论 著

基于深度测序和生信分析挖掘高原汉族人群红细胞增多症变异的研究

王思源，罗勇军
陆军军医大学 陆军卫勤训练基地军事医学地理教研室（重庆 400038）

【摘　要】目的 通过小样本外显子测序与公用数据库比对进行生物信息挖掘，探寻汉族人群高原红细胞增多症（high altitude polycythemia，HAPC）变异与发病的相关性。 **方法** 纳入汉族高原红细胞增多症男性患者4例和高原健康男性5例，采集静脉血并提取DNA进行全外显子组测序；将测序数据进行功能学富集分析（gene ontology，GO & kyoto encyclopedia of genes and genomes，KEGG）构建突变基因蛋白质互作网络（protein protein interaction ，PPI），选择显著性最高的基因，筛选潜在关键HAPC相关变异位点，初步探究相关变异与HAPC发病可能的机制。 **结果** 通过全外显子测序，发现HAPC相关的突变基因216个，其中表皮生长因子（epidemal growth farctor，EGF）基因在功能学富集分析（GO & KEGG），蛋白质互作网络（PPI）中具有高度显著性，初步筛选出5个潜在关键HAPC相关变异位点，通过生物信息学预测及分析，最终筛选出EGF基因上3个可能的变异c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）以及c.2759A>T（p.Glu920Val），可能是HAPC的突变位点，与HAPC发病相关。 **结论** 本研究表明汉族EGF基因的变异与HAPC发病存在相关性。EGF基因的变异可能通过干扰RNA剪接，改变RNA二级结构，影响蛋白质结构，进而影响EGF的生成及HAPC的发生发展。

【关键词】高原红细胞增多症；EGF基因；生物信息学；全外显子测序
【中图分类号】R811.4　　**文献标志码**　A　　DOI：10.3969/j.issn.2096-3351.2024.03.007

A study of polycythemia variants in the plateau Han population based on deep sequencing and bioinformatics analysis

WANG Siyuan，LUO Yongjun
Department of Military Medical Geography，Army Health Medical Service Training Base，Army Medical University，Chongqing 400038，China

【Abstract】Objective By comparing small-sample exome sequencing with public databases，bioinformatics mining is conducted to explore the correlation between variations and the onset of high altitude polycythemia （HAPC） in the Han population. **Methods** Firstly，four cases of Han Chinese men with HAPC and five cases of healthy men with HAPC were enrolled，and venous blood was collected and DNA was extracted for whole exome sequencing；the sequencing data were subjected to functional enrichment analysis （Gene Ontology （GO） & Kyoto Encyclopedia of Genes and Genomes （KEGG）） to construct a protein protein interaction （PPI） network for mutated genes，and the most significant genes were selected to screen for potential key HAPC-related variants and the possible mechanisms of HAPC. We constructed a protein protein interaction （PPI） network，selected the most significant genes，screened potential key HAPC-associated mutation sites，and initially investigated the possible mechanisms of HAPC pathogenesis. **Results** Through whole exome sequencing，216 HAPC-related mutated genes were identified，among which the EGF gene was highly significant in the Functionality Enrichment Analysis （GO & KEGG），Protein Protein Interaction Network （PPI），and 5 potentially key HAPC-related variant sites were initially screened，and through bioinformatic prediction and analysis，3 possible EGF genes on the EGF gene were finally screened out variants c.2124G > A （p.Met708Ile），c.2351A > T （p.Asp784Val），and c.2759A > T （p.Glu920Val），which might be mutation sites for HAPC and were associated with HAPC pathogenesis. **Conclusion** The present study demonstrated a correlation between variants in the EGF gene and the development of HAPC in the Han Chinese population. Variants in the EGF gene might affect protein structure by interfering with RNA splicing and altering the secondary structure of RNA，thereby affecting the production of EGF and the development of HAPC.

【Key words】 Haigh altitade polycythemia（HAPC）；EGF gene；bioinformatics；whole exon sequencing

高原性红细胞增多症（high altitude polycythemia，HAPC）是机体在高原缺氧环境中，由于红细胞过度增生导致血液粘度增高和血流阻力增加引起的一种常见特发性慢性高原疾病，是一种代偿性反应，可引起人体多器官系统损害[1-2]。该病主要发生在海拔2 500 m以上的移居和世居人群，男性发病率高于女性，且在不同

基金项目：国家科技部第二次青藏高原综合科学考察研究专题（2019QZKK0607）
通信作者：罗勇军，E-mail：luoyj@tmmu.edu.cn
引用本文：王思源，罗勇军. 基于深度测序和生信分析挖掘高原汉族人群红细胞增多症变异的研究[J]. 西南医科大学学报，2024，47（3）：215-220. DOI：10.3969/j.issn.2096-3351.2024.03.007.

