgyz highlander population. All individuals gave informed consent. A detailed phenotypic characterization of the cohort, including right heart catheterization studies are reported elsewhere [3, 15]. In brief, the cohort consisted of four groups, which included (a) HAPH ($n = 9$), (b) controls (No-HAPH, $n = 9$) consisting of healthy highlanders that age matched to the HAPH group, (c) HRs ($n = 7$), consisting of relatively younger individuals showing a significant increase in their pulmonary artery pressure (PAP, mmHg) when exposed to 11% oxygen and (d) normal-responders (NR, $n = 9$), were the control group (age matched to HR) not hyper-responding to lower oxygen. All the subjects were carefully analyzed by an expert person. The study was approved by an institutional review board.

Library construction and sequencing

Blood sample (10 mL) was collected from each subject for DNA extraction. The whole-genome sequencing was carried out at HLI (Human Longevity Inc., San Diego). Library preparation was carried out using the TruSeq Nano DNA HT kit (Illumina Inc.). Manufacturer's instruction was strictly followed at all steps. Whole-genome sequencing were done at a mean coverage of 40.3× on the Illumina HiSeqX sequencer utilizing a 150 base paired-end single index read format. The additional details of the library construction and the quality control are described in the Supplementary Material. The sequencing data from this study have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/datasets/) under accession number EGAS00001003171.

Admixture analysis and population structure

ADMIXTURE [16] was used to measure the genetic affinity of the Kyrgyz individuals with other major populations from the 1000 Genome Project [17]. To calculate ancestry proportions, we first pooled Kyrgyz samples with individuals from 1000 Genome Project populations, including YRI (Yoruba in Ibadan, Nigeria), CEU (Utah Residents with Northern and Western European Ancestry), PJL (SAS) (Punjabi from Lahore, Pakistan) and JPT (Japanese in Tokyo, Japan). Then we filtered out variants with allele frequency of <0.05 and ran ADMIXTURE program with the parameter $K = 3$. Additional details on population structure is described in the Supplementary Material.

Selection scan

A detailed explanation on selection scan is provided in the Supplementary Material. Briefly, we computed the test statistics using a sliding window of 50 kb, with steps of 10 kb, over the autosomal genome. D [18] and H [19] statistics were computed on a site frequency spectrum of a 50 kb window and did not require any post-processing. However, in integrated Haplotype Score (iHS), number of segregating sites by length (nSL), population branch statistics (PBS) and (Cross Population Extended Haplotype Homozygosity) (XP-EHH) scans where the statistic was computed for a single variant, we computed the average of the scores for each window. Subsequently, for each method, we took genome-wide top 0.1 percentile to be our significance cutoff, and merged overlapping significant windows to obtain distinct set of significant intervals for each scan.

Prioritization of the selected intervals

To prioritize genomic intervals of different lengths and levels of variation, we assigned a P-value to each of the 71 selected intervals, using a case–control style association. We then used false discovery rate (FDR) [20] of 10% to choose a significance cutoff. Under the null model, we expect that allele frequencies of case and control groups to be similar. Additionally, because case and control individuals belong to ethnic and local populations, our null model posits that allele frequencies of the control population is closer to the allele frequencies of the case population than that of the outgroup population, a genetically distant population. We use three populations: case (Healthy), control (Sick) and outgroup (JPT), as defined previously. We use the PBS statistic, which uses the "length" of the case-lineage with F_{st} as a measure of genomic distance [21]. Specifically, PBS $= -\log[1 - F_{st}(\text{case, control})] - \log[1 - F_{st}(\text{case, outgroup})] + \log[1 - F_{st}(\text{control, outgroup})]$.

We used all variants in the selected region to compute F_{st}, defined by Weir et al. [22]. For each selected interval, we computed a windowed PBS statistic and then computed its significance by testing against an empirical null distribution of that interval. We calculated the empirical null distribution by permuting samples of case and control populations (1000 times) and keeping the outgroup population fixed.

Haplotyping and other pre-processing

In order to evaluate iHS, nSL and XP-EHH statistics for the Kyrgyz dataset, we first inferred the phased haplotypes. We used Beagle v4.1 program [23] where a reference panel of EAS population from 1000 Genome Project is provided as an input parameter. We computed the *frequency of each* *haplotype* in a population as the statistical mode of the frequencies of the single nucleotide polymorphisms (SNPs) carried by the haplotype. Also, to identify the derived allele, we used the Ensembl Compara 59 database [24], which has inferred the ancestral allele on six primates.

eQTL analysis

Expression quantitative trait locus (eQTL) for different tissues were obtained from GTEx database [25]. We limited our eQTL analysis to tissues that were in the GTEx data and were directly related to PH. The list of eQTL SNPs with significant P-values ($P < 0.05$) were obtained for each candidate genes. The sample size for related tissues reported in the current study are Lung ($n = 383$) and whole blood ($n = 369$).

Results and discussion

To our knowledge, this is the first whole-genome sequence study performed on a Central Asian population, Kyrgyz highlanders (Figs. 1a, b). Additionally, we used a case–control study design, comparing HAPH with healthy controls. The subjects in these cohorts have been very well phenotyped (Fig. 1c) and cardiac catheterization has been done on all subjects [3]. As this is a study of only 33 subjects, a question could be asked about a potential bias because the sample size is relatively small. Although large-scale association studies of urban populations could potentially provide means for determining genetic architecture of *common* complex traits, studying of locally adapted ethnic populations can be used to increase statistical power and target less common phenotypes [5, 7, 21, 26–28]. This paradigm reduces the number of loci for association from millions to tens or thousands of genetic loci. We note that although our sample size is limited, we have access to whole-genome sequence covering all variants (40.3× coverage) to perform selection scan on 66 haplotypes to reduce the number of loci for statistical test for association. By utilizing a pipeline that identifies regions under positive selection through an analysis of all variants in large (>50 kb) segments, and an integrated statistical test, we have been successful in the past identifying candidate genes involved in HA adaptations both in Ethiopian and Andean populations and these were subsequently validated [6, 7, 10, 11]. We use a similar approach here, with some modifications (Methods).

