scientific reports

OPEN

Establishing a prediction model of severe acute mountain sickness using machine learning of support vector machine recursive feature elimination

Min Yang[1]✉, Yang Wu[1], Xing-biao Yang[1], Tao Liu[1], Ya Zhang[1], Yue Zhuo[1], Yong Luo[1] & Nan Zhang[2]

Severe acute mountain sickness (sAMS) can be life-threatening, but little is known about its genetic basis. The study was aimed to explore the genetic susceptibility of sAMS for the purpose of prediction, using microarray data from 112 peripheral blood mononuclear cell (PBMC) samples of 21 subjects, who were exposed to very high altitude (5260 m), low barometric pressure (406 mmHg), and hypobaric hypoxia (VLH) at various timepoints. We found that exposure to VLH activated gene expression in leukocytes, resulting in an inverted CD4/CD8 ratio that interacted with other phenotypic risk factors at the genetic level. A total of 2286 underlying risk genes were input into the support vector machine recursive feature elimination (SVM-RFE) system for machine learning, and a model with satisfactory predictive accuracy and clinical applicability was established for sAMS screening using ten featured genes with significant predictive power. Five featured genes (EPHB3, DIP2B, RHEBL1, GALNT13, and SLC8A2) were identified upstream of hypoxia- and/or inflammation-related pathways mediated by microRNAs as potential biomarkers for sAMS. The established prediction model of sAMS holds promise for clinical application as a genetic screening tool for sAMS.

Acute mountain sickness (AMS) is believed to be a self-limiting syndrome of nonspecific symptoms concerning fatigue, headache, nausea, and dizziness, which may occur in nonacclimatized individuals under acute exposure to high altitudes above 2500 m[1]. In certain conditions without medical care and/or in certain groups with high risks, it is possible to develop severe AMS (sAMS), sometimes even accompanied by life-threatening situations such as cerebral edema and/or pulmonary edema[2]. Generally, symptoms of mild to moderate AMS can occur early and peak within 24–72 h post high-altitude exposure, which typically overlaps the time duration when cerebral edema or pulmonary edema occurs alongside[1,3,4]. The potential continuum from AMS to sAMS, cerebral edema or pulmonary edema suggests that preventing sAMS may hold promise to avoid related events at high altitude. The occurrence of these severe disorders, to a great extent, is determined by the planned altitude, the ascent speed and the individual susceptibility; thus, the incidence of these conditions may vary greatly in different studies[3]. Accordingly, it is usually difficult to predict who is at risk of developing sAMS for prevention purposes.

Partial pressure of oxygen (PaO_2)[5], the partial pressure of carbon dioxide ($PaCO_2$)[6], the saturation of oxygen (SaO_2)[7,8], arterial oxygen content (CaO_2)[9], oxygen tension at 50% haemoglobin saturation (P50)[10] and haemoglobin[11] were thought to be hypoxia-sensitive and evidenced either as independent predictors or factors related to the subsequent development of AMS; however, speculation remains regarding their importance in the prediction of AMS. To date, studies concerning the predictive value of blood gas testing are rather limited. Blood gas findings are usually inconsistent for possible interference from field or laboratory conditions or individual reasons. Additionally, statistically significant differences usually require large-scale and/or randomized controlled trials, which are currently almost impossible to complete under high-altitude circumstances. In addition, pulmonary function testing[12], cardiopulmonary exercise testing[13], and hypoxic exercise testing[14] have been

[1]Department of Traditional Chinese Medicine, Rheumatology Center of Integrated Medicine, The General Hospital of Western Theater Command, PLA, Chengdu 610083, China. [2]Department of Hematology, The General Hospital of Western Theater Command, PLA, Chengdu 610083, China. ✉email: translie@live.cn

used to assess the risk of hypoxemia, but the applicability of these measurements to high-altitude exposure has not been fully established[2].

Despite the variety of the ascent plan and the individual baseline medical conditions, genetic susceptibility is addressed to explain why AMS and the related events may still occur in certain groups[15]. Genetic issues concerning AMS have been previously discussed[16–20], however, evidence for the genetic susceptibilities to AMS is still very rare, even less to sAMS. In the study, microarray data of GSE103927[21] were explored for the genetic background of AMS, and a prediction model of sAMS was established by machine learning of the support vector machine recursive feature elimination (SVM-RFE) method[22], which was clinically applicable as tested within the timeline of the GSE103927 cohort and validated in an isolated cohort GSE52209[23]. Five featured genes (EPHB3, DIP2B, RHEBL1, GALNT13, and SLC8A2) were identified as important regulators of hypoxia-related processes, including erythrocyte differentiation, alpha–beta T cell differentiation, and secretion of histamine by mast cells. The study was a preliminary attempt to explore the genetic susceptibility of sAMS, which occurred in almost half of the GSE103927 subjects exposed to very high altitude (5260 m), low barometric pressure (Pb, 406 mmHg), and hypobaric hypoxia (VLH).

Results

Exposure to VLH activated gene expression in leukocytes. The microarray data of GSE103927 were based on 112 peripheral blood mononuclear cell (PBMC) samples collected from 21 subjects at seven time points of VLH exposure. At baseline, subjects were studied at sea level (130 m, Pb = 749 mmHg). At the first day noon (d1noon) or post meridiem (d1pm), d7, d16, post-decent day 7 (post7), or post21, subjects ascended or reascended to the target altitude (5260 m, Pb = 406 mmHg). All the subjects were sampled and studied following each rapid ascending to, or prolonged stay at the target altitude (Fig. 1a). Laboratory values concerning PaO_2 (mmHg), $PaCO_2$ (mmHg), SaO_2 (%), CaO_2 (mL/dL), P50 (mmHg), hemoglobin (g/dL), Lake Louise Questionnaire (LLQ)-AMS score, and AMS-C-Composite score were detected and recorded. Principal component analysis (PCA) was conducted to determine the timeline differences in gene expression (Fig. 1b). Gene expression patterns at d1pm and d7 indicated an apparent distinction from baseline (increasing along the PC1 axis), but this trend was reduced along both the PC1 and PC2 axes at d16 and post7, suggesting that acute exposure to VLH may change the gene expression pattern in monocytes. Then, the differentially expressed genes (DGs) across the baseline, d1pm, and d7 were determined by P value (< 0.05) and Log2(fold change, FC) (> 1.3 or $< - 1.3$). There were 512 overlapping genes with consistent significance along the exposure timeline from d1pm to d7 (Fig. 1c, Supplementary Table S1). We then conducted gene ontology (GO) analysis to identify the function of the 512 DGs, and it was indicated that most of them were involved in leukocyte activation (Fig. 1d). Twelve DGs enriched in the GO term "leukocyte activation involved in immune response" were further explored for their expression level along the timeline (Fig. 1e). These leukocyte activation-related genes were upregulated along the timeline with a peak at d7, suggesting that leukocytes were activated upon VLH, but they recovered along the exposure timeline.

T cells dominated the genetic responses upon VLH exposure. To investigate the functional differentiation of leukocytes upon VLH exposure, 1644 genes (Supplementary Table S2) with significant differences from the baseline were functionally clustered along the timeline, and six clusters were obtained (Fig. 2a). Both cluster 1 (299 genes) and cluster 6 (384 genes) were inversely changed across the timeline, with a peak decrease in cluster 1 and an increase in cluster 6 at d7. Different expression dynamics were also observed in cluster 2 (213 genes), cluster 3 (281 genes), cluster 4 (123 genes), and cluster 5 (344 genes). The function of each cluster was annotated using leukocyte 22 data matrix (LM22)[24] (Fig. 2b). As expected, each cluster was functionally associated with different leukocyte types (two-tailed spearman correlation test, N = 22, $P < 0.05$), with CD4/CD8 T cells dominating in clusters 1/2/4/5, regulatory T cells in clusters 4/5, gamma delta T cells in clusters 2/4/5/6, resting NK cells in cluster 6, monocytes in clusters 2/6, activated dendritic cells in cluster 3, and neutrophils in cluster 6. We then compared the biological functions among the clusters using the GO analysis (Fig. 2c). No significant terms were enriched in cluster 5. No relationship was observed in the expression pattern between clusters 2, 3, and 4. Both clusters 1 and 6 overlapped in T cell-associated functions, oxidative stress, and blood coagulation. Cluster 1 was specifically enriched in cell death in response to hydrogen peroxide and regulation of response to reactive oxygen species. Cluster 6 was specifically enriched in platelet activation. Both clusters 1 and 6 were indicated with dramatic altitude or hypoxia sensitivity, and both were functionally associated with T cell activities, suggesting the dominant roles of T cells in the genetic responses to VLH exposure.

The inverted CD4/CD8 ratio may function as a risk factor for sAMS. To investigate the timeline abundance of various leukocytes in subjects (N = 11) exposed to VLH, the cibersort algorithm[25] was applied in combination with LM22, and expression profiling was performed at baseline, d1pm, d7, d16, and post7 (Fig. 3a and Supplementary Table S3). CD8 and CD4 T cells, resting NK cells, and monocytes were indicated as the major cell types in PBMC samples, with a significant decrease (paired t test, $P < 0.05$) in the estimated proportion of CD4 cells at d1pm and d7 vs. baseline, which resulted in significantly inverted CD4/CD8 ratio at d1pm and d7 (N = 11, paired t test, $P < 0.05$; Fig. 3b, Supplementary Table S4). When predicted using the area under the receiver operating characteristic curve (ROC AUC) (N = 11), SaO_2 (AUC = 0.833), P50 (AUC = 0.833), and the serum level of hemogbin (AUC = 0.792) performed well in differentiating subjects with a normal or inverted CD4/CD8 ratio (Fig. 3c), suggesting their potential effects on CD4/CD8 balance. SaO_2 (AUC = 0.617), CaO_2 (AUC = 0.783), P50 (AUC = 0.667), hemogbin (AUC = 0.850), and the estimated proportion of T cells (AUC = 0.633) were also indicated as possible binary classifiers for sAMS (sAMS or non-sAMS) (Fig. 3d). The interplay effects between