人群中差异显著。针对移居汉族人群，目前有研究表明促红细胞生成素（erythropoietin，EPO）、缺氧诱导因子（hypoxia inducible factor，HIF）等基因的变异可能促进HAPC的进展，而某些炎症通路也被证明与HAPC的发生有关[3-5]；同时也有研究通过数据库分析表明甲基化变异等可能通过影响基因表达进而参与HAPC的进程[6-8]。全外显子测序作为一项使用成熟的技术，已经在临床研究实验室中得到广泛的使用。本研究希望通过外显子测序进行潜在HAPC相关变异位点筛选，探寻HAPC相关基因或通路，进而从高原低氧环境视角丰富对HAPC发生发展的认识。

1 材料与方法

1.1 外显子测序

本研究共纳入4例高原HAPC患者（HAPC定义为男性血红蛋白浓度高于210 g/L）与5例高原健康者进行对照，两组对象均为汉族成年男性。排除标准：①慢性呼吸系统疾病、原发性心脑血管疾病等临床表现相似的其他疾病；②炎症发热创伤等急性应激反应。采集血液样本进行DNA提取并进行质检，检测合格的DNA样品片段化处理后制备文库，通过QUBIT对文库进行定量，并用Agilent 2100 Biioanalyzer检测文库插入片段大小情况，最后库检合格样品进行双端测序。本研究获得陆军军医大学伦理委员会批准（2020第001-02），所有参与者均签署知情同意书。

1.2 潜在关键HAPC相关突变基因分析

测序得到的原始双端序列进行质量评估和过滤后，将高质量的测序结果与参考序列进行比对，利用ANNOVAR和CIRCOS进行变异注释和全局变异总览，利用ANNOVAR对检测出的SNV和InDel进行注释，并筛选位于基因功能区域的位点，使用fisher test比较两组中基因或基因型有差异的位点（$P<0.05$），并使用多重检验BH方法矫正，选择$P<0.05$作为最后的差异具有统计学意义，筛选出变异位点数有较大差异的基因。利用基因本体论数据库（gene ontology database，GO）[9]与KEGG（https:www.genome.jp/kegg/）进行通路的分析[10]，初步筛选出潜在关键HAPC相关突变基因。

1.3 潜在HAPC相关变异位点的生物学功能预测

运用STRING构建汉族HAPC突变基因相关PPI[11]，通过Cytoscape软件中的MCODE插件进行模块化分析，将筛选得到的变异位点进行生物信息学预测[12]。使用RNAsnp软件通过RNA折叠算法预测变异位点对局部RNA二级结构影响[13-14]，以判断其对后续蛋白质合成的影响；使用PESX及RESCUE-ESE软件综合分析变异位点是否对剪接调节功能造成影响；使用ESE finder软件预测变异位点对剪切增强子干扰的具体影响[15]。

2 结果

2.1 外显子测序

对HAPC组及对照组进行了外显子测序分析，平均深度分别为177.39和163.83。HAPC组和对照组存在差异的突变基因216个（图1A），涉及INDELs突变的基因210个（图1B），SNVs突变的基因9个（图1C），位于差异基因上影响基因功能的突变位点数目为871个。