Ancestry and population structure of Kyrgyz individuals

To identify the genetic history of the Kyrgyz population and to find a close outgroup population for selection scan, we performed admixture and principal component analysis (PCA). Chromosome-wide admixture analysis of Kyrgyz samples, along with samples from African (YRI, $n = 108$), European (CEU, $n = 99$), South Asian (PJL, $n = 96$) and East Asian (JPT, $n = 104$) (Fig. 2a) demonstrated that Kyrgyz highlanders (KGZ, $n = 33$) have a strong East Asian component ($\mu = 0.76$, $\sigma = 0.09$ for the distribution of the proportion of the East Asian Ancestry in Kyrgyz samples, where μ is mean and σ is standard deviation) along with some European ancestry ($\mu = 0.24$, $\sigma = 0.09$ for the distribution of the proportion of the European Ancestry in Kyrgyz samples). As East Asian ancestry is significantly higher in the Kyrgyz samples (two sample Kolmogorov–Smirnov $P = 0.00015$), we used JPT as reference population in our current study because along with others, we have found that this population was relatively closer to Kyrgyz than the other East Asian reference population, i.e., Han Chinese [29] (Supplementary Fig. S1). The admixture from East and West is consistent with the mitochondrial DNA ancestry analysis of the Kyrgyz highlanders [30]. No sign of shared ancestry was detected with YRI (two sample Kolmogorov–Smirnov $P = 3.98\text{E}{-}20$). Similar to KGZ, PJL showed European and East Asian ancestry, but with inverse proportions that were relatively inverted ($\mu = 0.82$, $\mu = 0.18$ for European and East Asian ancestry, respectively; Fig. 2a). Interestingly, despite the prehistoric and historic eastward migrations, the JPT proportion of admixture dominated both the mitochondrial/maternal lineage [30], as well as the nuclear DNA lineage (Fig. 2a).

We then performed PCA on the Kyrgyz sample along with 1000 Genome Project super-populations AFR ($n = 661$), EAS ($n = 504$), EUR ($n = 503$) and SAS ($n = 489$), and observed that the Kyrgyz individuals were located between EAS and SAS super-populations (Fig. 2b). A previous study on the population structure and genetic ancestry of Central Asia, which also included the Kyrgyz population has revealed similar findings [29]. Finally, we tested if there is a population structure within Kyrgyz populations that is correlated with phenotypic groups. To do this, we computed the PCA of the Kyrgyz samples and stratified the PCA projection by four phenotypic groups: No-HAPH, HAPH, HR and NR. As shown in Fig. 2c, the Kyrgyz subgroups do not reveal any significant substructure (one-way analysis of variance $P = 0.3$, 0.29 for the first and second principal components, respectively).

Selected intervals

We computed six different statistics that captured regions under selection over sliding windows of 50 kbp with steps of 10 kbp over each autosomal genome (Supplementary Fig. 2 and Methods). The phased and unphased dataset contain 6,841,212 and 11,703,698 variants, respectively,

