数据库查询示例

NCBI：

NCBI 数据库，全称 National Center for Biotechnology Information（国家生物技术信息中心），是隶属美国国家卫生研究所的国家医学图书馆(NLM)的分部，拥有包括 PubMed、PubMed Central 和 GenBank 在内的大约 40 个在线文献和分子生物学数据库。数据库资源涵盖基因组、变异、基因和基因表达、蛋白质、小分子和生物信息学分析等关键词，汇集生物分子结构、元件和数据分析工具。

（1） 查文献

文献检索是科研工作绕不开的内容，NCBI 的文献资源包括：NLM Catalog、PubMed、PubMed Central (PMC)、Bookshelf 以及补充资源 MeSH、PubMed DTD、JATS、NIHMS。最常用的是 PubMed

以下是文献检索步骤：

1. 选择 PubMed
2. 在搜索框输入关键词，以 OCA2 基因为例
3. 通过增加关键词缩小范围

两种常用方式：

1) 布尔逻辑检索符：AND（与）、OR（或）、NOT（非）

注意，逻辑字符要大写

以 OCA2 基因 rs1800414 位点为例进行检索：

2) “Advanced”高级检索，通过不同的字段来增加检索条件

4. 跟踪进展
   可注册 NCBI 账户，然后点击 Create alert

5. 设置显示的模式

1) 显示是否按时间、相似度等排序，是否显示摘要