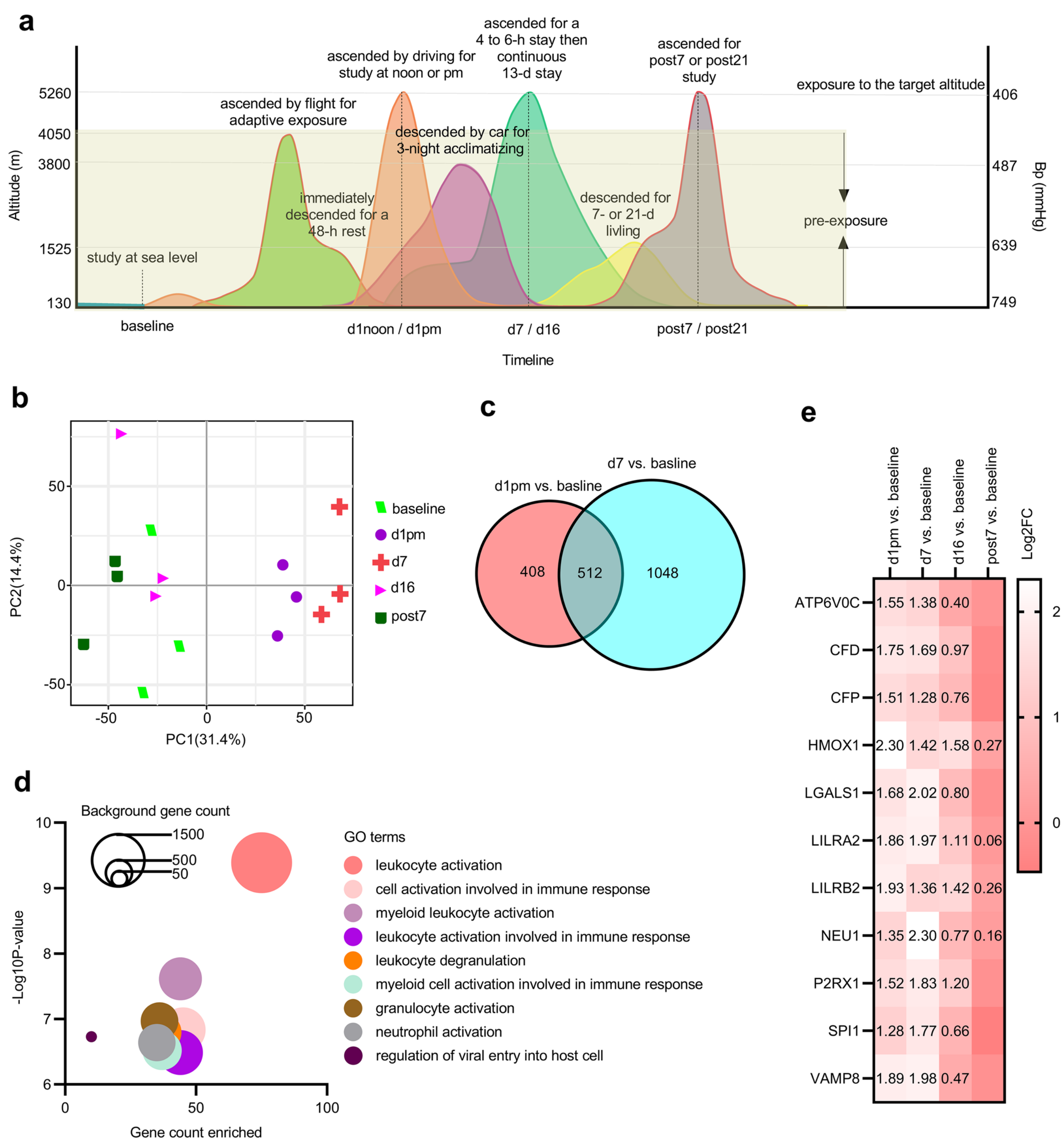

Figure 1. Acute exposure to VLH caused leukocyte activation. (**a**) Altitude-timeline plan of GSE103927. Subjects were studied at sea level for 30 days and then ascended to 4050 m by flight, followed by an immediate descent to 1525 m for a 48-h rest. After that, subjects ascended to the target altitude 5260 m for the d1 study at noon and pm, then decended to 3800 m for 3-night acclimatizing, allowing a 4- to 6-h exposure at the target altitude, then a continuous 13-day exposure at 5260 m for d7 and d16 studies. After that, subjects were allowed to stay at 1525 m again for 7 or 21 days and then reascended for the post7 or post21 study. (**b**) PCA for timeline gene expression patterns. (**c**) Venn analysis. (**d**) GO analysis. (**e**) Timeline changes in 12 DGs related to leukocyte activation.

the CD4/CD8 ratio, SaO_2, P50, hemoglobulin, the estimated proportion of T cells, and CaO_2 implied that the inverted CD4/CD8 ratio may function as the potential risk factor for sAMS.

Genetic profiling for sAMS. To uncover the gene signature underlying the phenotypes of sAMS, subjects in the training cohort were binarily subgrouped with the best predicted thresholds of various classifiers. The interplay effects between classifiers were investigated regarding the expression pattern of DGs across the binary

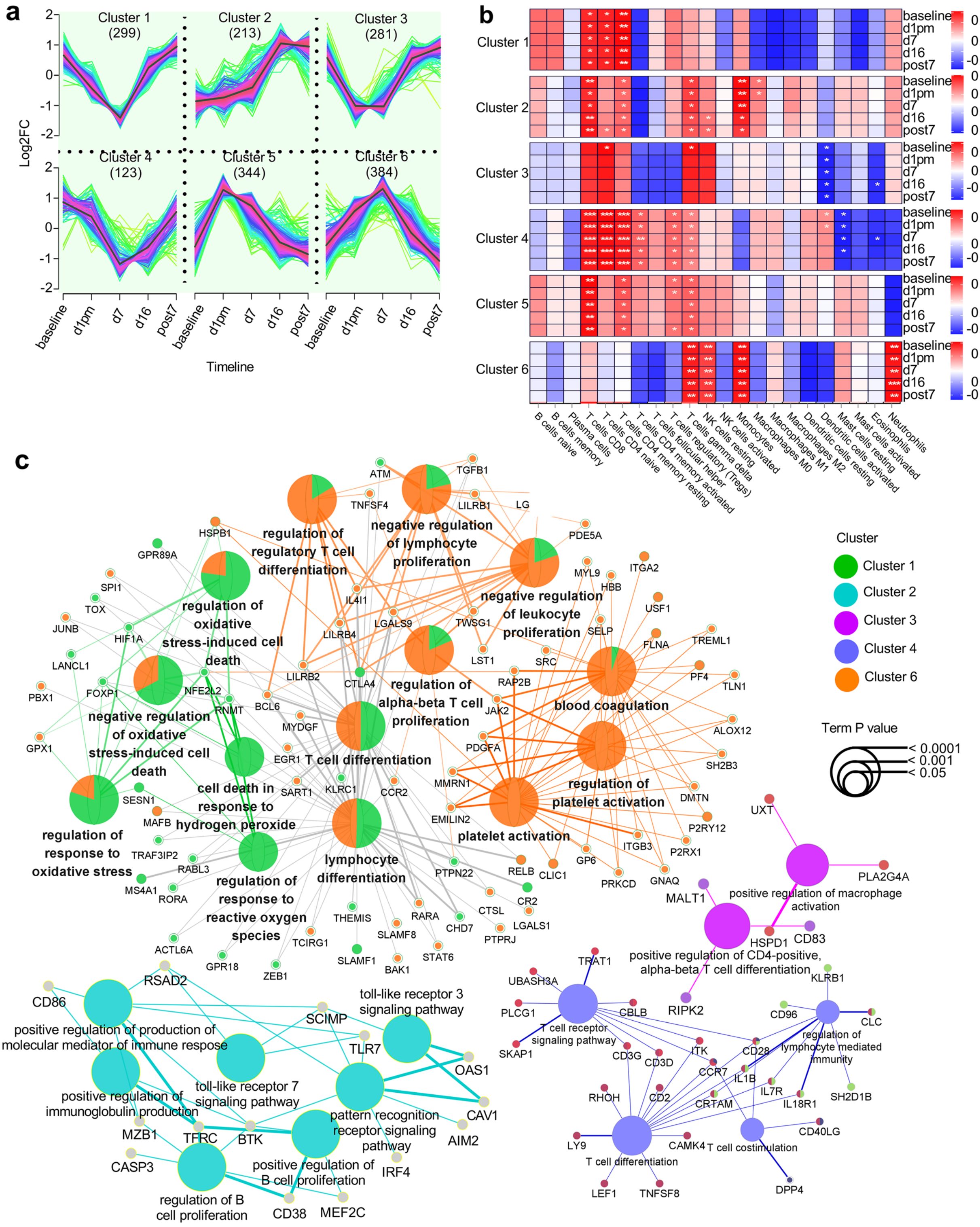

Figure 2. The genetic and immune functional responses to VLH exposure. (**a**) Timeline genetic responses of 1644 DGs in the six clusters. Gene expression changes were normalized using log2. (**b**) Dynamic leukocyte functional differentiation in each cluster. Backgrounded by LM22, each cluster was significantly annotated with genetic associations to different leukocyte types (two-tailed spearman correlation test, N = 22, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). (**c**) Functional enrichment analysis based on the GO databases. A built-in ClueGOPlugin (v.2.5.9) in the software Cytoscape (3.9.1) was applied in the GO analysis.