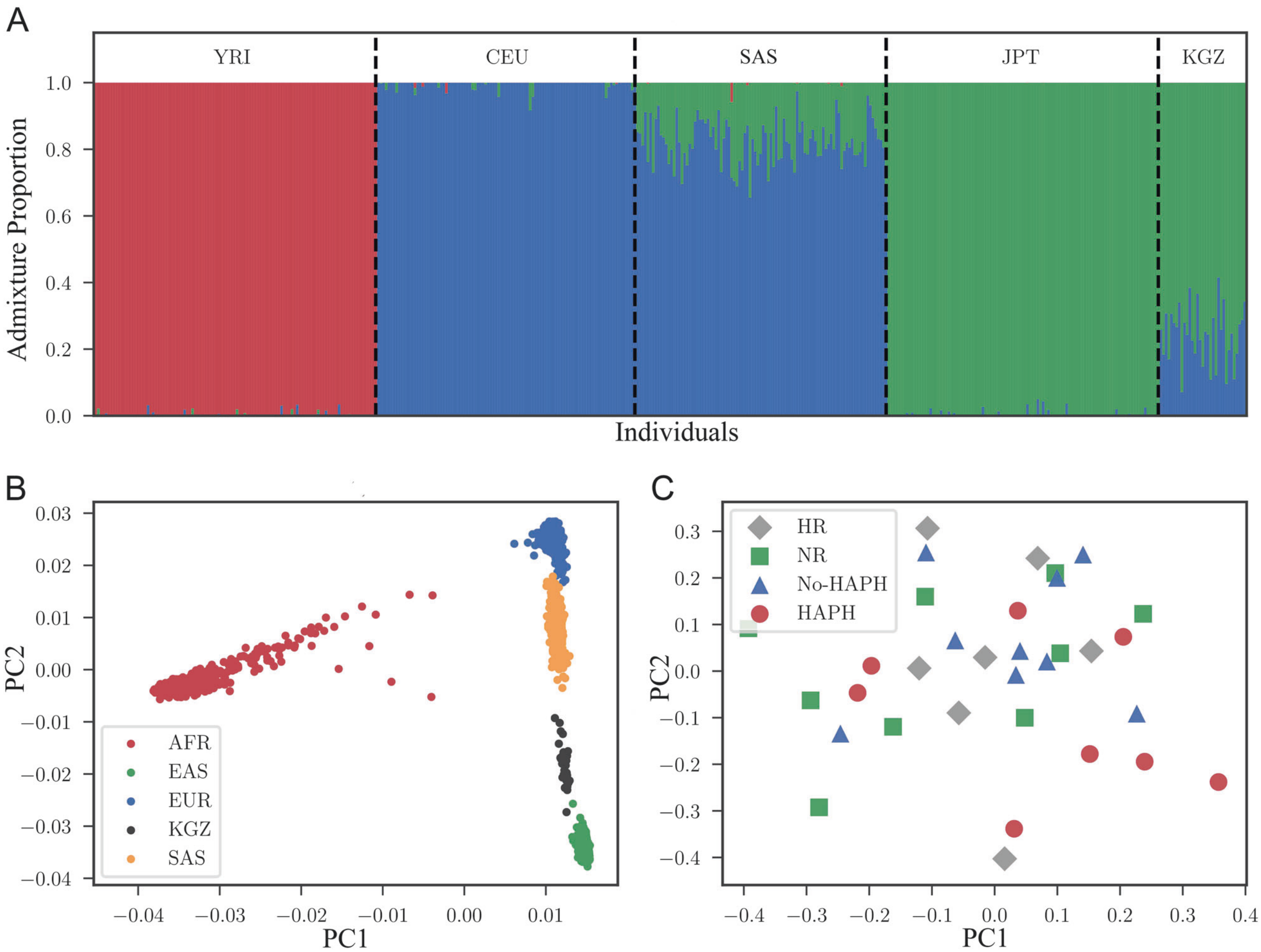

Fig. 2 Admixture and PCA depicting the genetic relatedness of Kyrgyz Population (KGZ), which includes HAPH, No-HAPH, HR and NR groups, to other major populations. **a** Admixture analysis shows that Kyrgyz population consists of major genetic proportion from East Asian lineage with minute contributions from the European genetic ancestry. **b** PCA reveals that the Kyrgyz population is located between SAS and EAS but more closely related to EAS. **c** Within Kyrgyz cluster the subjects are randomly distributed. SAS South Asian, EAS East Asian

resulting in 284,906 overlapping 50 kbp windows. For each test statistic, we identified windows that scored in the top 0.1 percentile (~ top 285) among all windows. We then merged the identified windows from the six different methods. This resulted in 71 distinct intervals. Next, we computed empirical *P*-values for each interval (Methods) and used 10% FDR cutoff [20] to identify top intervals (Table 1) for further analysis.

The most significant interval, interval-1 (Figs. 3a, b), located on chromosome 17, contains 147 SNPs with the haplotype extending to 892 kbp. The haplotype frequency difference between the cases and the controls was ~ 60% (Fig. 3b and Methods). The region contained 10 genes (Table 1, Fig. 3a). In order to make biological sense of the candidate interval, we performed an extensive literature survey looking for all possible aspects that would connect this interval to HAPH. Accumulating evidence suggests *MTMR4* (myotubularin-related protein 4, highlighted in Fig. 3a) as one of the biologically plausible candidate gene of HAPH. It contains tyrosine/dual-specificity phosphatase activity and is known to dephosphorylate SMAD1/2/3 (Mothers Against Decapentaplegic Homolog 1/2/3 (Drosophila)) [31, 32]. Previous studies have shown that transforming growth factor β (TGFβ)/Smad2/3 signaling is disrupted in a monocrotaline based rodent model of PAH [33]. Similarly, the disruption of bone morphogenetic proteins (BMPs) signaling, also a member of TGFβ superfamily, is a known factor that initiate PH [31]. Studies have also shown that *MTMR4* is an essential negative regulator of BMP signaling pathway [32]. Unlike TGFβ/Smad2/3 signaling, here it binds and dephosphorylates the activated Smad1 to attenuate BMP signaling [31, 32]. This activity was indeed confirmed by overexpressing *MTMR4* that led to repressed BMP-induced gene expression, and by *MTMR4* specific small interfering RNA enhancing BMP signaling [32]. Therefore, genetic selection involving selective inhibition of this gene would lead to an enhancement of both TGFβ and BMP signaling, which may protect No-HAPH from developing PH under chronic environmental hypoxia.

Table 1 Top intervals prioritized by windowed PBS empirical *P*-value

Region	Chrom	Start (Mbp)	End (Mbp)	Length (kbp)	Genes
1	17	56.41	57.242	832	*RNF43*, *HSF5*, ***MTMR4***, *SEPT4*, *C17orf47*, *TEX14*, *RAD51C*, *PPM1E*, *TRIM37*, *SKA2*
2	15	52.004	52.279	275	*SCG3*, *LYSMD2*, *TMOD2*, ***TMOD3***, *LEO1*
3	19	50.454	50.582	128	*IL4I1*, *NUP62*, *CTC-326K19.6*, *ATF5*, ***SIGLEC11***, *VRK3*, *ZNF473*
4	1	101.078	101.312	234	***VCAM1***, *DPH5*
5	1	31.702	31.968	265	*NKAIN1*, *SNRNP40*, *ZCCHC17*, *FABP3*, *SERINC2*, *AC114494.1*
6	6	54.839	55.147	308	*HCRTR2*
7	6	135.282	135.404	123	*ALDH8A1*, *HBS1L*
8	17	61.69	62.014	324	*DCAF7*, *TACO1*, *MAP3K3*, *LIMD2*, *RP11-51F16.8*, *STRADA*, *CCDC47*, *DDX42*, *FTSJ3*, *PSMC5*, *SMARCD2*, *CSH2*, *GH2*, *CSH1*, *CSHL1*, *GH1*, *CD79B*, *SCN4A*