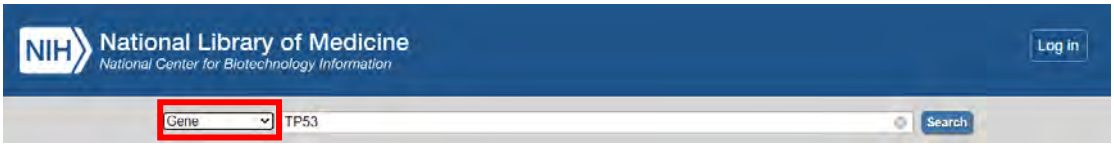
2) 根据文章类型、是否提供免费全文、出版日期等

在文章搜索界面左侧边框当中尝试勾选设置不同参数

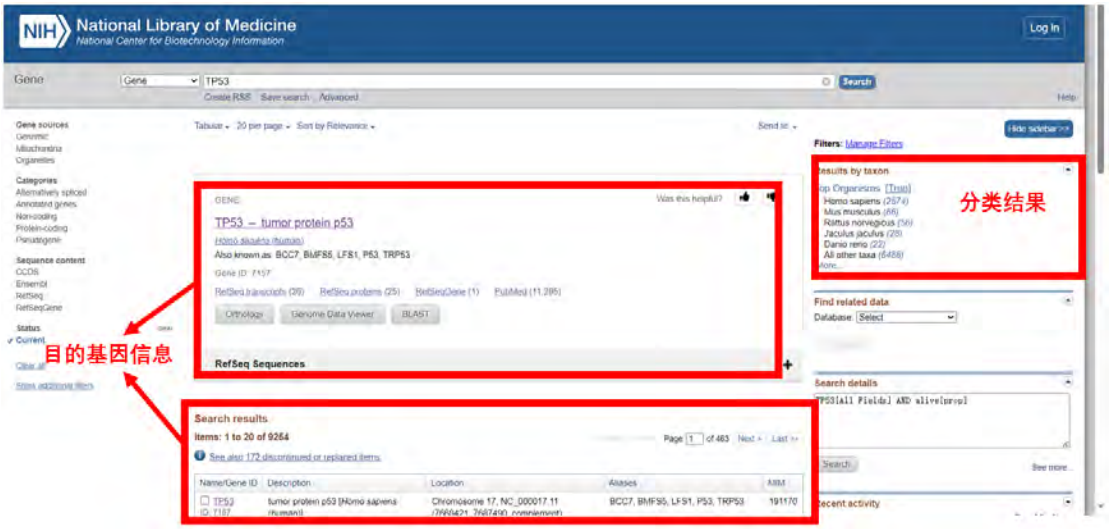
6. 保存检索结果

（2） 查基因（以抑癌基因 TP53 为例）

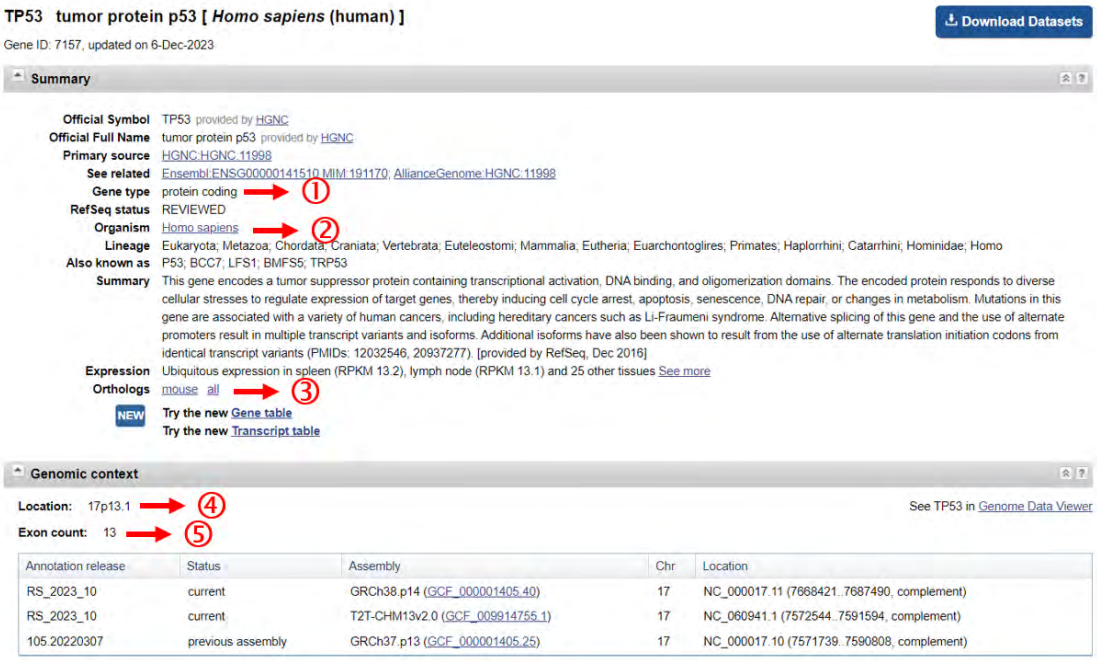
查询时选择数据库类型为“Gene”，搜索框输入想查询的基因名字，如 TP53。



检索后会得到多条与输入内容有关的信息。相同基因名可能存在于多个物种中。



点击“TP53-tumor protein p53”获得更多基因信息。



① 基因类型，TP53 为蛋白质编码基因。

常见基因类型有 Non-coding RNA Gene (非编码 RNA 基因，如 rRNA, tRNA, snRNA, miRNA 登)，Pseudogene (假基因，一类失去功能的基因副本，可能是由于基因重复、突变等原因而失去了编码功能)，Transposon (转座子，移动元件，可以在基因组中移动，并且可能对基因的结构和调控产生影响)，Viral Gene(病毒基因，与病毒有关的基因，包括病毒自身的基因或与宿主相互作用的基因)，Mitochondrial Gene (线粒体基因，存在于线粒体中的基因，编码线粒体所需的蛋白质、rRNA 和 tRNA) 等。