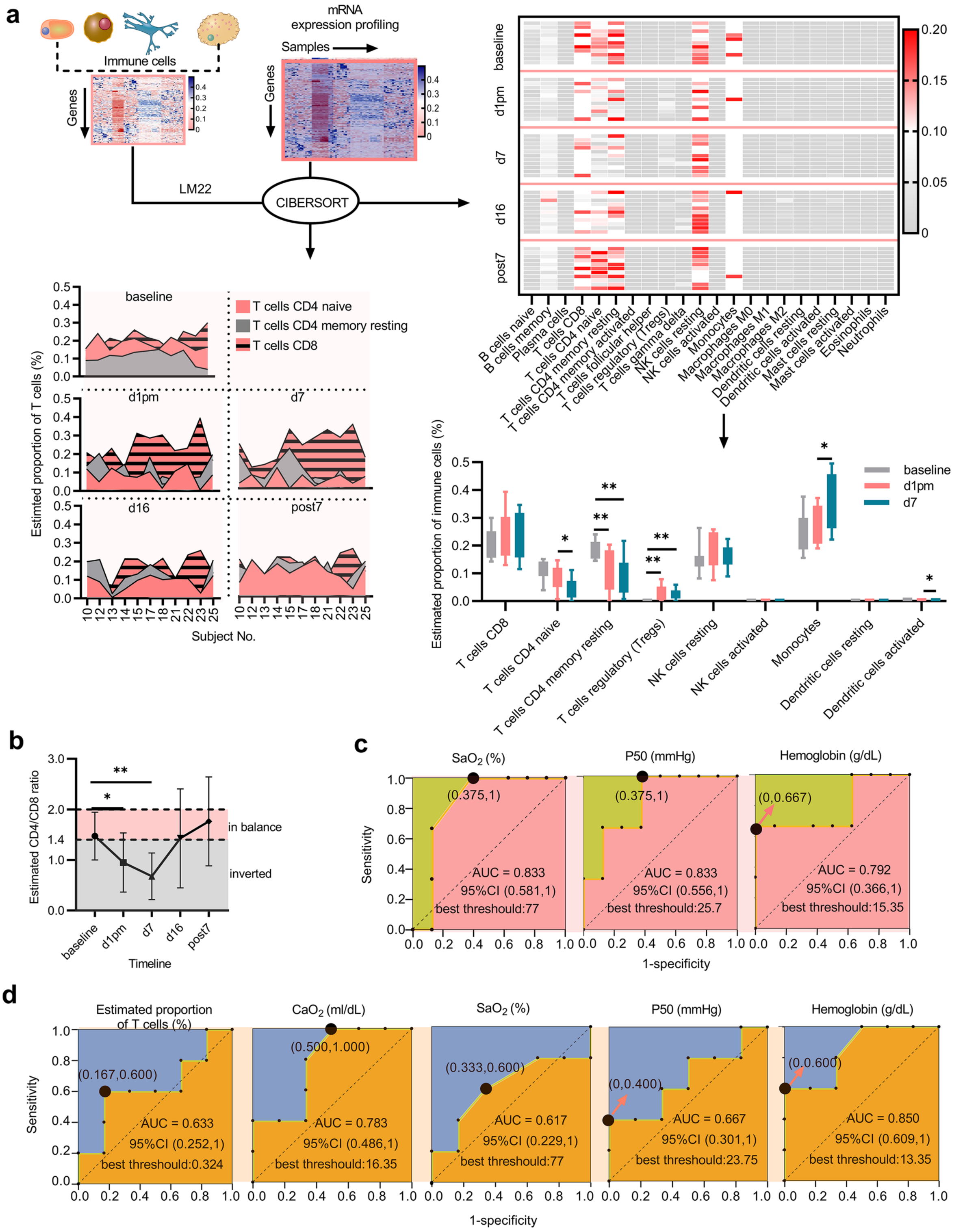

Figure 3. The potential risk factors for sAMS. (**a**) CIBERSORT algorithm coupled with LM22 revealed the timeline proportion of leukocytes in each subject. As indicated in the heatmap, resting T cells and NK cells were dominantly proportioned in subjects exposed to VLH. The proportion of CD4 T cells was predicted to be significantly decreased at d1pm and d7 compared with that at baseline. The bar chart indicates the changes in the average cell proportions across baseline, d1pm, and d7 (N = 11, paired t test, **P* < 0.05, ***P* < 0.01). The error bars indicate the 95% CI. The timeline changes in CD4 and CD8 T cells in each subject in the training cohort are displayed with the area chart. (**b**) The inverted CD4/CD8 ratio. The average estimated CD4/CD8 ratio decreased from baseline to d1pm and then increased from d7 to post7 (N = 11, paired t test, **P* < 0.05, ***P* < 0.01). (**c**) Binary classifiers of the CD4/CD8 ratio and the best thresholds. (**d**) Binary classifiers of sAMS and the best thresholds.

subgroups (Fig. 4a). Significant correlations were observed between SaO_2, P50, hemoglobin, and the CD4/CD8 ratio in the expression pattern of related DGs (Pearson correlation test, $P<0.001$). Meanwhile, as the potential risk factors of sAMS, the CD4/CD8 ratio, the estimated proportion of T cells, SaO_2, P50, hemoglobin, and CaO_2 were indicated with significant correlations with the LLQ-AMS score (Pearson correlation test, $P<0.001$). A total of 2286 risk factor-related DGs were identified, with 328 in the set of CaO_2, 346 in the CD4/CD8 ratio, 682 in hemoglobin, 964 in P50, 515 in SaO_2, 263 in the estimated proportion of T cells, and 508 in the LLQ-AMS score (Supplementary Table S5), which were intersected in 2–6 ways under Venn analysis[26] (Fig. 4b). To further identify function modules among the 2286 DGs and the relationship to the risk factors of sAMS, weighted correlation network analysis (WGCNA)[27] was performed using the values of PaO_2, $PaCO_2$, SaO_2, CaO_2, P50, hemoglobin, LLQ-AMS score, AMS-C-composite score, CD4/CD8 ratio, and the estimated proportion of T cells as a trait. A gene coexpression network was constructed, and 7 modules were identified using hierarchical clustering, dynamic tree cut, and merged dynamics tools (Fig. 4c). Next, we established module-trait relationships (Fig. 4d). The royalblue module was negatively correlated with SaO_2 but positively correlated with the CD4/CD8 ratio. Both the purple and pink modules were negatively correlated with hemoglobin but positively correlated with the LLQ-AMS score and AMS-C-composite score. Moreover, the purple module seemed to be sensitive to CaO_2; blue, to hemoglobin; turquoise, to P50; tan, yellow and turquoise, to CD4/CD8 ratio; and red and cyan, to LLQ-AMS score. The results demonstrated that SaO_2, CaO_2, P50, hemoglobin, LLQ-AMS score, AMS-C-composite score, CD4/CD8 ratio, and the estimated proportion of T cells, as potential risk factors for sAMS, may impact the disease outcome at the genetic level.

Machine learning to establish the prediction model of sAMS. To identify the marker genes of sAMS, we constructed a prediction model of sAMS using the machine learning of SVM-RFE, which consists of the classification algorithm and the feature selection algorithm wrapped around, strategized to select or remove some features from the high-dimensional feature set, and obtain the optimum feature subset from various candidate subsets generated. Therefore, SVM-RFE is actually designed to find a hyperplane of the maximized marginal distance with the best differentiating performance between the two categories of the dataset, which is represented with the weight vector W^T, feature vector X, and threshold b as follows: $W^TX+\mathrm{b}=0$ (Fig. 5a). Obviously, when $W^TX+\mathrm{b}=0$, the sum of the marginal distances from the hyperplane to the closest features (D1 + D2) is maximized; however, it would be indicated with the poorest differentiating performance and accuracy whenever $W^TX+\mathrm{b}=1$ or − 1. In machine learning of SVM-RFE (Fig. 5b), the initial features ($M=2286$) were input for classifier training, with the relevance of the n-th entry of X determined by the corresponding value W_n in W^T($\mathrm{n}=1, 2,\ldots M$). Then, in each fold ($\mathrm{k}=15$) of cross validation (CV), the concrete number of features ($\tau=30$) with the lowest absolute values of W_n were rejected. The maximum accuracy was determined by the entire feature selection and error estimation process (five-fold CV). The top 14 ranked features (Supplementary Table S6) with the highest five-fold CV accuracy (Fig. 5c) or the lowest error (Fig. 5d) were selected for further analysis. Ten of 14 with significant predictive power ($N=19$, univariate logistic regression test, $P<0.05$) (Supplementary Table S7) were used to build the model (C-index = 1, $P<0.01$) (Supplementary Table S8) and nomogram (Fig. 5e). In the training cohort of d1, the model had excellent prediction performance for sAMS as analysed using ROC (AUC = 1) (Supplementary Fig. S1a), calibration curve (Fig. 5f), and survival analysis ($R^2=5.011$, $P=0.024$) (Fig. 5g). When tested using ROC within the timeline of the training cohort over baseline (N = 21, AUC = 0.600), d7 (N = 18, AUC = 0.691), d16 (N = 21, AUC = 0.673), post7 (N = 14, AUC = 0.633), and validated in the validation cohort (N = 31, AUC = 0.626) (Supplementary Fig. S1b–f), the model was indicated with satisfactory predictive accuracy between the actual probability and the predicted probability. To assess the clinical applicability of the model, we also established a single-gene (OR10G8) model (C-index = 0.764) using the baseline data of the training cohort (Supplementary Fig. S2, Supplementary Table S9) and a three-gene model (B4GALT4, DIP2B, GALNT13) (C-index = 0.897) (Supplementary Fig. S3) based on the validation cohort data (Supplementary Table S10, S11). All the models were indicated with overall net benefits varying from 53 to 100% when assessed using decision curve analysis (DCA) (Fig. 5h).