The candidate genes within top regions are highlighted in bold. Note, *SIGLEC11* was identified in Ref. 17. The genome coordinates are based on hg19

Chrom chromosome

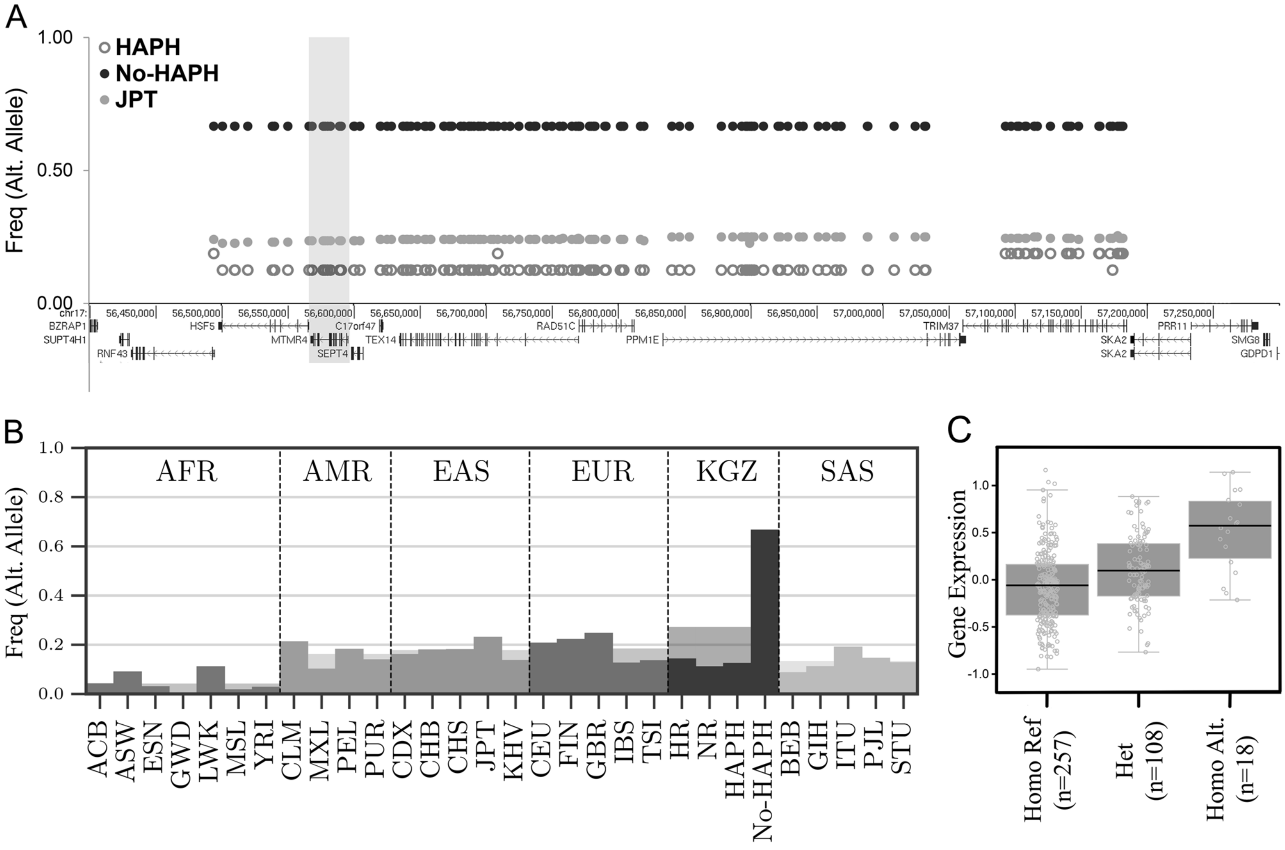

Fig. 3 Layout of genetic variation in the top selected interval-1 in the HAPH, No-HAPH and JPT (outgroup) populations (**a**). The haplotype frequencies among No-HAPH is higher compared with HAPH and JPT. **b** Frequency of the top selected haplotype (interval-1) among Kyrgyz highlanders and populations from the 1000 Genome Project. The *y* axis is frequency of one of the SNPs (out of 147 fully linked SNPs) of the selected haplotype. **c** A representative box plot showing the genotype of an eQTL SNP in interval-1 and the respective expression of gene in Lung (*P* = 1.5e-11)

Additionally, the haplotype interval also consists of two missense mutations located in *HSF5* (heat shock transcription factor 5, rs3803752 and rs117817367). Not much is reported on *HSF5*, but being from the HSF family and having transcription factor activity [34], its role in gene regulation events can be explored. We also performed eQTL analysis for these SNPs with gene expression levels in the tissues related to cardio-respiratory system.