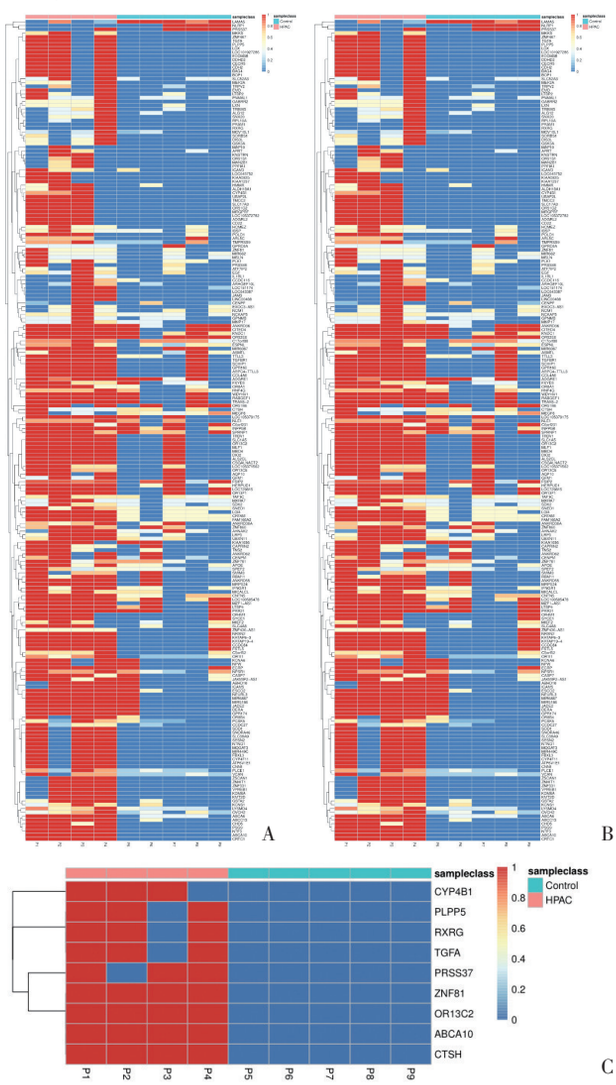


图1 外显子测序法检测具有SNV和INDELs突变的病例对照存在差异基因

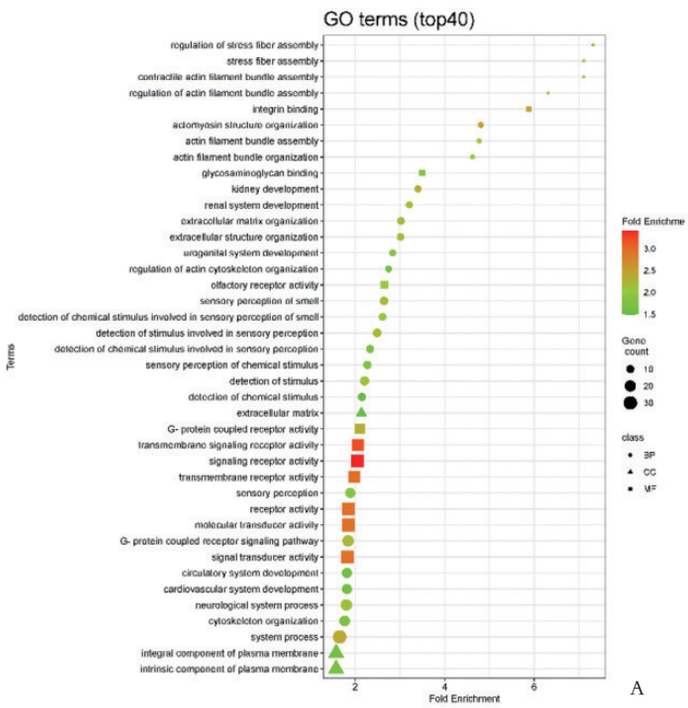
Figure 1 Exome sequencing was used to detect differential genes in case-controls with SNV and INDELs mutations

注：A：病例对照存在差异基因216个；B：涉及INDELs突变的基因210个；C：涉及SNVs突变的基因9个。

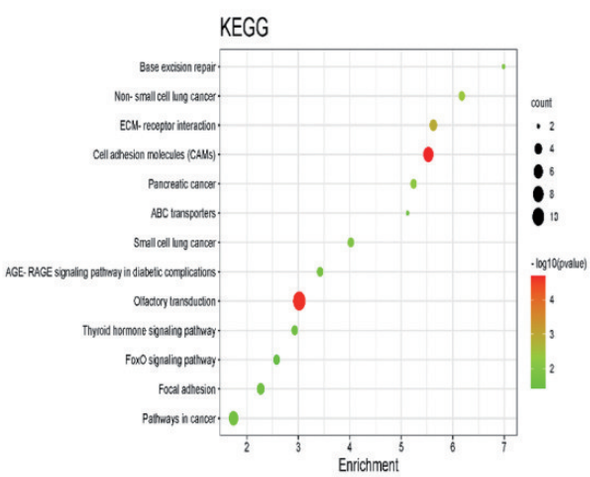
2.2 潜在关键HAPC相关突变基因分析

对汉族HAPC突变相关基因进行功能学富集GO（图2A）和KEGG（图2B）分析。根据P值筛选最显著的40条GO条目，结果发现有关肾脏、循环发育等GO条目在HAPC突变基因中富集，提示HAPC突变基因与肾脏合成EPO有关。选择符合筛选条件（$P\leqslant0.05$）且