MicroRNAs (miRs) mediated the effects of the featured genes in the development of sAMS. There are 29 homo sapiens (hsa)-miRs identified in five featured genes, including 2 from EPHB3, 3 from DIP2B, 8 from RHEBL1, 3 from GALNT13, and 13 from SLC8A2, which were targeted to 3710 miR targets (Supplementary Fig. S4, Supplementary Table S12). We next wanted to determine the biological functions of the miR targets. As expected, most targets were enriched in terms related to lymphocyte activities under GO analysis (73.69%) (Fig. 6). Furthermore, 5.26% were enriched in histamine secretion, 2.56% in erythrocyte differentiation, and 15.79% in the regulation of myeloid cell differentiation (Supplementary Table S13). Accordingly, several meaningful pathways of the featured genes were identified, including GALNT13-(hsa-miR-124-3p/506-3p)–RCOR1, SLC8A2/DIP2B-(hsa-miR-133a-3p/133b)-RVAMP2/SLC4A1 (Fig. 7a), RHEBL1-(hsa-miR-19a/b-3p)-HIF1A, and EPHB3-(hsa-miR-149-5p)-IL6.Under GO and gene set enrichment analysis (GSEA), different expression profiles of the miR targets were observed between people with and without sAMS along the timeline (Fig. 7b), with pathways related to heme metabolism (P < 0.001, normalized enrichment score (NES) = − 1.703), G2M checkpoint (P = 0.023, NES = − 1.341), and coagulation (P = 0.040, NES = − 1.328) significantly down-regulated in sAMS at d1noon, while with oxidative phosphorylation (P < 0.001, NES = 1.624), inflammatory response (P < 0.001, NES = 1.543), IL6/JAK/STAT3 signaling (P = 0.012, NES = 1.492), IFNɣ response (P = 0.007, NES = 1.516), TNFα/NFκB signaling (P < 0.001, NES = 1.661) significantly up-regulated at d1pm, when sAMS occurred in 10 of 21 subjects observed (Fig. 7c). Furthermore, most of the 14 featured genes were observed with

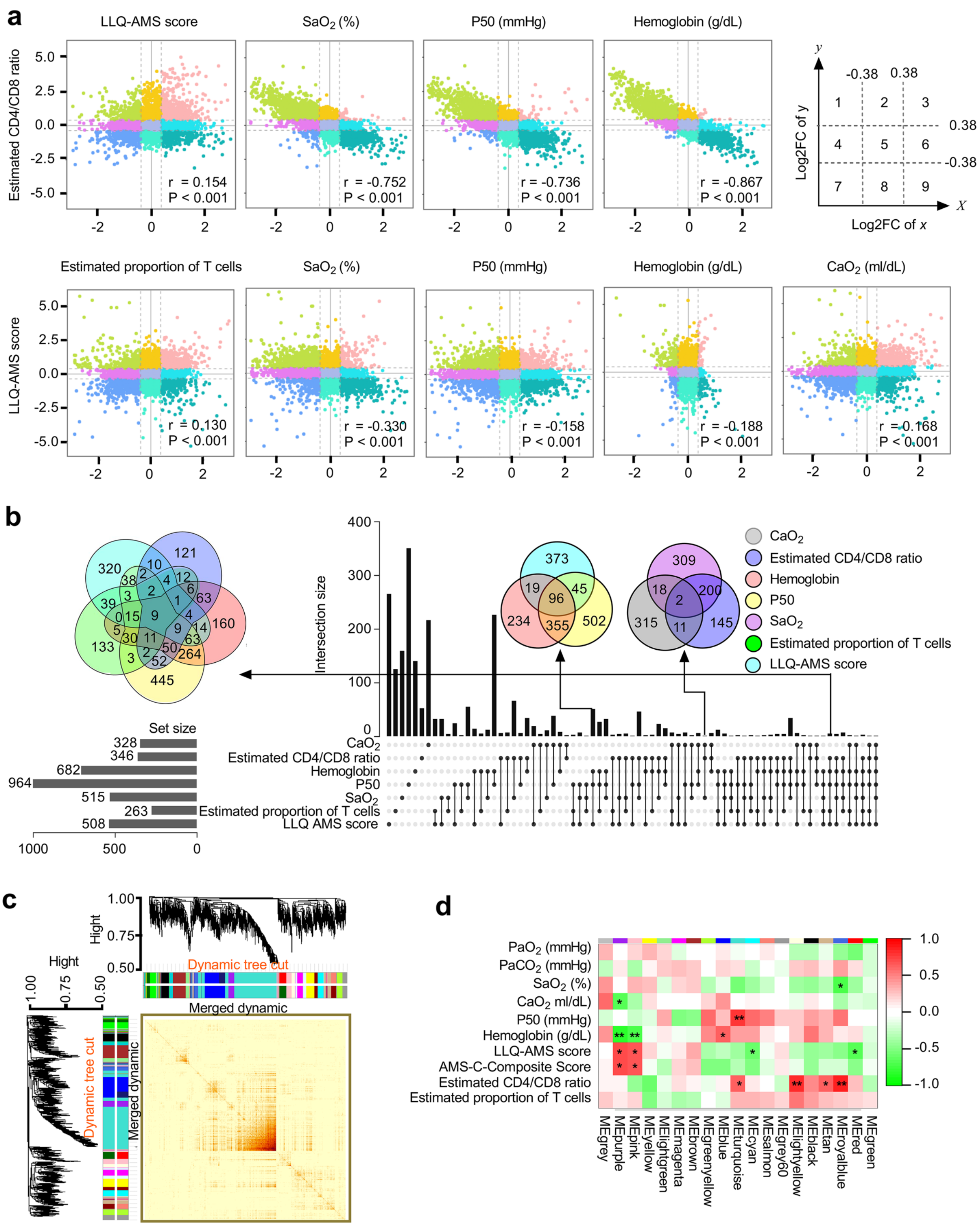

Figure 4. The gene signature underlying the phenotypes of sAMS. (**a**) Nine-quadrant diagram of the expression patterns of DGs across subgroups. The expression pattern was indicated with the log2 normalized FC. All DGs with log2FC > 0.38 or < -0.38 were visualized. Quadrant 1, DGs upregulated in classifier y but downregulated in x; quadrant 9, upregulated in x but downregulated in y; quadrant 3, upregulated in both x and y; quadrant 7, downregulated in both x and y. No significant correlations were observed in quadrants 2, 4, 5, 6 and 8. A total of 2286 DGs with log2FC > 1.3 or < -1.3 were selected for further analysis. (**b**) Venn analysis. The 2–6 intersections of DGs are indicated in the UpSet Venn diagram. The intersection size was displayed using upset bars and 3-way or 6-way Venn diagrams. (**c**) Network heatplot of the selected genes; (**d**) Module-trait relationships (Pearson correlation test, *$P < 0.05$, **$P < 0.01$).

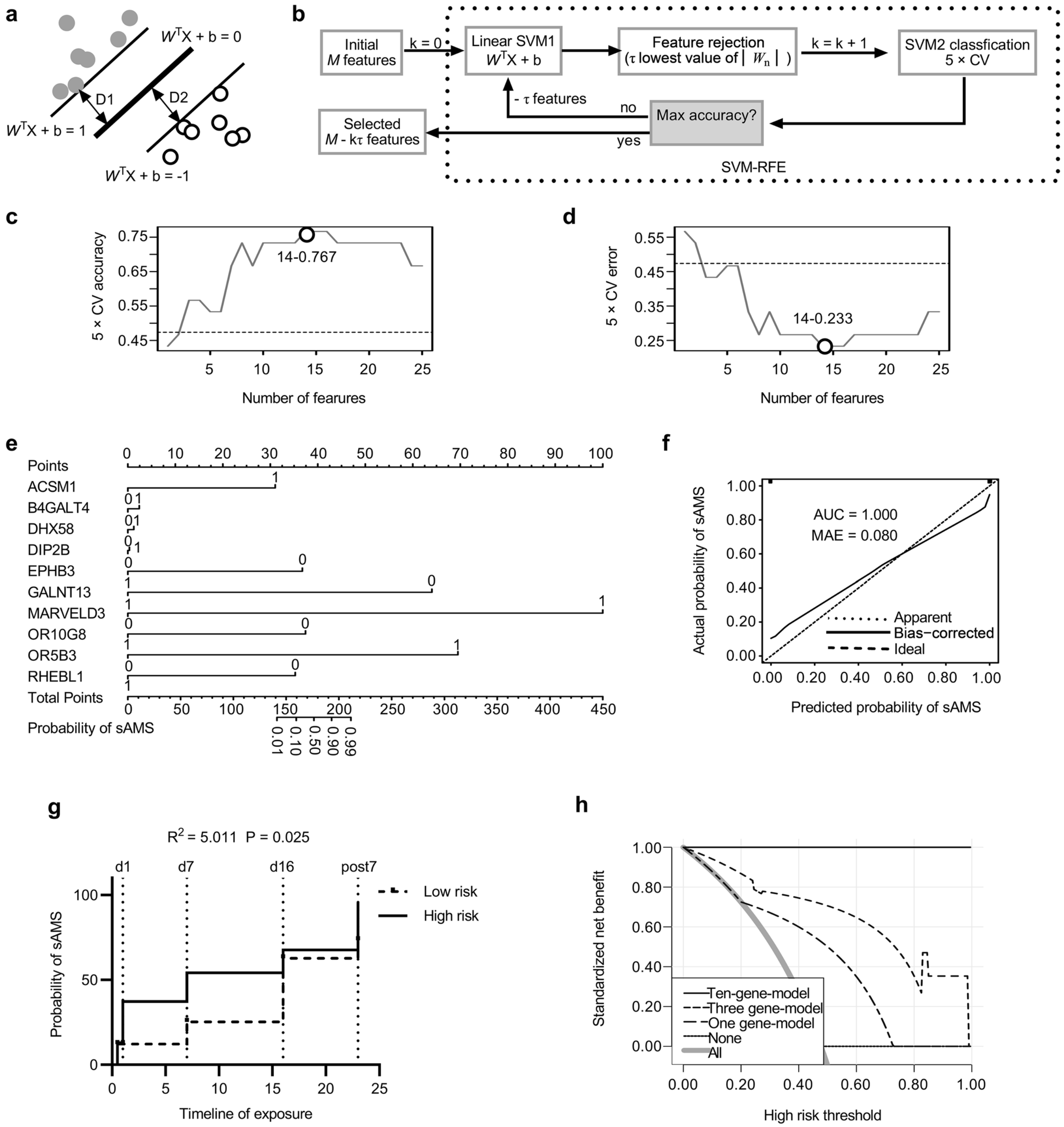

Figure 5. Establishing the model of genetic susceptibility to sAMS. (**a**) Hyperplane of SVM-RFE. Thicker line, hyperplane; thinner lines, margin limits; W^T, the weight vector; *X*, the feature vector; b, the threshold; D1 and D2, marginal distances. (**b**) SVM-RFE method. The SVM1 parameters (absolute values of W_n) were computed to determine the relevance of all the input features, and then the entire feature selection and error estimation process were performed during SVM2 classification as a guide to choose the optimal number of features. *M*, number of input features; k, number of folds of CV; τ, number of features to be rejected. (**c**,**d**) The estimated accuracy and error. (**e**) A nomogram for the prediction of sAMS upon rapid exposure to VLH. (**f**) The calibration performance of the model. (**g**) Survival analyses for the subjects in the training cohort. (**h**) DCA of the three models. The high-risk threshold was predicted as 47%.

different expressions in sAMS vs. non-sAMS at d1noon, d1pm and/or d7 (Fig. 7d). All the featured genes and their miR targets were functionally related to erythrocyte differentiation, alpha–beta T cell differentiation, and histamine secretion by mast cells (Fig. 8). These results suggested the important roles of the featured genes in sAMS, which were mediated by miRs and their downstream targets.