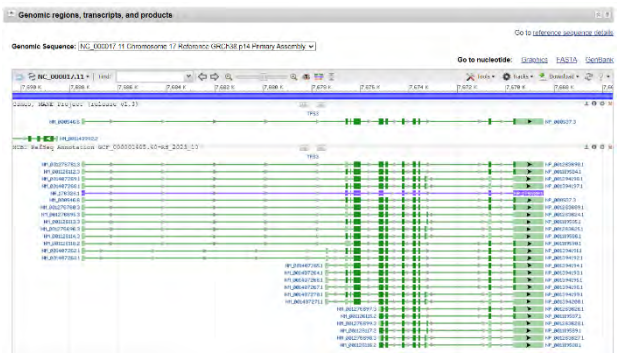
② “Organism”生物体，表示检索基因所属的生物种类或生物体名称，示例的“Homo sapiens”为智人。

③ “Orthologs”同源基因，指与查询基因在不同物种中具有共同祖先的基因。实例中老鼠中的 TP53 与智人的基因存在同源关系。同源基因是在进化过程中由共同祖先基因演变而来的基因，它们在不同物种中可能具有相似的功能。

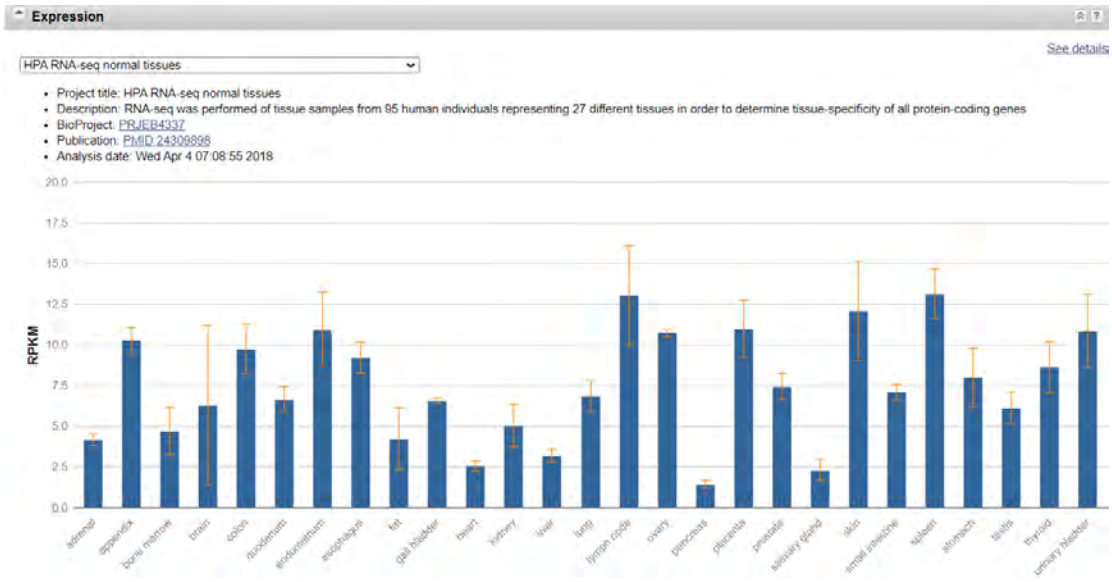
④ 基因所在位置，示例的“17p13.1”表示该基因在第 17 号染色体的短臂上的带 13 的子带的第一亚带。