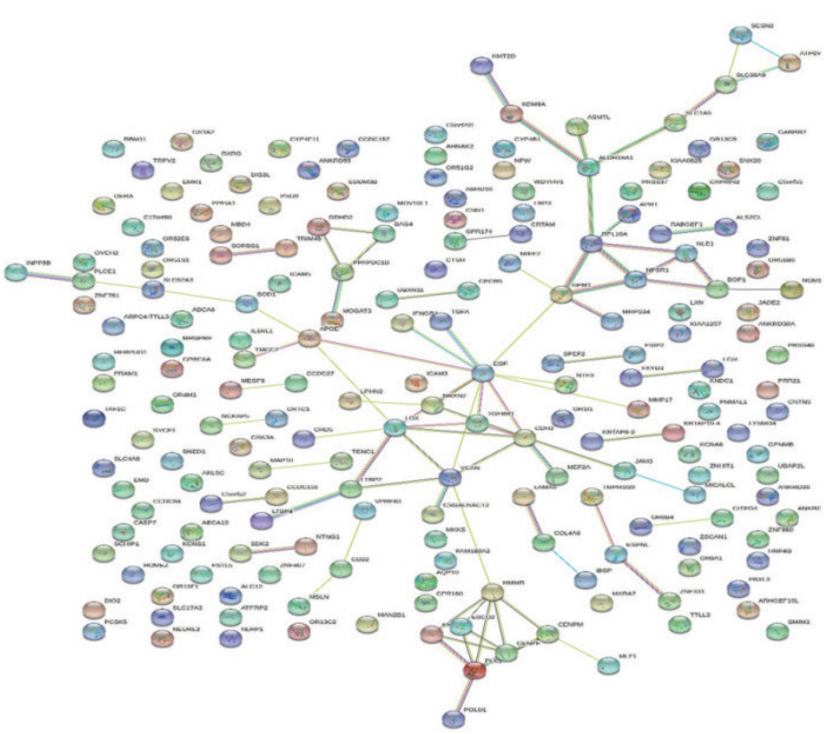
富集倍数较高的信号通路，结果显示CAM等信号途径在突变基因中富集倍数均大于1.5，证明显著富集，提示红细胞增多与细胞增殖机制有关系。运用STRING构建突变基因相关PPI（图2C），共192点，88边。通过网络深度分析发现，网络特征拓扑参数节点度（degree）和介数中心性（between ness）分布呈幂指数分布（图2D），运用韦恩图将节点度和介数中心性值前30%取交集，结果发现25个基因具有高节点度和高介数中心性值。通过Cytoscape软件中的MCODE插件进行模块化分析，从PPI网络中筛选出最重要的第一个模块，它包括11个节点和44条边（图2E）。其中，EGF、LOX等8个基因（表1）既具有网络高拓扑特征值又位于第一模块（图2F）。在这8个基因中，EGF是节点度和介数中心性值最高者，且GO显示其具有高度显著性，提示EGF可能是HAPC的重要变异基因。在样本中，检测到HAPC组EGF基因主要有5个变异位点突变，因此，我们将EGF基因的5个变异位点作为我们后续研究对象。