Interestingly, a large number of these SNPs, i.e., from interval-1, were identified as eQTLs where the ancestral allele was significantly associated with lower *RAD51C* levels in the lungs (Fig. 3c; Supplementary Fig. S3 and Supplementary Table S1). A higher frequency of derived allele among the healthy Kyrgyz highlanders can be correlated with a higher *RAD51C* expression in the lungs. Previous studies have shown that an increased *RAD51C* expression is associated with lung cancer [35]. Given that immune and inflammatory processes triggered by cancer cells can also lead to PH [36], it is intriguing to discover a significantly higher frequency of derived alleles in the No-HAPH group as compared with all the other populations (Fig. 3b).

The second significant interval (Table 1 and Fig. 4a) is located on chromosome 15 (position 52007217–52268916, hg19) spanning 315 kbp. It overlays five genes, i.e., *SCG3*, *LYSMD2*, *TMOD2*, *TMOD3* and *LEO1* (Fig. 4a and Table 1). The haplotype consisted of 61 SNPs differing significantly ($P < 0.05$) between the No-HAPH and controls (Fig. 4b). All the SNPs were in the non-coding regions and two SNPs located in the promoter region of *TMOD3* (rs11637876 and rs12913583) had CADD's PHRED value of 17.2 and 14.2, respectively. Large number of SNPs, which also included rs11637876 and rs12913583, were identified as eQTLs linked to *TMOD3* expression (Figs. 4c, d; Supplementary Fig. S3) in specific tissues associated with PH, e.g., whole blood (Fig. 4c; $P = 5.5\text{e}{-7}$), aorta (Fig. 4d; $P = 1.0\text{e}{-7}$).

When we explored the biological significance of this gene in term of HAPH, we found that the expression of *TMOD3* is increased in patients with idiopathic PAH [37]. This gene is expressed in the motile endothelial cells where it caps the pointed ends of actin filaments [38]. The capping inversely correlates with endothelial cell migration rates and the overexpression of *Tmod3* inhibits cell migration [38]. Keeping in mind the importance of cell migration in tissue repair responses, e.g., vascular repair and regeneration in PH [39], a higher expression of *TMOD3* in the PH patients [37] may be linked to *TMOD3*-related delay in cell migration for tissue repair process. A complete knockout of this gene in mice is lethal due to embryonic anemia and defects in fetal erythropoiesis [40] and therefore further studies involving tissue-specific differential expression will be needed to specify its role in PH pathogenesis.