⑤ 外显子数量。示例外显子数量为 13。

下图是该基因不同的转录本。

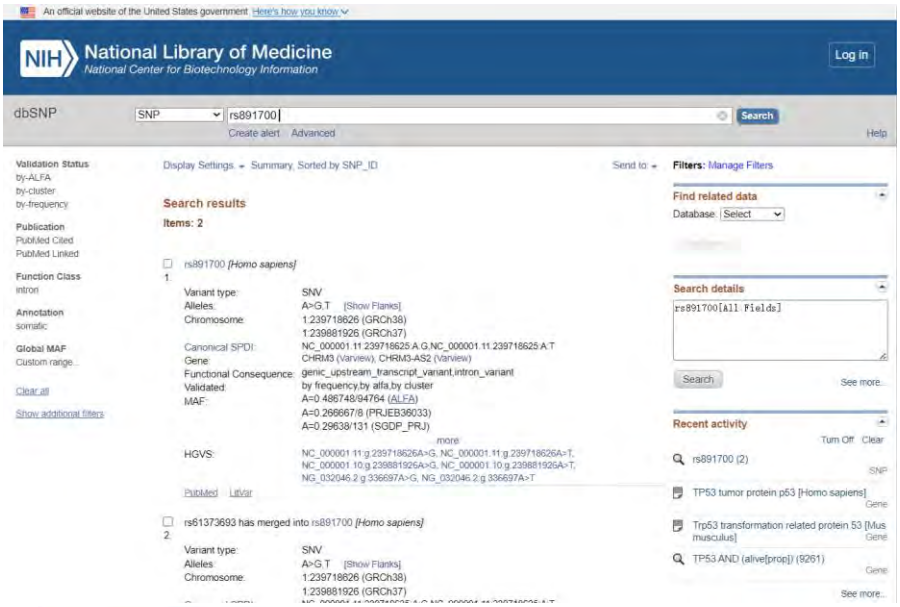


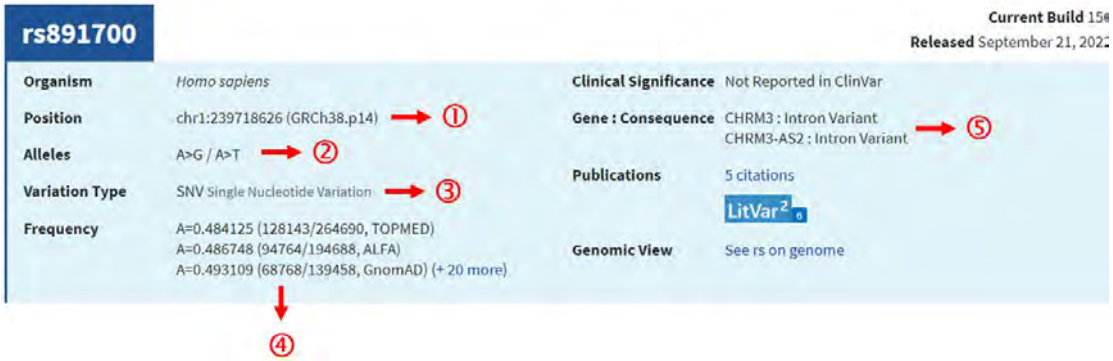
该基因在不同组织内的表达情况。



（3） 查 SNP（以个体识别位点 rs891700 为例）

选择数据库类型为 SNP，输入 SNP 的位点编号。





① SNP 的具体位置。示例为 1 号染色体第 239718626 个碱基

② 等位基因。示例中参考基因组碱基为 A，部分人群该位点变异为 G，极少量人群该位点 b 变异为 T。

③ 变异类型。示例为 SNV（单核苷酸替代，最常见的 SNP 类型）

其他变异类型还有插入 (insertion，在某个位置插入了一个或多个额外的碱基。导致该位置上的序列长度增加)、缺失(deletion，在某个位置删除了一个或多个碱基导致该位置上的序列长度减少)、复合变异(complex variation，包括了两个以上的变异类型的组合，例如插入和替代的组合)、多态性变异(在种群中存在多个等位基因，并且每个等位基因的频率超过 1%)、微卫星(microsatellite，短重复序列的变异，其中一个短序列单元的重复次数发生变化)等.