A



B



C

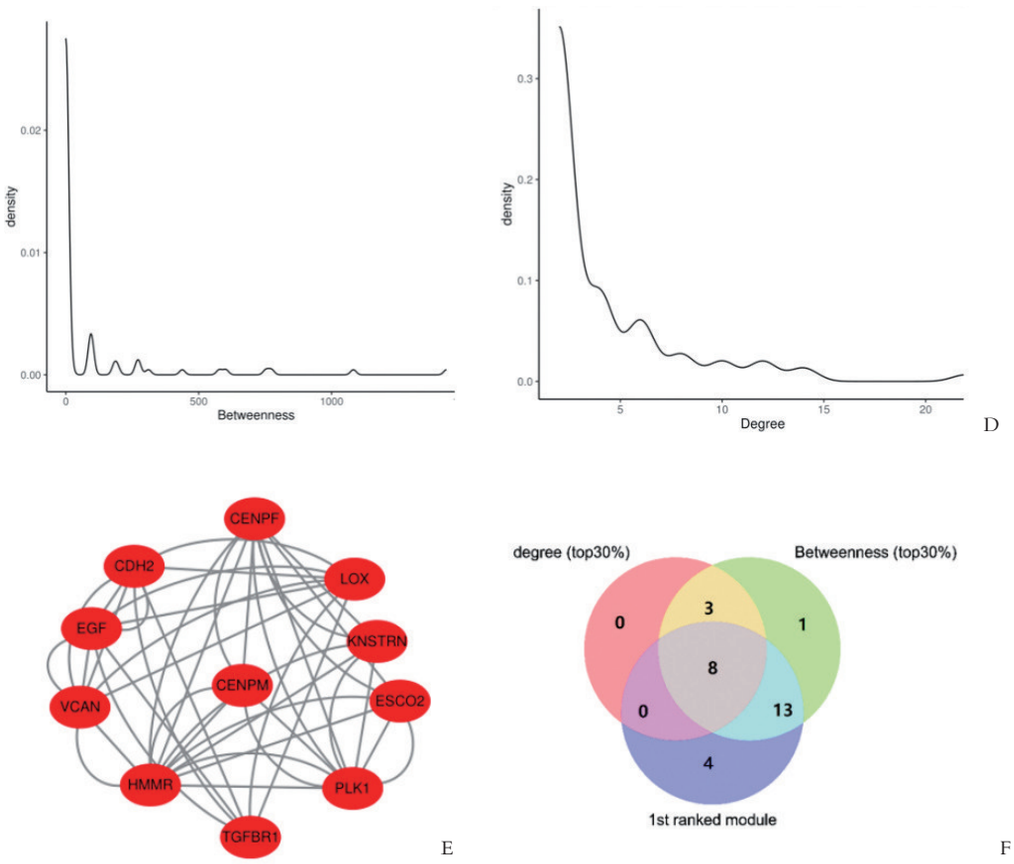


图2　突变基因的生物信息学分析

Figure 2 Bioinformatics analysis of mutant genes

注：A：功能学富集GO；B:KEGG分析；C：运用STRING构建突变基因相关PPI；D: 网络特征拓扑参数节点度和介数中心性分布；E: 第一模块；F：韦恩图。

表1　8个潜在关键HAPC相关突变基因的信息

Table 1 Information on 8 potentially key HAPC-associated mutation genes

基因名	基因ID	定位	SNP位点数	GO条目数	KEGG条目数	高节点度	高介数中心性值
EGF	1950	4q25	5	8	5	22	1433
CDH2	1000	18q12	2	2	1	14	311
LOX	4015	5q23	1	5	–	14	275
VCAN	1462	5q14	7	4	1	12	774
HMMR	3161	5q34	3	1	1	12	577
PLK1	5347	16p12	2	1	1	12	97
CENPF	1063	1q41	24	3	–	10	3
CENPM	79019	22q13	4	–	–	8	94

2.3　潜在HAPC相关变异位点的生物学功能预测

2.3.1　千人基因组计划中全球不同人群的等位基因分布　通过千人数据基因库查找比对全球不同人群的等位基因分布（图3），发现亚洲汉族人群中rs11568943、rs11569017、rs4698803与其他人群存在较大差异，进一步提示这三个SNP位点可能与HAPC发生有关。

2.3.2　5个HAPC相关变异位点生物信息学预测　为进一步验证5个突变位点与HAPC患病风险是否显著相关，将筛选得到的5个变异位点进行生物信息学预测，见表2。①变异位点蛋白编码功能分类：5个位点均为非同义突变，且得分较高，证明为有害突变的可能性较大；②变异容忍度：5个位点得分较低，不能提供明确依据证明有害；③变异位点蛋白编码功能预测：rs11569017、rs4698803位点得分较小，表明这两个突变位点导致蛋白结构或功能改变的可能性大；④致病性分析：rs4698803综合得分较高，进一步证明该位点的有害性较高。综合上述分析，rs764312466、rs11568943与rs2237051位点在预测中得分较高，提示其改变对疾病发生影响较大，可能是导致HAPC发生的变异位点。