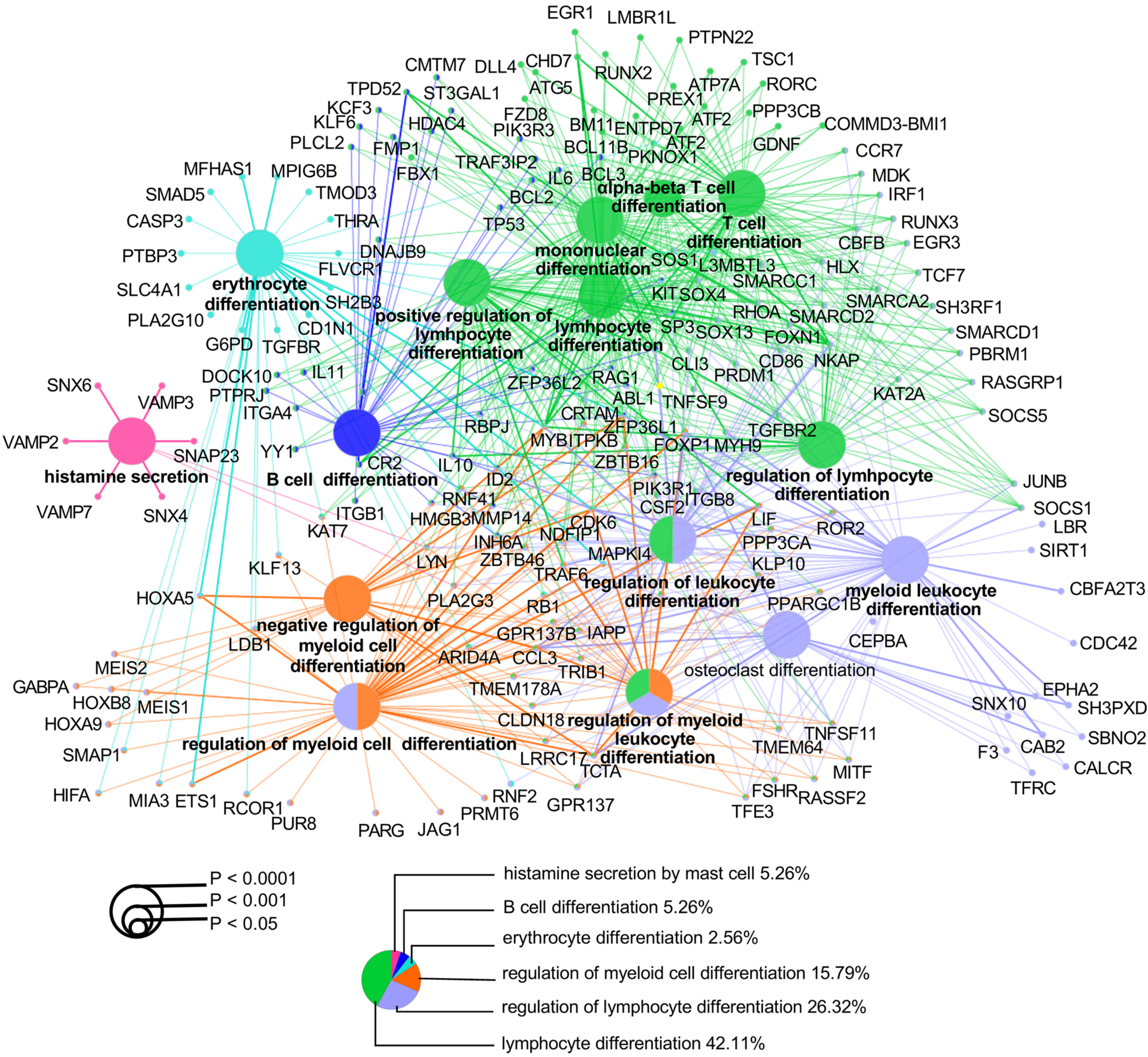

Figure 6. Functional enrichment of the miR targets.

Discussion

The study was based on the microarray dataset abstracted from GSE103927[21], a well-established dataset including 112 PBMC samples from 21 subjects who were rappidly exposed or re-exposed to the very high altitude of 5260 m after multiple periods of hypoxia acclimatization varying from 48 h to 21 days. Data at baseline, d1pm, d7, d16, and post7 from 21 subjects who completed all the planned tests were extracted for further analysis. On the first day of exposure, 10 of 21 were diagnosed with sAMS (LLQ-AMS score ≥ 6), a severe condition implying the possibility of life-threatening events. Although all of them recovered from a 3-night acclimatization at 3800 m followed by a prolonged stay at 5260 m for 13 days (d16), we still wondered why some of them were at risk of sAMS, but others were not, even under almost the same exposure conditions. In this study, the genetic basis underlying the pathological and physiological responses to VLH exposure was investigated to identify those who are vulnerable to sAMS or related events.

The gene expression patterns at d1pm and d7 varied from baseline but recovered after several acclimatization days at d16 and post7 (Fig. 1b), suggesting genetic responses upon acute VLH exposure. Most DGs across baseline, d1, and d7 (Fig. 1c) were involved in immune cell activation (Fig. 1d), with a continuous upregulation from baseline to the peak of d7 in those related to leukocyte activation upon immune response (Fig. 1e). Similar patterns of immune activation triggered by high-altitude exposure (3232 m) were also observed in another study[28], with immune responses sensitized at the early phase of high-altitude exposure. Furthermore, peak changes in clustered DGs were observed at d7 (Fig. 2a), with functions related to T cells (gamma delta, CD8, CD4 naive, CD4 memory resting, and CD4 memory activated) dominating in DG clusters (Fig. 2b). Accordingly, these up- or downregulated DGs were functionally related to platelet activation, oxidative stress, and/or T cell differentiation

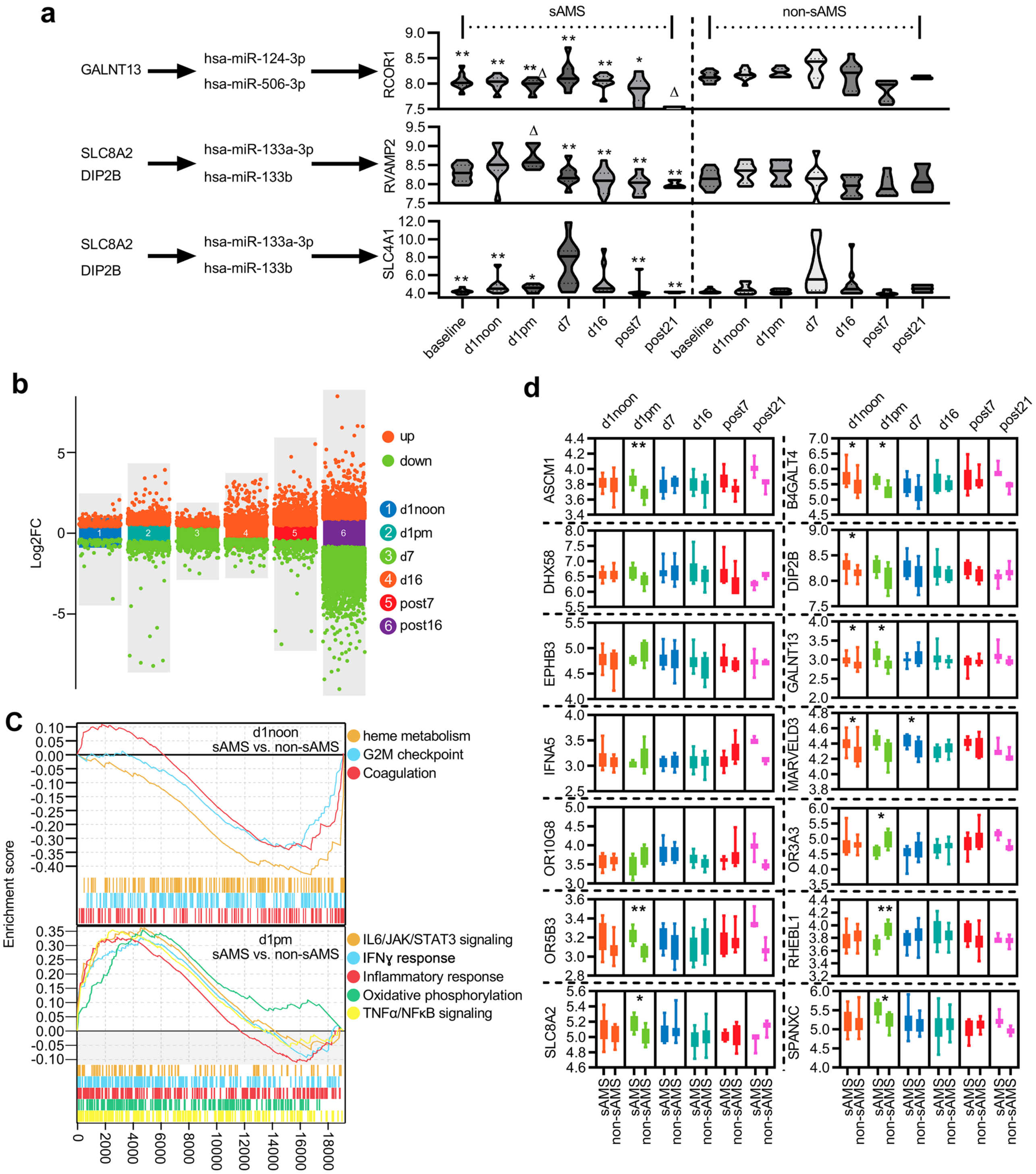

Figure 7. Changes of the featured genes and their miR targets in sAMS. (**a**) Timeline changes of miR targets. The miR products hsa-miR-124-3p/506-3p of GALNT13 mediated the expression changes in their target RCOR1 in sAMS, which peaked at d7 (N = 9) and recovered at post21 (N = 3). Compared with post21, **P* < 0.05, ***P* < 0.01; in comparison with non-sAMS, $^{\Delta}P < 0.05$ (Tukey's multiple comparisons test). As the shared miR products of SLC8A2 and DIP2B, hsa-miR-133a-3p/133bregulate the expression of their targets, RVAMP2 and SLC4A1, which were upregulated at d1pm (N = 6) or d7 (N = 9). **P* < 0.05, ***P* < 0.01 vs. d1pm or d7. $^{\Delta}P < 0.05$ vs. non-sAMS (Tukey's multiple comparisons test). (**b**) Between-group changes of miR targets. (**c**) GSEA analysis. (**d**) Between-group changes of the 14 featured genes. The featured genes were compared in the relative expression level between subjects with (N = 9 (d1noon), 6 (d1pm), 9 (d7), 10 (d16), 7 (post7), 3 (post21)) and without sAMS (N = 10 (d1noon), 6 (d1pm), 9 (d7), 11 (d16), 7 (post7), 3 (post21)). Compared with non-sAMS, **P* < 0.05, ***P* < 0.01 (Wilcox-test).