④ 不同碱基在种群中的频率。

⑤ “gene: consequence”表示关于该 snp 的基因和其影响的信息，用于描述 SNP 如何影响与之相关的基因。示例“CHRM3 : Intron Variant”表示该 snp 在基因 CHRM3 内，该变异位于基因内的内含子区域。

其他可能的“consequence”包括但不限于：Synonymous-导致编码相同氨基酸的变异；Intronic-位于基因内的内含子区域；Upstream/Downstream-位于基因上游或下游区域; Splice site-影响剪切位点的变异等。

OMIM：

OMIM 数据库，全称 Online Mendelian Inheritance in Man（在线人类孟德尔遗传数据库）
OMIM 数据库包括:
1.gene entry 基因条目；2.allelic variations 等位基因变异；3.gene map 基因图谱；4.phenotypic series 表型系列；5.phenotype entry 表型条目；6.clinical synopsis 临床提要；7.external links 外部链接

基因型与表型的关系:

以 OCA2 为例，可按以下路径查询:

https://www.omim.org/entry/611409?search=OCA2&highlight=oca2

HGNC Approved Gene Symbol: OCA2

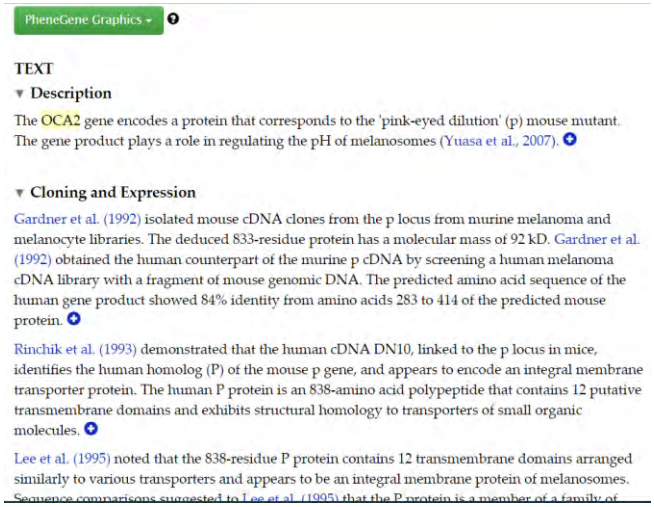
Cytogenetic location: 15q12-q13 ***Genomic coordinates (GRCh38): 15:27,719,007-28,099,341*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype Clinical Synopses	Phenotype MIM number	Inheritance	Phenotype mapping key
15q12-q13	[Skin/hair/eye pigmentation 1, blond/brown hair]	227220	AR	3
	[Skin/hair/eye pigmentation 1, blue/nonblue eyes]	227220	AR	3
	Albinism, brown oculocutaneous	203200	AR	3
	Albinism, oculocutaneous, type II	203200	AR	3

location 代表相关基因在染色体中的位置；phenotype 代表基因相关的表型；phenotype MIM number 代表表型的 MIM 编号；inheritance 代表遗传，是指该基因的遗传类型，如 AR 是指常染色体隐性遗传，AD 是常染色体显性遗传，SMu 是指体细胞突变，鼠标点击缩写符号就会出现不同缩写代表的具体含义；phenotype mapping key 代表表型映射关键，3 代表

该疾病的分子基础是已知的；Gene/Locus 代表对应的基因或位点；Gene/Locus MIM number 代表对应的基因或基因座 MIM 编号。点 location 还能显示该位置在同一个染色体相邻的基因列表以及引发的各种疾病。


PheneGene Graphics

TEXT

▼ Description

The OCA2 gene encodes a protein that corresponds to the 'pink-eyed dilution' (p) mouse mutant. The gene product plays a role in regulating the pH of melanosomes (Yuasa et al., 2007).

▼ Cloning and Expression

Gardner et al. (1992) isolated mouse cDNA clones from the p locus from murine melanoma and melanocyte libraries. The deduced 833-residue protein has a molecular mass of 92 kD. Gardner et al. (1992) obtained the human counterpart of the murine p cDNA by screening a human melanoma cDNA library with a fragment of mouse genomic DNA. The predicted amino acid sequence of the human gene product showed 84% identity from amino acids 283 to 414 of the predicted mouse protein.