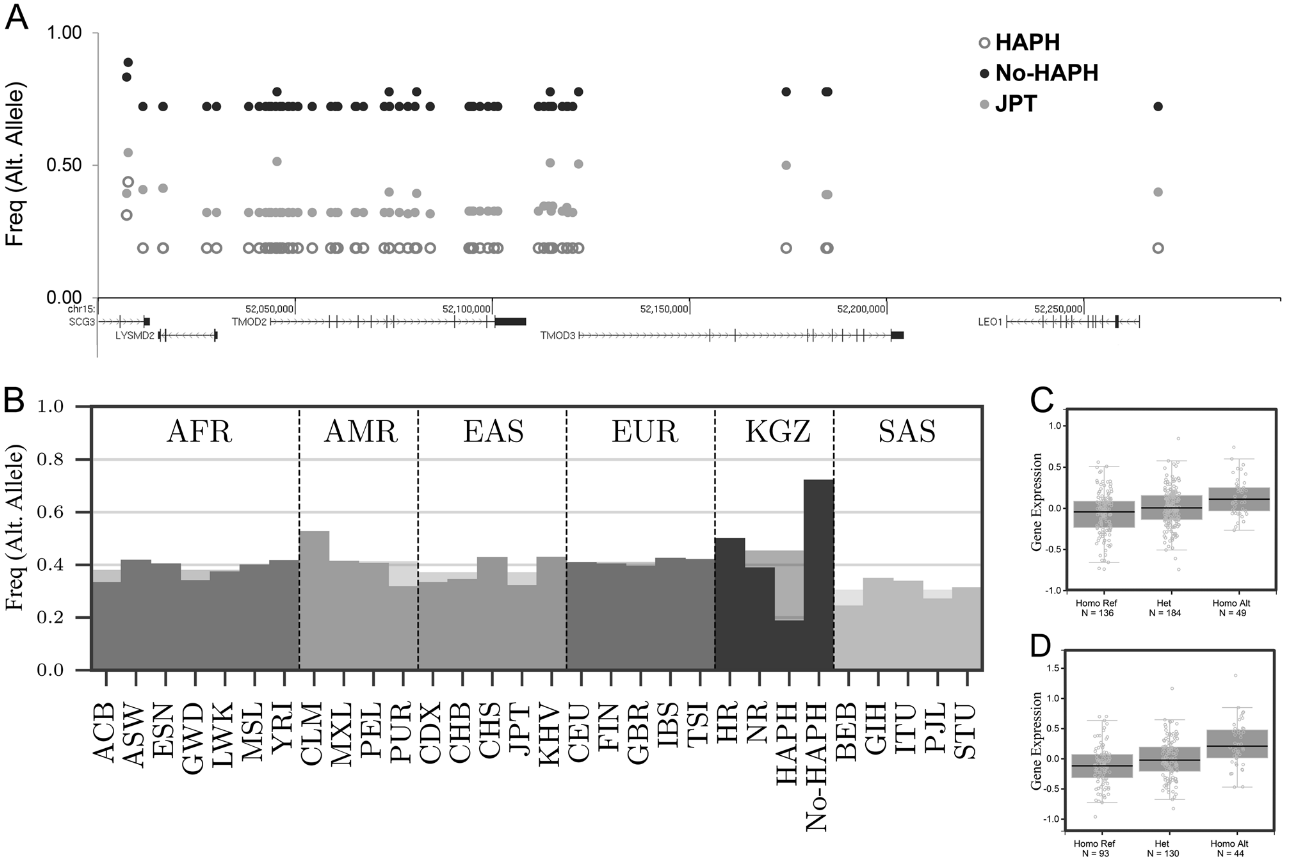

Fig. 4 Layout of genetic variation in selected interval-2 in the HAPH, No-HAPH and JPT (outgroup) populations (**a**). Haplotype frequencies among No-HAPH are higher compared with HAPH and JPT. **b** Frequency of a SNP from a set of perfectly linked SNP of interval-2 among Kyrgyz highlanders and populations from the 1000 Genome Project (**b**). Box plot of SNPs, rs11637876 (**c**) and rs12913583 (**d**) from interval-2 that was identified as eQTL for the expression of *TMOD3* ($P = 5.5\text{e}{-7}$ for both SNPs)

The involvement of *TMOD2*, also from the same tropomodulin family, in HAPH is less likely because of its restricted expression in neural tissues [41].

Other candidate intervals

The third top interval consists of seven genes. Interestingly, *SIGLEC11* (sialic acid binding Ig-like lectin 11), which mediates anti-inflammatory and immunosuppressive signaling was previously found to be associated with HAPH [13]. There was also ample evidence that *VCAM1* (vascular cell adhesion molecule-1) from “interval 4” (Table 1) has a role in HAPH. The gene is a member of the immunoglobulin superfamily and is known to be a marker of endothelial cell inflammation, and mediate adhesion of white blood cells to the endothelium. *VCAM1* has three different splice variants and the transcription levels when measured in different cell lines with ENCODE reveals higher expression levels in the *Homo sapiens* endothelial of umbilical vein primary cell (HUVEC) and *Homo sapiens* skeletal muscle myoblast primary (HSMM) cells (Supplementary Fig. S4). Interestingly, the H3K4me1 and H3K27Ac tag densities, which often found near regulatory elements are higher only in the HUVEC cells and the peaks also align with few of the selected SNPs (Supplementary Fig. S4). This could indicate differential regulation of the selected SNPs specific to endothelial cells. Higher levels of VCAM1 are associated with renal dysfunction, hepatic impairment and more importantly correlated with the severity of PH in patients with sickle cell disease [42]. The levels were also found high in peripheral blood [43] and lung fibroblasts [44] of patients suffering from idiopathic pulmonary fibrosis. The study also shows that TGF-β1 treatment leads to an increase in VCAM1 levels at transcriptional, as well as protein levels while silencing VCAM1 expression inhibits fibroblast proliferation [44] and the role of PH in idiopathic pulmonary fibrosis etiology is well recognized. Furthermore, its role in allergic lung diseases [45] or in systemic sclerosis [46], where PAH is the leading cause of death [47], clearly indicates evidence of a positive selection sweep on genetic variants that would protect the No-HAPH subjects from developing PH. Furthermore, studies in mice with Lipopolysaccharide-induced acute lung injury have clearly shown an upregulation of VCAM1 [48]. Similarly, transcriptome analysis of patients diagnosed with HA pulmonary edema (HAPE), also an acute lung injury, depicts maximum fold change for VCAM1 along with MAPK10 [49]. This upregulation in HAPE patients provides a direct link between VCAM1 and HAPH because PH is a hallmark in patients with HAPE. In-vitro analysis modeling PAH by CypA-induced oxidative stress also revealed increased VCAM1 [50].

Overall, our analysis of whole-genome sequence clearly indicates that there are multiple genomic regions under selection (Table 1). This is also supported by a recent whole-exome analysis of the same population where they have identified 33 candidate genes linked to HAPH [13]. Interestingly, only one gene out of these 33 candidates was among the top candidate interval in the current study. These differences could be due to the input data and computational pipelines, which are completely different in the two studies. For example, a stronger genetic selection in the regulatory regions such as promoters and enhancers, which the whole-exome sequencing fails to capture. In our previous studies on selection scan at HA, we observed similar signals, i.e., variants in the non-coding region, which were subsequently validated in different model systems [5–7, 10, 11]. Additionally, because of weaker signals of selection, other genes may not have passed multiple scans of selection and the three-way association filtering criteria that we applied in the current analysis. The genes like *DST*, *SDK1* and *OSBPL9* from the previous study were part of our initial gene list but were subsequently dropped out as they failed the subsequent stringent PBS association filter. In order to recapitulate previous findings in our study, we took the specific variants (chromosome position and rsID#) from the previous study [13] and systematically scan all the 33 signals in our current subjects/samples. In all cases (including *SIGLEC11*), we found that the frequency of heterozygotes was at most 3/33 (Supplementary Table 2) and did not meet the criterion established by the previous study [13]. Finally, there were additional known genes located in different intervals where the haplotype frequencies were significantly different between No-HAPH vs HAPH comparisons (Empirical F_{st}, $P < 0.05$) but not significant when compared with the outgroup population, such as with the gene *EDNRB* (Supplementary Fig. S5). Even though the difference remains significant (Empirical F_{st}, $P < 0.05$) when No-HAPH is compared with other major populations, i.e., AFR, EUR and SAS, it is not consistent with our “local adaptation model”, i.e., natural selection is specific to the adapted (No-HAPH) population’s environment. A simple explanation could be an involvement of other selection pressures on this region in the JPT, totally unrelated to HAPH that may have impacted the genomic structure in this region among JPT to behave in a manner similar to our No-HAPH group.