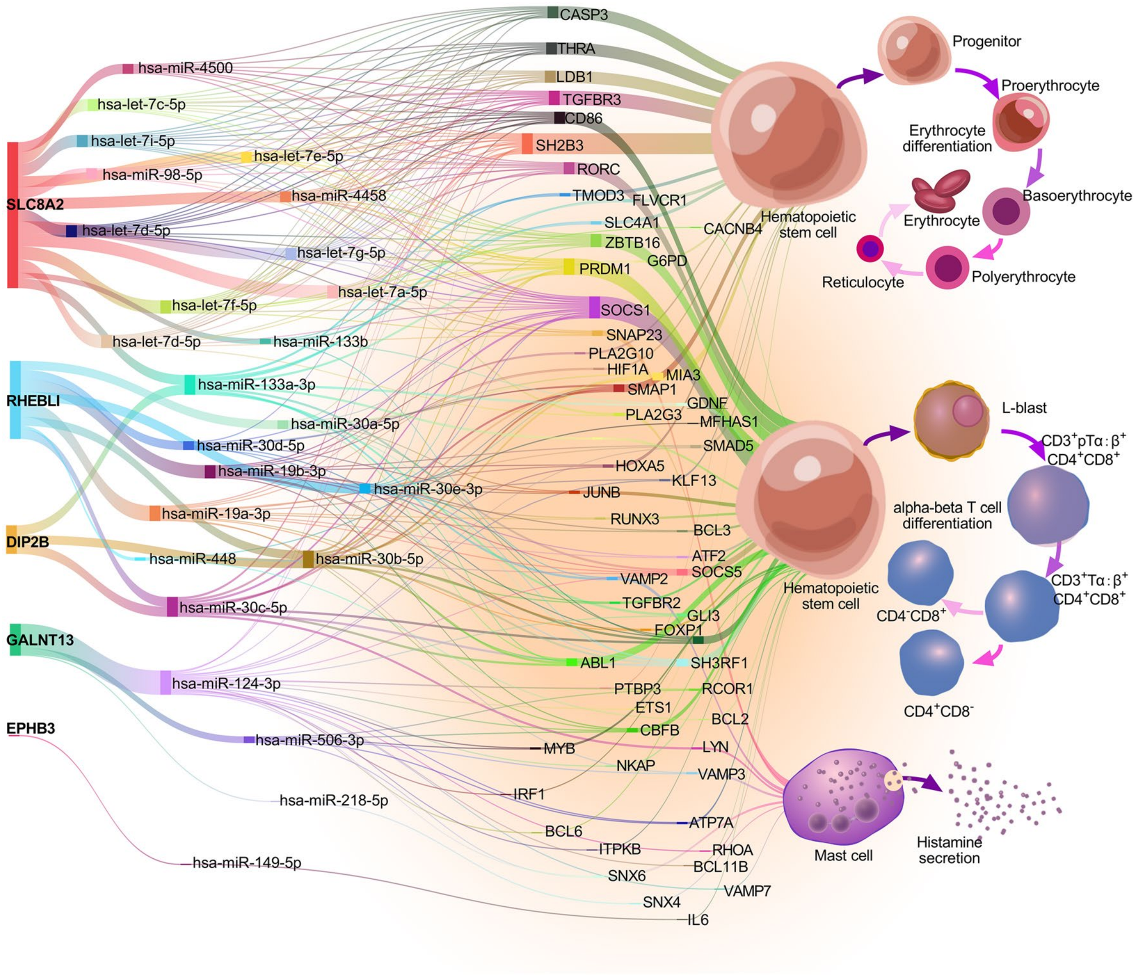

Figure 8. The identified pathways related to erythrocyte differentiation, alpha-bata T cell differentiation, and histamine secretion by mast cells.

(Fig. 2c), which are considered essential to the occurrence of AMS, the subsequent development of sAMS, and/or related events[29–33]. These results implied that T cells dominated the genetic responses to VLH exposure.

The proportion of CD4 T cells was indicated by a significant decrease in the subjects from baseline to d7 (Fig. 3a) and the inverted CD4/CD8 ratio (Fig. 3b), which was also reported in other high-altitude populations[33], reminding us of the enhanced immunity and susceptibility to sAMS. Interestingly, similar timeline changes were observed in the CD4/CD8 ratio, LLQ-AMS score, and other laboratory values detected, all with peaked values at d1pm or d7, then recovered from d7 to d16 (Supplementary Fig. S5), implying their contributions to sAMS as risk factors. Then, the CD4/CD8 ratio, SaO_2, CaO_2, P50, hemoglobulin, and the estimated proportion of T cells were further investigated for the interplay effects between them, aiming to identify the underlying risk genes of sAMS (Fig. 3c,d). A total of 2286 risk genes were mapped (Fig. 4) for classifier training, and 14 gene classifiers were identified to establish the model using SVM-RFE (Fig. 5a–d). We established a ten-gene model of genetic susceptibility to sAMS (Fig. 5e) with excellent discrimination (C-index = 1, AUC = 1) and satisfactory predictive accuracy as assessed using ROC and survival analysis (Fig. 5f,g). We also constructed one-gene and three-gene models. All the models were indicated to have good clinical applicability as assessed by the overall net benefits over risks (Fig. 5h), suggesting the roles of the modeled genes as predictive markers for sAMS.

Limited evidence has indicated that certain miRs may function as biomarkers for AMS[34,35] or play roles in acute hypoxia and hypoxia-induced pulmonary vascular leakage[36]. In subjects exposed to a height of 3100 m, miR-424 was overexpressed in a HIF1A-dependent manner, which in turn can stabilize HIF1A. In our study, 29 miRs and 3710 miR targets were identified from five genes (EPHB3, DIP2B, RHEBL1, GALNT13, and SLC8A2) (Supplementary Fig. S4), which were associated with multiple biological processes, as evidenced by GO analysis (Fig. 6). We identified 260 important miR-mediated signalling pathways concerning erythrocyte differentiation, alpha–beta T cell differentiation, and histamine secretion (Fig. 8). As one of the sodium-calcium exchangers, SLC8A2 has been previously shown to be a nuclear translocation regulator of HIF1A[37], which was significantly downregulated upon SLC8A2 overexpression[38]. Our study indicated that SLC8A2 acts upstream of multiple hypoxia- and/or altitude-sensitive miR targets, such as RCOR1 (a transcription rheostat essential for normal myeloerythroid lineage differentiation)[39,40], PRDM1 and LDB1 (both are involved in high-altitude adaptation)[41,42], and CASP3 (a member of the hypoxia-activated mitochondrial apoptosis pathway)[43]. We noticed that both hsa-miR-133a-3p and hsa-miR-133b-3p mediate the signals from SLC8A2 or DIP2B to SLC4A1, a biomarker of AMS, which was correlated with various AMS symptoms and plays important roles in CO_2 gas transport in erythrocytes[44]. The shared miR targets of SLC8A2 and DIP2B also include TMOD3, which has been shown to be a candidate biomarker for high-altitude pulmonary hypertension in Kyrgyz highlanders[45]. Interestingly, GALNT13 has been previously identified as a risk gene relevant to sickle cell disease-associated pulmonary hypertension, which may play roles in endothelial permeability[46,47]. Our results showed that GALNT13 interacts with multiple miR targets related to histamine secretion and hypoxia-induced activities in erythrocytes and T cells, suggesting its potential effects in pulmonary vascular and ventricular injuries. More importantly, as miR products of RHEBL1 (a member of Ras superfamily)[48], both hsa-miR-19a-3p and hsa-miR-19b-3p were indicated as mediators of HIF1A[49,50], suggesting important roles in hypoxia-related biological processes. Furthermore, the EPHB3 (a proliferation suppressor in ambient and hypoxic environments)[51]-(hsa-miR-149-5p)-IL6 pathway was also believed to be essential for hypoxic responses as the underlying association between hypoxia and inflammation. We failed to identify miR products or miR-mediated signals related to the other 9 featured genes, whereas some of them were previously argued to be hypoxia-sensitive genes, such as ACSM1, a member of the lipoic acid salvage pathway controlling HIF1 activation[52,53]. Obviously, as potential predictors or biomarkers of sAMS, the 14 featured genes still remain far from being uncovered regarding their roles and mechanisms in the development of sAMS.

Because of the infeasibility, it is almost impossible to conduct large-scale trials at high altitudes, especially under extreme conditions such as VLH. In our study, data of very small number of cases were used, which may result in biased machine learning performance. SVM seems good at dealing with small samples and large numbers of features, but the use of CV methods seems not sufficient to control overfitting[54]. Furthermore, higher dimensional interactions should be considered in machine learning, as it is possible that linear kernels based SVMs may not capture all non-linearity inherently. Though, in our study, completely separating validating and training data were applied and satisfactory prediction accuracy of the model was abtained, other model selection strategies may have superiority over SVM, like StepSVM[55], Random Forest[56], and Xgboost[57]. Fortunately, part of the selection results were repeated when using bagging-based Random Forest, and boosting-based Xgboost instead of SVM (Supplementary Fig. S6), suggesting the rationality of SVM-RFE strategy applied in sAMS prediction. We have explored the interplay effects between the laboratory values and LLQ-AMS score at the genetic level. It was indicated that SaO_2 (AUC = 0.617), CaO_2 (AUC = 0.783), P50 (AUC = 0.667), hemogbin (AUC = 0.850) functioned as possible binary classifiers for sAMS, suggesting their roles in determining sAMS. However, for inconsistent testing conditions and findings, these laboratory values may still be not available in the prediction of sAMS for prevention purposes.