Rinchik et al. (1993) demonstrated that the human cDNA DN10, linked to the p locus in mice, identifies the human homolog (P) of the mouse p gene, and appears to encode an integral membrane transporter protein. The human P protein is an 838-amino acid polypeptide that contains 12 putative transmembrane domains and exhibits structural homology to transporters of small organic molecules.

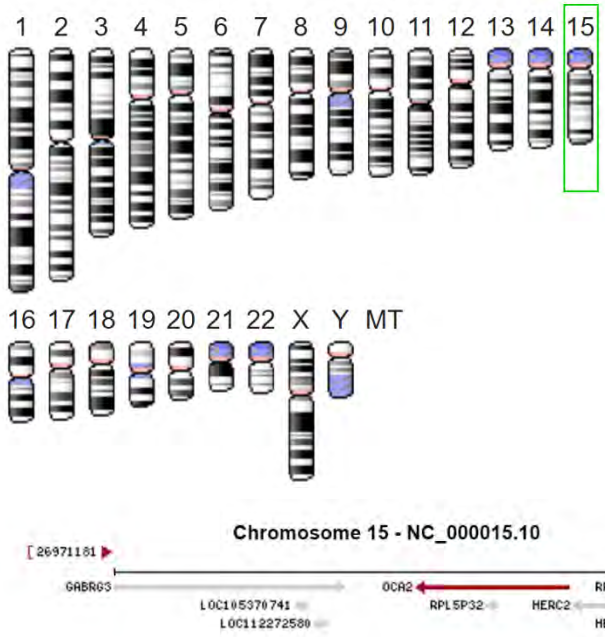
Lee et al. (1995) noted that the 838-residue P protein contains 12 transmembrane domains arranged similarly to various transporters and appears to be an integral membrane protein of melanosomes. Sequence comparisons suggested to Lee et al. (1995) that the P protein is a member of a family of