2.3.3　5个HAPC相关变异位点对其RNA二级结构的影响　利用RNAsnp对RNA二级结构进行构建后，除了c.562 G>A对RNA二级结构无影响外，其余所有变异位点对RNA二级结果影响较大（见图4）。与野生型相比，c.1292 G>A在主茎环结构外增加了两种类似的“茎泡”样结构且减少了两个“茎泡”样结构；c.2124 G>A在主茎环结构上减少了一个“茎泡”样结构，且其附近的茎环结构发生了显著改变；c.562G>A、c.2759A>T突变体的主茎环结构无显著差异，提示其对RNA二级结构差异影响不大；c.2351 A>T无法做出二级结构图，因此

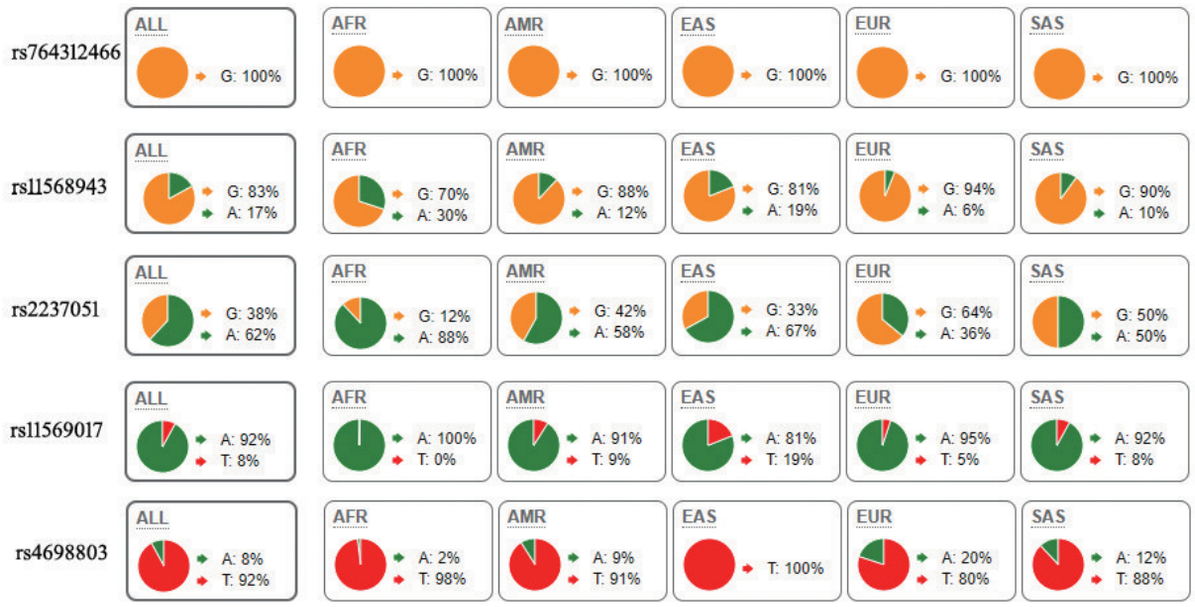


图3　千人基因组项目中世界各地不同人群中突变位点等位基因的分布

Figure 3 Distribution of mutation alleles in different populations around the world in the 1000 Genomes Project

表2　EGF基因中5个HAPC相关突变位点的生物学分析

Table 2 Biological analysis of five HAPC-related mutation sites in EGF gene

定位	rs764312466 Chr.4:109943894	rs11568943 Chr.4:109961965	rs2237051 Chr.4:109980042	rs11569017 Chr.4:109980955	rs4698803 Chr4:109993271
突变位点	c.562G>A (p.Asp188Asn)	c.1292G>A (p.Arg431Lys)	c.2124G>A (p.Met708Ile)	c.2351A>T (p.Asp784Val)	c.2759A>T (p.Glu920Val)
功能分类	错义突变	错义突变	错义突变	错义突变	错义突变
MAF	< 0.01	0.38	0.50	0.23	0.30
变异容忍度	0.125	0.152	0.004	0.261	0.195
−CADD	0.093	3.276	0.045	15.98	0.677
−GERP	−5.56	−1.92	−5.56	0.40	−5.02
变异位点蛋白编码功能预测					
−SIFT	1	0.89	1	0.01	0.15
−PloyPhen	0	0	0	0.005	0
致病性分析					
−REVEL	0.342	0.171	0.207	0.105	0.082
−MetaLR	0.465	0	0.252	0	0.67

