实验 2.3 变异检测

一、实验背景

变异检测是指通过高通量测序技术对某一物种个体或群体的基因组进行测序及差异分析， 获得大量的遗传变异信息，如单核苷酸多态性（single nucleotide polymorphisms，简称 SNP）、插入缺失（Insertion–deletion mutations，简称 InDel）、结构变异（structural variation，简称 SV）、拷贝数变异（copy number variation，简称 CNV）等用于开发分子标记建立遗传多态性数据库，为后续揭示进化关系、挖掘功能基因等奠定数据基础。本章节是在序列数据比对的基础上进行变异的识别、过滤和注释。

二、教学目标

变异是什么？变异的类型有哪些？如何基于序列比对结果进行变异的识别？常见的变异检测的方法和工具有哪些？

本节课程，我们将基于序列比对的结果进行变异检测分析，学习变异检测的思路和操作方法。

三、实验原理

第一部分 中心法则

中心法则是指遗传信息从 DNA 传递给 RNA，再从 RNA 传递给蛋白质的转录和翻译的过程，以及遗传信息从 DNA 传递给 DNA 的复制过程，这个过程决定了蛋白质的特异性。这是所有有细胞结构的生物所遵循的法则。

遗传密码（Genetic code）又称遗传编码，是遗传信息的传递规则，将 DNA 或信使 RNA(mRNA)序列以三个核苷酸为一组的“密码子（codon）”翻译为氨基酸序列，以用于指导蛋白质合成。

遗传密码以 DNA 密码子表的形式表示，这是因为在细胞核糖体制造蛋白质时，指导合成蛋白质的是 mRNA。mRNA 的序列则由基因组 DNA 决定。随着计算生物学和基因组学的兴起，现今可在 DNA 水平上发现大多数基因，因此 DNA 密码子表变得愈加有用。标准密码子表如下：

碱基1	碱基2								碱基3
	T		C		A		G		
T	TTT	(Phe/F) 苯丙氨酸	TCT	(Ser/S) 丝氨酸	TAT	(Tyr/Y) 酪氨酸	TGT	(Cys/C) 半胱氨酸	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) 亮氨酸	TCA		TAA[B]	终止（赭石）	TGA[B]	终止（蛋白石）	A
	TTG		TCG		TAG[B]	终止（琥珀）	TGG	(Trp/W) 色氨酸	G
C	CTT		CCT	(Pro/P) 脯氨酸	CAT	(His/H) 组氨酸	CGT	(Arg/R) 精氨酸	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) 谷氨酰胺	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) 异亮氨酸	ACT	(Thr/T) 苏氨酸	AAT	(Asn/N) 天冬酰胺	AGT	(Ser/S) 丝氨酸	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) 赖氨酸	AGA	(Arg/R) 精氨酸	A
	ATG[A]	(Met/M) 甲硫氨酸	ACG		AAG		AGG		G
G	GTT	(Val/V) 缬氨酸	GCT	(Ala/A) 丙氨酸	GAT	(Asp/D) 天冬氨酸	GGT	(Gly/G) 甘氨酸	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) 谷氨酸	GGA		A
	GTG		GCG		GAG		GGG		G

第二部分 变异的概念和分类

人类基因组上的变异主要分为三大类：

1. 单核苷酸变异（SNV），是单个碱基的改变，当 SNV 在人群中的频率大于 1%时被称为单核苷酸多态性，简称 SNP；

2. Indels，是 Insertion 和 Deletion 的简称，表示在基因组上某个位置上所发生的较短长度的线性片段的插入或缺失，长度通常在 50bp 以下，这个长度范围的变异通常可以采用 Smith-Waterman 的比对算法来精准获得，可以在目前短读长的测序数据中较好的检测出来；