Conclusions

Our study provides the first whole-genome sequence analysis of a Central Asian population from the Kyrgyz highland and also for HAPH with prospect of identifying novel genes for this disease. Admixture analysis revealed that the Kyrgyz highlanders consisted of both European and East Asian ancestry with an overwhelming contribution from the East Asian gene pool. Evolutionarily, the typical form of HAPH is distinctive to Kyrgyz highlanders and in the current study, we utilized this feature and discover novel genetic markers with a potential to give insight into therapy for PH. From the top candidate genes detected, we raise here three possibilities that may individually or together lead to HAPH in Kyrgyz highlanders. This includes, *first*, a dysfunctional BMP signaling involving *MTMR4* overexpression as an alternate gene regulating BMP signaling. *Second*, *TMOD3*-related delayed cell migration for tissue repair process. And *third*, an abnormal endothelial cell function with elevated *VCAM1* (a schema is presented in Supplementary Fig. S6). However, despite an unbiased identification approach with ample literature evidence, we cannot rule out the involvement of additional genes or gene interactions in HAPH. Future studies targeting these genes may strengthen our findings and will provide a better understanding of HAPH.

Acknowledgements This work was supported by NIH grant (R01 HL127403-02 and R21NS101652-01) GGH. VB and AI were supported by grants from the NSF (NSF-DBI-1458557 and NSF-IIS-1318386) and NIH (R01GM114362). We thank Prof. Erkin Mirrakhimov, Kyrgyz State Medical Academy and National Center and Dr. Aibek Mirrakhimov for their help in recruiting the volunteers.

Compliance with ethical standards

Conflict of interest VB is a co-founder, has an equity interest and receives income from Digital Proteomics, LLC and Pretzel Genomics Ltd. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. Digital Proteomics, LLC was not involved in the research presented here. EHMW, EFK and JCV are employees of Human Longevity Inc.

References

1. Lai Y-C, Potoka KC, Champion HC, Mora AL, Gladwin MT. Pulmonary arterial hypertension: the clinical syndrome. Circ Res. 2014;115:115–30.
2. Gräf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat Commun. 2018;9:1416.
3. Aldashev AA, Sarybaev AS, Sydykov AS, Kalmyrzaev BB, Kim EV, Mamanova LB, et al. Characterization of high-altitude pulmonary hypertension in the Kyrgyz: association with angiotensin-converting enzyme genotype. Am J Respir Crit Care Med. 2002;166:1396–402.
4. Stobdan T, Karar J, Pasha MAQ. High altitude adaptation: genetic perspectives. High Alt Med Biol. 2008;9:140–7.
5. Stobdan T, Akbari A, Azad P, Zhou D, Poulsen O, Appenzeller O, et al. New insights into the genetic basis of Monge's disease and adaptation to high-altitude. Mol Biol Evol. 2017;34:3154–68.
6. Udpa N, Ronen R, Zhou D, Liang J, Stobdan T, Appenzeller O, et al. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 2014;15:R36.
7. Zhou D, Udpa N, Ronen R, Stobdan T, Liang J, Appenzeller O, et al. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. Am J Hum Genet. 2013;93:452–62.
8. Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. 2010;6:e1001116.
9. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, et al. Genetic evidence for high-altitude adaptation in Tibet. Science. 2010;329:72–75.
10. Stobdan T, Zhou D, Ao-Ieong E, Ortiz D, Ronen R, Hartley I, et al. Endothelin receptor B, a candidate gene from human studies at high altitude, improves cardiac tolerance to hypoxia in genetically engineered heterozygote mice. Proc Natl Acad Sci USA. 2015;112:10425–30.
11. Azad P, Zhao HW, Cabrales PJ, Ronen R, Zhou D, Poulsen O, et al. Senp1 drives hypoxia-induced polycythemia via GATA1 and Bcl-xL in subjects with Monge's disease. J Exp Med. 2016;213:2729–44.
12. Azad P, Stobdan T, Zhou D, Hartley I, Akbari A, Bafna V, et al. High-altitude adaptation in humans: from genomics to integrative physiology. J Mol Med. 2017;95:1269–82.
13. Wilkins MR, Aldashev AA, Wharton J, Rhodes CJ, Vandrovcova J, Kasperaviciute D, et al. α1-A680T variant in GUCY1A3 as a candidate conferring protection from pulmonary hypertension among Kyrgyz highlanders. Circ Cardiovasc Genet. 2014;7:920–9.
14. Rotta A, Cánepa A, Hurtado A, Velásquez T, Chávez R. Pulmonary circulation at sea level and at high altitudes. J Appl Physiol. 1956;9:328–36.
15. Kojonazarov BK, Imanov BZ, Amatov TA, Mirrakhimov MM, Naeije R, Wilkins MR, et al. Noninvasive and invasive evaluation of pulmonary arterial pressure in highlanders. Eur Respir J. 2007;29:352–6.
16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68.
18. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genet Genet Soc Am. 1989;123:585–95.
19. Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genet Genet Soc Am. 2000;155:1405–13.
20. Storey JD, Tibshirani R. Statistical significance for genome wide studies. Proc Natl Acad Sci USA. 2003;100:9440–5.
21. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329:75–78.
22. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evol [Soc Study Evol, Wiley]. 1984;38:1358–70.
23. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
24. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012;40:D84–90.
25. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45:580–5.