This study was based on microarray data from 112 PBMC samples of 21 subjects exposed to VLH and aimed to explore the genetic susceptibility of sAMS. Using the machine learning of SVM-RFE, we identified 14 classifier genes and established a prediction model of sAMS, which performed well in predicting or differentiating subjects suffering from sAMS, and hold promise to be clinically applied in the early screening for sAMS risks. However, more studies are still needed to corroborate the existing findings related to the predictive or differentiating power of the model and to establish the role of the modeled genes as biomarkers for sAMS.

Methods

Collection and preprocessing of VLH microarray data.

VLH microarray data were explored in Gene Expression Omnibus[58] using the keywords "AMS" and "high altitude" and were collected from the platform GPL6244 in MINiML format under the accession numbers GSE103927 and GSE52209. GSE103927 was structured on 112 PBMC samples from 21 subjects exposed to sea level then VLH at seven time points, including baseline, the first day noon (d1noon) or post meridiem (d1pm), day 7 (d7), day 16 (d16), post7 or post21. Information concerning age, sex, height, detected levels of PaO_2, $PaCO_2$, SaO_2, CaO_2, P50, hemoglobin, LLQ-AMS score, and AMS-C-Composite score were included (Supplementary Table S14). Data of d1 were used to train and establish the model. Data from baseline, d7, d16, and post7 were used to test the model. Isolated data from GSE52209 were applied to validate the model. All the raw data were extracted and preprocessed using the package in RStudio (2022.07.1 + 554) and then normalized by log2 transformation using the normalize quantiles function of the preprocessCore package. The normalized data were annotated in GPL6244 for the conversion of all probes into gene symbols. Probes mapping to multiple genes were filtered out. The final gene expression value was determined by the mean over multiple detection. The removeBatchEffect function of the limma package was applied to remove the batch effects. PCA was performed prior to further analysis of all the data, to determine whether gene expression changes over time or whether timeline datasets depend on one another. No live human/blood sample were used for the study.

Identification of DGs.

DGs across baseline, d1pm and d7 in the training cohort were identified using the Limma package (version: 3.52.3) in R software. False positive results were corrected via P-value adjustment. The thresholds for the screening of DG mRNAs were defined as $P < 0.05$ and $\log2FC > 1.3$ or $\log2FC < -1.3$.

Function and pathway enrichment.

GO analysis, GSEA and their visualization were performed using the ClueGo application in Cytoscape software (version 3.9.1)[59]. Various types of evidence (experimental, computational, author statement from publication, and curatorial statement) were used for analysis. Ontologies, pathways, and annotation files were updated before each analysis using the GO annotation database (UniProt-GOA)[60]. To identify the representative pathways, medium network specificity was selected, and GO levels varying from 3 to 8, with a minimum of 3 genes per term and at least 4% of the total associated genes, were mapped.

The GO term fusion threshold was 50% for group merging. Only terms with a P value < 0.05 were displayed with statistical significance.

Cluster analysis.

DGs across baseline, d1pm, and d7 in the training cohort were explored for functional differentiation of leukocytes using fuzzy c-means clustering. The mean expression value of DGs at each time point was calculated using the avereps function in the limma package. The expression patterns along the exposure timeline were detected with the Mfuzz package. The filter threshold for the expression value was 0.25. According to the results of multiple rounds of training, the number of clusters was expected to be 6. Each cluster was compared in relation to immune signatures using GO analysis and LM22, a leukocyte gene signature matrix containing a total of 547 leukocyte markers[24]. Pearson's test was used to estimate the timeline correlation of each cluster to LM22 signatures.

Cibersort algorithm for the abundance of leukocyte types in PBMC samples.

Coupled with LM22, the cibersort algorithm was used to distinguish 22 types of leukocytes, including B cells, plasma cells, T cells, NK cells, monocytes, macrophages, dendritic cells, mast cells, eosinophils, neutrophils, and the subtypes described above[25]. The timeline expression data from 112 samples were extracted for analysis. To improve the accuracy of the deconvolution algorithm, 1000 permutations from the default signature matrix were applied to compute the P value and root-mean-square deviation for samples at each time point. The scores for each signature were summarized and median centered to permit timeline comparisons. The paired T test was used to discover the significance between comparisons, with parametric test performed to assume Gaussian distribution. The total estimated sum of CD4 T cells (naive, memory resting, and memory activated) and CD8 T cells was used to calculate the CD4/CD8 ratio. All T cell subtypes were used to estimate the total proportion of T cells. ROC was used to predict the best thresholds of SaO_2, P50, hemoglobin, CaO_2, and estimated proportion of T cells, as binary classifiers of CD4/CD8 ratio, and sAMS (LLQ-AMS score ≥ 6).

Risk gene mapping.

The interplay effects between the estimated CD4/CD8 ratio, LLQ-AMS score, SaO_2, P50, hemoglobin, CaO_2, estimated proportion of T cells, and underlying hub genes were visualized using a nine-quadrant diagram and Venn diagram[26]. WGCNA[27] was conducted to discover the relationships between gene expression patterns and phenotypes. Genes with expression values above 1 were applied for further analysis. The soft power was estimated at 7. An unsigned scale-free coexpression network for the genes was constructed using the minimum module size (minModuleSize) of 100 and the threshold of 0.25 for merging of modules. Pearson's correlation value was applied to establish the similarity matrix, adjacency matrix, and topological overlap matrix between each pair of genes across all samples. The gene coexpression module was detected using the dynamic tree cut algorithm and was constructed with a cut height of 0.975.

Machine learning to establish the prediction model of sAMS.

The phenotype-validated DGs were input into the SVM-RFE system for machine learning, which was run within the e1071 and msvmRFE packages. SVM-RFE is an SVM-based iterative algorithm that works backward from an initial set of features and is applied to find the optimal hub gene by deleting feature vectors. The input data were from all samples of the training cohort of d1, containing 19 observations (individual subjects) and 2286 features (DGs). We used k = 15 for the k-fold CV and halve.above = 30 to cut the features in half each round until there were fewer than 30 remaining features. The entire feature selection and generalization error estimation process was wrapped for five-fold CV. Feature ranking was performed using the lapply function based on the average rank across the five folds of accuracy and error estimation. Univariate Cox hazard analysis was applied to assess the prediction performance of the selected features. Multinomial logistics regression, nomogram, survival analysis, AUC, and calibration curve were used to establish, test, and validate the model. DCA was used to assess the clinical applicability of the prediction model. The model was tested within the timeline of the training cohort, including baseline, d7, d16, and post7, and validated in another cohort, which comprised 17 subjects who developed high-altitude pulmonary edema within 48–72 h after exposure to VLH, 14 normal controls, and 14 high-altitude natives (GSE52209)[23]. R packages including Hmisc, lattice, survival, Formula, ggplot2, rmda, ggDCA, rms, SparseM, caret, and pROC were applied in the model development.

MiR analysis.

To further explore the roles of the 14 featured genes in the development of sAMS, miRs and their targets were predicted using the miR function of FunRich (version 3.1.3). Timeline changes in miR targets in sAMS were analysized using ordinary one-way analysis of variance. Tukey's multiple comparisons were performed to compare the mean expression of each targets with the mean value of other targets. Testing the homogeneity of variances (equality variances) among different time points was done using the Brown–Forsythe test. All the miR targets were compared in the expression profiles between sAMS and non-sAMS using GO and GSEA tools. The between-group changes of the featured genes were analyzed using Wilcox-test, with the hypothetical value set as zero. All the values matching zero in the datasets were entirely ignored in the analysis. All the methods used in the study were outlined in the flowchart (Supplementary Fig. S7).

Data availability

Data sets described in this article are available from Gene Expression Omnibus (http://www.ncbi.nih.gov/geo) via the accession number GSE103927 and GSE52209. All data generated in this article are freely accessible in supplementary tables to any scientist wishing to use them noncommercially. On reasonable request, the corresponding author can provide further information.

Received: 13 January 2023; Accepted: 17 March 2023
Published online: 21 March 2023