3. 基因组结构性变异（Structure Variations，简称 SVs），包含长度在 50bp 以上的长片段序列的插入或者缺失、串联重复、染色体倒位。

人类遗传变异类型众多，但 SNP 变异可解释近 90%的个体表型差异，因此成为众多疾病研究关注的热点，也是我们本次课程的学习重点。

SNP 在群体中的发生频率不小于 1 %，其种类包括单个碱基的替换、插入和缺失等类型，其中替换又分为转换和颠换：

转换：同型碱基之间的替换，即嘌呤与嘌呤(G/A) 、嘧啶与嘧啶(C/ T) 间的替换；

颠换：发生在嘌呤与嘧啶(A/T、A/C、C/G、G/T) 之间的替换。

依据排列组合原理，SNP 一共有 6 种替换情况，即 A/ G、A/ T、A/ C、C/ G、C/ T 和 G/ T ，但事实上，由于不同碱基的化学结构和分子特性有差异，导致转换的发生频率多于颠换；在转换当中，C/T 转换又多于 G/A 转换。人类基因组上共有约 300 万个 SNP 位点，每隔 100 至 300 个碱基就会发生一处 SNP 变异。每 3 个 SNP 变异中有 2 个会是 C/T 转换。

第三部分 SNP 的功能分类

根据 SNP 变异的发生是否会影响个体表型，我们可将其分为 2 类：

同义突变：即 SNP 突变所致的编码序列的改变并不引发其所翻译的氨基酸序列改变， 但近年来有研究表明，同义突变可以通过涉及自缠结的局部界面错误折叠来改变蛋白质二聚化（参考文献：doi：10.1101/2021.10.26.465867）

非同义突变：指碱基序列的改变可使其对应翻译的氨基酸序列发生改变，从而影响蛋白质功能，这种改变常是导致生物性状改变的直接原因。非同义突变又可以分为错义突变、无义突变两种类型。错义突变是指编码的某种氨基酸的密码子变成另一种氨基酸密码子，使得多肽链的氨基酸种类和序列发生改变，错义突变通常会使多肽链丧失原有功能。无义突变是指编码某一氨基酸的密码子变成终止密码子 UAA、UGA 或 UAG，导致多肽链翻译中止，从而形成一条不完整的多肽链，使蛋白质的生物活性和功能改变。

通常我们还可按应用对 SNP 进行分类，以下是常见的几类：

（1）个体识别 SNP（IISNPs）

SNP 作为第三代遗传标记，具有较高遗传稳定性，几乎为零的循环突变率、所需扩增片段长度短、遍布全基因组，更加适用于高度降解 DNA 样本的鉴定。但是由于 SNP 为双等位基因，其多态性的信息含量低，需要联合多个位点才能达到个体识别的要求。在个体识别 SNP 位点的筛选研究中，耶鲁大学的 Kidd 实验室开展了系统深入的研究,并于 2006 年界定了筛选个体识别 SNP 位点的标准[2-5]：1.等位基因的平均杂合度>=0.4;2.不同人群间等位基因频率的差别小,Fst<0.06;3.筛选的 SNP 位点之间互不连锁。

（2）祖先信息 SNP（AIMSNPs）

在人类基因组的 SNPs 中有一部分是与种族起源相关的，不同人群之间基因频率差异非常大的多态性基因位点，被称为祖先信息标记-AIMs（Ancestry Informative Markers）. AIMs 包含有种群结构、种内和种间差异相关信息，可以定量的估计某个体可能的地域种族来源。本科课程中涉及的 AIMs 的筛选标准[6]：1.筛选出来的位点构建的体系符合 H-W 平衡和连锁平衡；2.人群特异性位点的选择中次等位基因频率（MAF）>0.01；3.等位基因频率差异（AFD）>0.5;4.群体遗传分化指数 Fst>0.3；