26. Tishkoff S. Strength in small numbers. Science. 2015;349:1282–3.
27. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015;349:1343–7.
28. Ilardo MA, Moltke I, Korneliussen TS, Cheng J, Stern AJ, Racimo F, et al. Physiological and genetic adaptations to diving in sea nomads. Cell. 2018;173:569–.e15.
29. Hodoğlugil U, Mahley RW. Turkish population structure and genetic ancestry reveal relatedness among Eurasian populations. Ann Hum Genet. 2012;76:128–41.
30. Peng M-S, Xu W, Song J-J, Chen X, Sulaiman X, Cai L, et al. Mitochondrial genomes uncover the maternal history of the Pamir populations. Eur J Hum Genet. 2017. https://doi.org/10.1038/s41431-017-0028-8
31. Newman JH, Wheeler L, Lane KB, Loyd E, Gaddipati R, Phillips JA 3rd, et al. Mutation in the gene for bone morphogenetic protein receptor II as a cause of primary pulmonary hypertension in a large kindred. N Engl J Med. 2001;345:319–24.
32. Yu J, He X, Chen Y-G, Hao Y, Yang S, Wang L, et al. Myotubularin-related protein 4 (MTMR4) attenuates BMP/Dpp signaling by dephosphorylation of Smad proteins. J Biol Chem. 2013;288:79–88.
33. Zakrzewicz A, Kouri FM, Nejman B, Kwapiszewska G, Hecker M, Sandu R, et al. The transforming growth factor-beta/Smad2,3 signalling axis is impaired in experimental pulmonary hypertension. Eur Respir J. 2007;29:1094–104.
34. Gomez-Pastor R, Burchfiel ET, Thiele DJ. Regulation of heat shock transcription factors and their roles in physiology and disease. Nat Rev Mol Cell Biol. 2018;19:4–19.
35. Chen X, Qian D, Cheng J, Guan Y, Zhang B, Ding X, et al. High expression of Rad51c predicts poor prognostic outcome and induces cell resistance to cisplatin and radiation in non-small cell lung cancer. Tumour Biol. 2016;37:13489–98.
36. Pullamsetti SS, Kojonazarov B, Storn S, Gall H, Salazar Y, Wolf J, et al. Lung cancer-associated pulmonary hypertension: role of microenvironmental inflammation based on tumor cell-immune cell cross-talk. Sci Transl Med. 2017;9. https://doi.org/10.1126/scitranslmed.aai9048
37. Rajkumar R, Konishi K, Richards TJ, Ishizawar DC, Wiechert AC, Kaminski N, et al. Genome wide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension. Am J Physiol Heart Circ Physiol. 2010;298:H1235–48.
38. Fischer RS, Fritz-Six KL, Fowler VM. Pointed-end capping by tropomodulin3 negatively regulates endothelial cell motility. J Cell Biol. 2003;161:371–80.
39. Farkas L, Kolb M. Vascular repair and regeneration as a therapeutic target for pulmonary arterial hypertension. Respiration. 2013;85:355–64.
40. Sui Z, Nowak RB, Bacconi A, Kim NE, Liu H, Li J, et al. Tropomodulin3-null mice are embryonic lethal with anemia due to impaired erythroid terminal differentiation in the fetal liver. Blood. 2014;123:758–67.
41. Cox PR, Zoghbi HY. Sequencing, expression analysis, and mapping of three unique human tropomodulin genes and their mouse orthologs. Genomics. 2000;63:97–107.
42. Kato GJ, Martyr S, Blackwelder WC, Nichols JS, Coles WA, Hunter LA, et al. Levels of soluble endothelium-derived adhesion molecules in patients with sickle cell disease are associated with pulmonary hypertension, organ dysfunction, and mortality. Br J Haematol. 2005;130:943–53.
43. Richards TJ, Kaminski N, Baribaud F, Flavin S, Brodmerkel C, Horowitz D, et al. Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2012;185:67–76.
44. Agassandian M, Tedrow JR, Sembrat J, Kass DJ, Zhang Y, Goncharova EA, et al. VCAM-1 is a TGF-β1 inducible gene upregulated in idiopathic pulmonary fibrosis. Cell Signal. 2015;27:2467–73.
45. Popper HH, Pailer S, Wurzinger G, Feldner H, Hesse C, Eber E. Expression of adhesion molecules in allergic lung diseases. Virchows Arch. 2002;440:172–80.
46. Shahin AA, Anwar S, Elawar AH, Sharaf AE, Hamid MA, Eleinin AA, et al. Circulating soluble adhesion molecules in patients with systemic sclerosis: correlation between circulating soluble vascular cell adhesion molecule-1 (sVCAM-1) and impaired left ventricular diastolic function. Rheumatol Int. 2000;20:21–24.
47. Chaisson NF, Hassoun PM. Systemic sclerosis-associated pulmonary arterial hypertension. Chest. 2013;144:1346–56.
48. Tao J, Nie Y, Hou Y, Ma X, Ding G, Gao J, et al. Chemomics-integrated proteomics analysis of Jie-Geng-Tang to ameliorate lipopolysaccharide-induced acute lung injury in mice. Evid Based Complement Altern Med. 2016;2016:7379146.
49. Sharma M, Singh SB, Sarkar S. Genome wide expression analysis suggests perturbation of vascular homeostasis during high altitude pulmonary edema. PLoS ONE. 2014;9:e85902.
50. Xue C, Sowden M, Berk BC. Extracellular cyclophilin A, especially acetylated, causes pulmonary hypertension by stimulating endothelial apoptosis, redox stress, and inflammation. Arterioscler Thromb Vasc Biol. 2017;37:1138–46.