推测其变异位点对二级结果不具有影响。

对变异位点剪接调节功能进行预测，通过ESEfinder检测到变异位点均有不同类型的SRSF序列的改变，提示其对外显子的剪切有明显的影响。运用ensemble发现EGF的位点c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）、c.2759A>T（p.Glu920Val）经过检测发生变化，可能会影响拼接精度或效率而引起表型变化；运用ESRrsearch发现c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）、c.2759A>T（p.Glu920Val）有变化，提示其在外显子剪切中会存在干扰；运用PESX及RESCUE-ESE软件中发现c.562G>A（p.Asp188Asn）、c.2351A>T（p.Asp784Val）以及c.2759A>T（p.Glu920Val）对剪切增强子均存在干扰。以上提示位点c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）、c.2759A>T（p.Glu920Val）可能通过对外显子的剪切造成影响，引起后续蛋白合成的变化，见表3。

表3　变异生物功能信息预测结果

Table 3 Prediction results of variant biological function information

变异位点	SNP ID	剪接增强器干扰		功能分类	ESEFinder	ESRSearch	PESX	RESCUE ESE
		剪切增强子获得	剪切增强子丢失					
c.562G>A(p.Asp188Asn)	rs764312466	SRSF5序列		非同义突变	C	N	C	N
c.1292G>A(p.Arg431Lys)	rs11568943	SRSF5序列		非同义突变	C	N	N	N
c.2124G>A(p.Met708Ile)	rs2237051	SRSF1序列		非同义突变	C	C	C	C
c.2351A>T(p.Asp784Val)	rs11569017	SRSF5序列		非同义突变	C	C	C	C
c.2759A>T(p.Glu920Val)	rs4698803		SRSF2序列	同义突变	C	C	C	C

注：C：有变化；N：无变化。

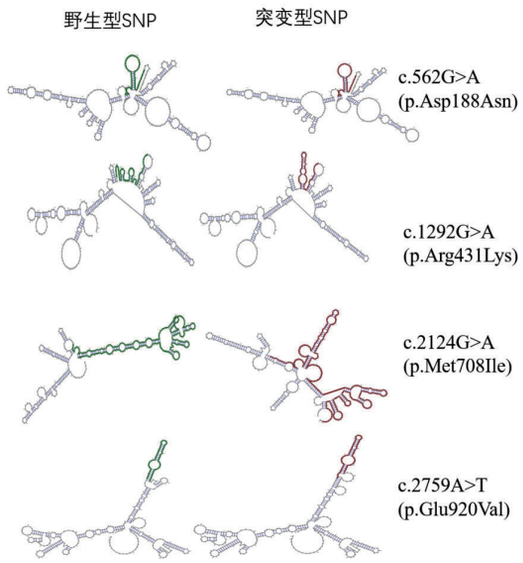


图4 变异位点对其RNA二级结构的影响

Figure 4 Effect of variant sites on the secondary structure of RNA

3 讨论

久居高原的人群在长期缺氧的自然环境中发生了一系列的病理生理变化，这些适应性改变促进其对氧的利用。目前对HAPC的研究主要集中于亚洲的青藏高原、南美洲的安第斯山脉和东非的埃塞俄比亚高原原住民，但对高原汉族的研究偏少。本研究通过外显子测序和生物信息学分析，发现了EGF基因可能参与了这一特殊人群HAPC的发生、发展。本研究结果表明EGF位点与HAPC患病风险显著相关，发现EGF的c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）以及c.2759A>T（p.Glu920Val）变异在HAPC组中分布频率升高，据此我们推测这三个突变可能通过对外显子的剪切造成影响，引起RNA二级结构的改变，影响蛋白合成的变化，增加EGF的合成进而激活EGF通路，刺激VEGF-A的分泌和VEGFR-2刺激细胞增殖[16-18]，增加高原低氧暴露环境下血管损伤和局部炎症反应，促进红细胞增多，最终导致HAPC的进展[19-20]。

本研究纳入样本量较少，今后将扩大样本量，并探讨不同高海拔高原患者是否存在差异。

4 结论

本研究确定了EGF这一与高原汉族人群HAPC风险相关的基因，并结合生物信息学分析，发现HAPC风险与EGF上的c.2124G>A（p.Met708Ile）、c.2351A>T（p.Asp784Val）以及c.2759A>T（p.Glu920Val）突变位点高度相关。研究结果揭示了这一特殊人群中可能存在的HAPC风险因素，提高了对汉族人群在缺氧条件下HAPC遗传变异位点的认识，将更有利于我们未来对慢性高原病的研究，并为探索HAPC的发病机制，以及可能的靶向治疗和预防方案提供了基础理论依据。