此即基因型与表型关系查询线索。

通过数据库检索对表型进行解读示例：

我们以东亚人先天肤色深浅度突变位点 rs1800414 为例，介绍查阅 NCBI 和 OMIM 数据库对位点进行解读的步骤和方法，解读结果涵盖以下要点：

1、该位点在基因组和基因序列中的位置；



位点：The T/C-27951890（rs1800414）（参考序列版本：GRch38）

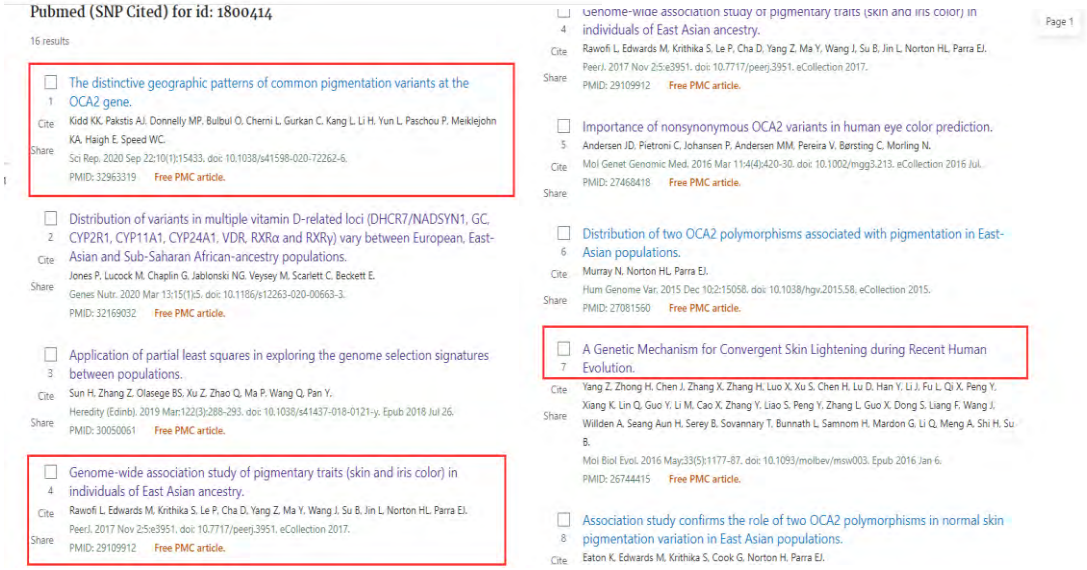
2、该位点对基因表达及性状的影响：

TT：黑色素前体跨膜转运蛋白表达正常，酪氨酸酶合成转运正常，黑色素合成顺畅，先天肤色深；

CC：黑色素前体跨膜转运蛋白表达量少，酪氨酸酶合成和转运受阻，黑色素合成少，先天肤色浅；
TC：黑色素前体跨膜转运蛋白表达量以及酪氨酸酶合成转运量一般，黑色素合成居于中等水平，先天肤色介于深色和浅色之间，表现为一般肤色。

2. 数据库检索-NCBI
检索 rs1800414 位点信息，获取该位点在基因组和基因区域的具体位置信息：

打开页面左下角 PubMed 文献库，选择与 rs1800414 位点及其所在基因、表型等相关的文献：



3. 文献解读

阅读文献，并形成以下位点解读报告

（1）性状背景：人皮肤色素相关背景介绍

人类的肤色是一种高度可遗传的表型特征，不同地理位置、种族的人群肤色是在长期自然选择和人群迁移、混合下形成的，这使得不同人群的肤色表型既有一些共同的遗传基础，又有鲜明的局部遗传差异。亚洲人的典型特征之一是有黄色皮肤，但人群中肤色深浅也有较大差异。

黑色素是决定人类肤色的主要色素，它由酪氨酸衍生而来，分为黑色真黑色素和黄红色脱黑色素两种不同类型。胚胎发育时期的神经嵴细胞分化出成黑素细胞，进一步发育成为皮肤黑素细胞，产生类似溶酶体的细胞器——黑素体，负责合成和储存黑色素。黑素体发育成熟后会向黑素细胞周围的角质细胞中转移 。不同个体皮肤角质层黑素体中黑色素的组成、丰度和分布不同，导致不同个体呈现出的肤色深浅各不相同。

（2）变异位点影响性状效应的解读要点

1）OCA2 基因负责编码跨膜转运蛋白，参与黑色素前体的跨膜转运，同时一定程度上通过加工和转运酪氨酸酶，控制黑素细胞中真黑色素的含量 ，从而影响肤色深浅。OCA2 基因位于第 15 号染色体长臂的第 13 号区域，简写为 15q13。

2）rs1800414 变异位点位于 15 号染色体上 27951890 号位置，落在 OCA2 基因第 17 号外显子区，为非同义突变，引起组氨酸（密码子 CAT/CAC）变为精氨酸（CGT/CGC），导致转运蛋白或酪氨酸酶合成加工相关的蛋白结构发生改变 ，黑色素前体转运及合成受阻。

3）rs1800414 位点 T/C 突变频率在欧洲人群中<0.3%，在中国汉族人群中高于 50%，该变异类型在非洲人群中缺失。（OCA2 基因位于 DNA 负链，与基因组正链碱基型互补）

(http://www.uniprot.org/). The geographic distribution of this SNP across world populations (data from HGDP database: http://hgdp.uchicago.edu) showed that the derived G allele of rs1800414 is highly prevalent in most East Asian populations (G allele > 50% in Han Chinese), as well as in American Indians, but is in low frequency in A-A populations (G allele = 16.0% (A-A from China); 21.3% (A-A from Cambodia)), totally absent in Africans, and absent or extremely rare in western Europeans (G allele < 0.3%) (fig. 2*B*). This piece of evidence suggests that rs1800414 is likely an East Asian specific mutation, consistent with the proposed local selection on this SNP and an independent evolution of skin pigmentation in the area (Edwards, et al. 2010).