References

1. Berger, M. M., Sareban, M. & Bärtsch, P. Acute mountain sickness: Do different time courses point to different pathophysiological mechanisms?. *J. Appl. Physiol.* **128**, 952–959 (2020).
2. Luks, A. M. & Hackett, P. H. Medical conditions and high-altitude travel. *N. Engl. J. Med.* **386**, 364–373 (2022).
3. Turner, R. E., Gatterer, H., Falla, M. & Lawley, J. S. High-altitude cerebral edema: Its own entity or end-stage acute mountain sickness?. *J. Appl. Physiol.* **131**, 313–325 (2021).
4. Swenson, E. R. Early hours in the development of high-altitude pulmonary edema: Time course and mechanisms. *J. Appl. Physiol.* **128**, 1539–1546 (2020).
5. Cobb, A. B. *et al.* Physiological responses during ascent to high altitude and the incidence of acute mountain sickness. *Physiol. Rep.* **9**, e14809 (2021).
6. Douglas, D. J. & Schoene, R. B. End-tidal partial pressure of carbon dioxide and acute mountain sickness in the first 24 hours upon ascent to Cusco Peru (3326 meters). *Wilderness Environ. Med.* **21**, 109–113 (2010).
7. Burtscher, M. *et al.* Physiological responses in humans acutely exposed to high altitude (3480 m): Minute ventilation and oxygenation are predictive for the development of acute mountain sickness. *High Alt. Med. Biol.* **20**, 192–197 (2019).
8. Mazur, K., Machaj, D., Jastrzębska, S., Płaczek, A. & Mazur, D. Prediction of the development and susceptibility to acute mountain sickness (AMS) by monitoring oxygen saturation (SpO_2)—Literature review. *J. Educ. Health Sport* **10**, 79–84 (2020).
9. Duffin, J., Hare, G. M. & Fisher, J. A. A mathematical model of cerebral blood flow control in anaemia and hypoxia. *J. Physiol.* **598**, 717–730 (2020).
10. Dominelli, P. B. *et al.* Dissociating the effects of oxygen pressure and content on the control of breathing and acute hypoxic response. *J. Appl. Physiol.* **127**, 1622–1631 (2019).
11. Zubieta-Calleja, G. R. & Zubieta-DeUrioste, N. High altitude pulmonary edema, high altitude cerebral edema, and acute mountain sickness: An enhanced opinion from the high Andes–La Paz, Bolivia 3,500 m. *Rev. Environ. Health* (2022).
12. Small, E. *et al.* Predictive capacity of pulmonary function tests for acute mountain sickness. *High Alt. Med. Biol.* **22**, 193–200 (2021).
13. Minder, L. *et al.* Cardiopulmonary response to exercise at high altitude in adolescents with congenital heart disease. *Congenit. Heart Dis.* **16**, 597–608 (2021).
14. Georges, T. *et al.* Contribution of hypoxic exercise testing to predict high-altitude pathology: A systematic review. *Life* **12**, 377 (2022).
15. MacInnis, M. J. & Koehle, M. S. Evidence for and against genetic predispositions to acute and chronic altitude illnesses. *High Alt. Med. Biol.* **17**, 281–293 (2016).
16. MacInnis, M. J., Wang, P., Koehle, M. S. & Rupert, J. L. The genetics of altitude tolerance: The evidence for inherited susceptibility to acute mountain sickness. *J. Occup. Environ. Med.* **53**, 159–168 (2011).
17. Ding, H. *et al.* Polymorphisms of hypoxia-related genes in subjects susceptible to acute mountain sickness. *Respiration* **81**, 236–241 (2011).
18. Liu, Z., Chen, H., Xu, T., Wang, X. & Yao, C. HSPA1A gene polymorphism rs1008438 is associated with susceptibility to acute mountain sickness in Han Chinese individuals. *Mol. Genet. Genom. Med.* **8**, e1322 (2020).
19. Yu, J. *et al.* Analysis of high-altitude syndrome and the underlying gene polymorphisms associated with acute mountain sickness after a rapid ascent to high-altitude. *Sci. Rep.* **6**, 38323 (2016).
20. Yu, J. *et al.* EDN1 gene potentially involved in the development of acute mountain sickness. *Sci. Rep.* **10**, 5414 (2020).
21. Subudhi, A. W. *et al.* AltitudeOmics: The integrative physiology of human acclimatization to hypobaric hypoxia and its retention upon reascent. *PLoS One* **9**, e92191 (2014).
22. Sanz, H., Valim, C., Vegas, E., Oller, J. M. & Reverter, F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinform.* **19**, 432. https://doi.org/10.1186/s12859-018-2451-4 (2018).
23. Tomar, A., Malhotra, S. & Sarkar, S. Polymorphism profiling of nine high altitude relevant candidate gene loci in acclimatized sojourners and adapted natives. *BMC Genet.* **16**, 112. https://doi.org/10.1186/s12863-015-0268-y (2015).
24. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Method.* **12**, 453–457 (2015).
25. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
26. Jia, A., Xu, L. & Wang, Y. Venn diagrams in bioinformatics. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbab108 (2021).
27. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 1–13 (2008).
28. Feuerecker, M. *et al.* Immune sensitization during 1 year in the Antarctic high-altitude Concordia Environment. *Allergy* **74**, 64–77 (2019).
29. Lackermair, K. *et al.* Combined effect of acute altitude exposure and vigorous exercise on platelet activation. *Physiol. Res.* **71**, 171 (2022).
30. Lackermair, K. *et al.* Effect of acute altitude exposure on serum markers of platelet activation. *High Alt. Med. Biol.* **20**, 318–321 (2019).
31. Pena, E., El Alam, S., Siques, P. & Brito, J. Oxidative stress and diseases associated with high-altitude exposure. *Antioxidants* **11**, 267 (2022).
32. Liu, B. *et al.* IL-10 dysregulation in acute mountain sickness revealed by transcriptome analysis. *Front. Immunol.* **8**, 628 (2017).
33. Bai, J., Li, L., Li, Y. & Zhang, L. Genetic and immune changes in Tibetan high-altitude populations contribute to biological adaptation to hypoxia. *Environ. Health Prev. Med.* **27**, 39–39 (2022).
34. Liu, B. *et al.* A signature of circulating microRNAs predicts the susceptibility of acute mountain sickness. *Front. Physiol.* **8**, 55 (2017).
35. Huang, H. *et al.* The role of salivary miR-134-3p and miR-15b-5p as potential non-invasive predictors for not developing acute mountain sickness. *Front. Physiol.* **10**, 898 (2019).
36. Tsai, S.-H. *et al.* Roles of the hypoximir microRNA-424/322 on acute hypoxia and hypoxia-induced pulmonary vascular leakage. *Available at SSRN 3221410* (2018).
37. Liu, H., Yu, J., Yang, L., He, P. & Li, Z. NCX2 regulates intracellular calcium homeostasis and translocation of HIF-1α into the nucleus to inhibit glioma invasion. *Biochem. Genet.* https://doi.org/10.1007/s10528-022-10274-9 (2022).
38. Qu, M. *et al.* The candidate tumor suppressor gene SLC8A2 inhibits invasion, angiogenesis and growth of glioblastoma. *Mol. Cells* **40**, 761–772 (2017).
39. Rivera, C. *et al.* Unveiling RCOR1 as a rheostat at transcriptionally permissive chromatin. *Nat. Commun.* **13**, 1–15 (2022).
40. Yao, H., Goldman, D. C., Fan, G., Mandel, G. & Fleming, W. H. The corepressor Rcor1 Is essential for normal myeloerythroid lineage differentiation. *Stem Cells (Miamisburg)* **33**, 3304–3314 (2015).
41. Stobdan, T. *et al.* New insights into the genetic basis of Monge's disease and adaptation to high-altitude. *Mol. Biol. Evol.* **34**, 3154–3168 (2017).
42. Jin, M. *et al.* Selection signatures analysis reveals genes associated with high-altitude adaptation in Tibetan Goats from Nagqu, Tibet. *Animals* **10**, 1599. https://doi.org/10.3390/ani10091599 (2020).

43. Hou, Y. *et al.* Establishment and evaluation of a simulated high-altitude hypoxic brain injury model in SD rats. *Mol. Med. Rep.* **19**, 2758–2766 (2019).
44. Yang, J. *et al.* Proteomic and clinical biomarkers for acute mountain sickness in a longitudinal cohort. *Commun. Biol.* **5**, 548. https://doi.org/10.1038/s42003-022-03514-6 (2022).
45. Iranmehr, A. *et al.* Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders. *Eur. J. Hum. Genet.* **27**, 150–159 (2019).
46. Desai, A. A. *et al.* A novel molecular signature for elevated tricuspid regurgitation velocity in sickle cell disease. *Am. J. Respir. Crit. Care Med.* **186**, 359–368 (2012).
47. Maron, B. A., Machado, R. F. & Shimoda, L. Pulmonary vascular and ventricular dysfunction in the susceptible patient (2015 Grover conference series). *Pulmon. Circ.* **6**, 426–438 (2016).
48. Zhang, Z. *et al.* Targeted sequencing identifies the genetic variants associated with high-altitude polycythemia in the Tibetan population. *Indian J. Hematol. Blood Transf.* **38**, 556–565 (2022).
49. Tian, H., Qiang, T., Wang, J., Ji, L. & Li, B. Simvastatin regulates the proliferation, apoptosis, migration and invasion of human acute myeloid leukemia cells via miR-19a-3p/HIF-1α axis. *Bioengineered* **12**, 11898–11908 (2021).
50. Liu, H., Shi, C. & Deng, Y. MALAT1 affects hypoxia-induced vascular endothelial cell injury and autophagy by regulating miR-19b-3p/HIF-1α axis. *Mol. Cell. Biochem.* **466**, 25–34 (2020).
51. Assis-Nascimento, P., Tsenkina, Y. & Liebl, D. J. EphB3 signaling induces cortical endothelial cell death and disrupts the blood–brain barrier after traumatic brain injury. *Cell Death Dis.* **9**, 1–15 (2018).
52. Bailey, P. S., Hiltunen, J. K., Dieckmann, C. L., Kastaniotis, A. J. & Nathan, J. A. Different opinion on the reported role of Poldip2 and ACSM1 in a mammalian lipoic acid salvage pathway controlling HIF-1 activation. *Proc. Natl. Acad. Sci.* **115**, E7458–E7459 (2018).
53. Paredes, F., Williams, H. & Martin, A. S. 258-Poldip2 is an oxygen-sensitive mitochondrial protein that controls oxidative/glycolytic metabolism balance and proteasome activity. *Free Radic. Biol. Med.* **112**, 173–174 (2017).
54. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **14**, e0224365. https://doi.org/10.1371/journal.pone.0224365 (2019).
55. Guo, C. Y. & Chou, Y. C. A novel machine learning strategy for model selections—Stepwise Support Vector Machine (StepSVM). *PLoS ONE* **15**, e0238384. https://doi.org/10.1371/journal.pone.0238384 (2020).
56. Darst, B. F., Malecki, K. C. & Engelman, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* https://doi.org/10.1186/s12863-018-0633-8 (2018).
57. Chen, T. *et al.* Xgboost: Extreme gradient boosting. *R package version 0.4-2* **1**, 1–4 (2015).
58. Chen, G. *et al.* Restructured GEO: Restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database https://doi.org/10.1093/database/bay145 (2019).
59. Mlecnik, B., Galon, J. & Bindea, G. Automated exploration of gene ontology term and pathway networks with ClueGO-REST. *Bioinformatics (Oxford, England)* **35**, 3864–3866 (2019).
60. Courtot, M. *et al.* UniProt-GOA: A central resource for data integration and GO annotation. *SWAT4LS* **2015**, 227–228 (2015).

Acknowledgements

This work was supported by the Natural Science Foundation of Si Chuan Province (2023NSFSC0528) and Institute Management Project (2021-XZYG-B34) of the General Hospital of Western Theater Command. We would like to give special thanks to all the contributors of GSE103927 and GSE52209 for their meaningful data. We thank Dr. Haijing Wang from the Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, for her helpful advice in regard to the use of R packages.

Author contributions

M.Y., N.Z., Y.W., XB.Y. conceived and designed the study. M.Y. and Y.L. acquired funding for the study and provided the necessary support. M.Y., T.L., and X.B.Y. collected and analyzed the microarray data. M.Y., Y.W., Y.Z. Yue.Z. and Y.L. conducted the image analysis and interpretation and wrote the manuscript. All the authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31797-0.

Correspondence and requests for materials should be addressed to M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.