5 参考文献

[1] DENG BN, LIU WL, PU LL, *et al*. Quantitative proteomics reveals the effects of resveratrol on high-altitude polycythemia treatment[J]. Proteomics, 2020, 20(14): e1900423.
[2] SÁNCHEZ K, BALLAZ SJ. Might a high hemoglobin mass be involved in non-cardiogenic pulmonary edema? The case of the chronic maladaptation to high-altitude in the Andes[J]. Med Hypotheses, 2021, 146: 110418.
[3] AZAD P, ZHAO HW, CABRALES PJ, *et al*. Senp1 drives hypoxia-induced polycythemia via GATA1 and Bcl-xL in subjects with Monge's disease[J]. J Exp Med, 2016, 213(12): 2729-2744.
[4] FAN XW, MA LF, ZHANG ZY, *et al*. Associations of high-altitude polycythemia with polymorphisms in PIK3CD and COL4A3 in Tibetan populations[J]. Hum Genomics, 2018, 12(1): 37.
[5] 申杨磊，方龙伟，央拉，等. 利用数据库解析高原红细胞增多症与肺动脉高压基因表达差异的关联性[J]. 检验医学与临床，2023，20(13): 1872-1877.
[6] 罗勇军，陈郁，蒲懿，等. 基于基因芯片技术的高原红细胞增多症基因组DNA甲基化差异分析[J]. 重庆医学，2023，52(14): 2132-2137，2142.
[7] 冯思维. 西藏地区高原红细胞增多症患者外周血mRNA的差异表达分析[D]. 咸阳: 西藏民族大学，2023.
[8] 高文宇，曾蓉，许梦娜，等. EPAS1基因rs1868092位点多态性与中国藏族人群高原红细胞增多症的相关性研究[J]. 大理大学学报，2023，8(4): 39-42.
[9] ZHAO YW, WANG J, CHEN J, *et al*. A literature review of gene function prediction by modeling gene ontology[J]. Front Genet, 2020, 11: 400.
[10] KANEHISA M, SATO Y, FURUMICHI M, *et al*. New approach for understanding genome variations in KEGG[J]. Nucleic Acids Res, 2019, 47(D1): D590-D595.
[11] SZKLARCZYK D, MORRIS JH, COOK H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible[J]. Nucleic Acids Res, 2017, 45(D1): D362-D368.
[12] REVA B, ANTIPIN Y, SANDER C. Predicting the functional impact of protein mutations: application to cancer genomics[J]. Nucleic Acids Res, 2011, 39(17): e118.
[13] HUA JT, AHMED M, GUO HY, *et al*. Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19[J]. Cell, 174(3): 564-575.e518.
[14] SABARINATHAN R, TAFER H, SEEMANN SE, *et al*. The RNAsnp web server: predicting SNP effects on local RNA secondary structure[J]. Nucleic Acids Res, 2013, 41(Web Server issue): W475-W479.
[15] FAIRBROTHER WG, YEO GW, YEH R, *et al*. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons[J]. Nucleic Acids Res, 2004, 32(Web Server issue): W187-W190.
[16] 陈郁. 高原移居汉族男性高原红细胞增多症遗传易感性研究[D]. 重庆: 第三军医大学，2016.
[17] 侯云鹏，董红梅，陈郁，等. 慢性高原病诊治研究进展[J]. 人民军医，2017，60(12): 1238-1242.
[18] 吴萍，赵艳霞，李永平，等. 基于因子分析的高原红细胞增多症中医证候量化诊断研究[J]. 西部中医药，2023，36(10): 104-111.
[19] NICOLAS S, ABDELLATEF S, HADDAD MA, *et al*. Hypoxia and EGF stimulation regulate VEGF expression in human glioblastoma multiforme (GBM) cells by differential regulation of the PI3K/rho-GTPase and MAPK pathways[J]. Cells, 2019, 8(11): 1397.
[20] 郭勇，王生艳，易静静，等. 红景天苷对高原红细胞增多症模型大鼠骨髓$CD71^+$有核红细胞凋亡的影响[J]. 吉林大学学报(医学版)，2023，49(5): 1174-1181.

（利益冲突：无）

（收稿日期：2023-08-10；修回日期：2024-03-24）