（3）表型相关 SNPs

SNP 是影响基因表达和功能的最主要遗传标记类型，能够预测个体的外部可视化特征，比如身高、肤色、瞳孔颜色、发色、脸部形态等。本课程中涉及到表型相关的 SNP 共有 24 个[7-10]，表征的相关表型有乳糖代谢能力、肌肉类型、酒精代谢能力、耳垢类型、头发颜色、瞳孔颜色。

（4）连锁信息性 SNP

线粒体（mtDNA）及 Y 染色体上的遗传标记带有家族特征，又被称为“谱系标记”。Y 染色体作为父系遗传，具有单倍型保持完整、突变率低、遗传稳定的特点，所以适合用作法医遗传标记，也可以用于人类进化中的物种起源、物种迁移以及遗传推断。由于集中在 mtDNA 和 Y 染色体上的 SNPs，对于个体识别能力较低，所以限制了这一类 SNPs 的应用范围。在本课程中的 SNPs 数据集没有涉及连锁信息性 SNP，但是随着人类单倍体型图计划(HapMap)的启动[11]，通过测定序列变异特征、变异频率以及其关联性，绘制人类基因组的单倍型块以及不同单倍型块的标记 SNPs，这将为寻找常染色体连锁信息的 SNPs 提供新的途径。

第四部分 SNP 与表型的关系

表型（phenotype），又称性状，是指一个生物体（或细胞）可以观察到的性状或特征，是特定的基因型与环境相互作用的结果。包括个体形态、功能等各方面的表现，如身高、肤色、血型、酶活力、药物耐受力乃至性格等。经典遗传学（genetics)是指由于基因序列改变(如基因突变等)所引起的基因功能的变化，从而导致表型发生可遗传的改变，基因对形状的控制可以通过酶的表达或者蛋白质的合成来实现，生物的表型主要通过蛋白质来表现。SNP 也决定了生物表型的

多样性。而表观遗传学（epigenetics）则是指在 DNA 序列没有发生改变的情况下，基因功能发生了可遗传的变化，并最终导致了表型的变化。

SNP 是影响基因表达和功能的最主要遗传标记类型，不仅能够预测个体的外部可视化特征，还可以对个体的营养代谢能力以及疾病风险进行预测。控制人类乳糖代谢能力、运动能力以及耳垢类型的一些 SNP 如下：

基因	SNP	基因型	表型
MCM6	rs4988235	GG	乳糖不耐受
MCM6	rs4988235	AA	乳糖耐受
MCM6	rs4988235	GA	乳糖耐受
MCM6	rs182549	CC	乳糖不耐受
MCM6	rs182549	TT	乳糖耐受
MCM6	rs182549	CT	乳糖耐受
ACTN3	rs1815739	TT	耐力型
ACTN3	rs1815739	CC	爆发型
ACTN3	rs1815739	CT	爆发型
ALDH2	rs671	GG	喝酒不会有或有较轻的脸红反应
ALDH2	rs671	AA	喝酒会有脸红反应
ALDH2	rs671	AG	喝酒不会有或有较轻的脸红反应
ABCC11	rs17822931	CC	很可能为湿型耳垢
ABCC11	rs17822931	TT	很可能为干型耳垢
ABCC11	rs17822931	TC	很可能为湿型耳垢

第五部分 变异的识别方法

freebayes 检测变异（freebayes v1.2.0）

单倍型，是单倍体基因型的简称，在遗传学上是指在同一染色体上进行共同遗传的多个基因座上等位基因的组合;通俗的说法就是若干个决定同一性状的紧密连锁的基因构成的基因型。 按照某一指定基因座上基因重组发生的数量，单倍型可以指至少两个基因座甚至整个染色体。

FreeBayes 是一个贝叶斯遗传变异检测器，旨在基于单倍型寻找小的多态性事件，如 SNP、indels，MNP（多核苷酸多态性）和复杂事件（复合插入和替换事件。 Freebayes 使用 reads 比对结果（带有 Phred+33 编码的质量分数的 BAM

文件，现在是标准的），对一个群体的任何数量的个体和参考基因组（FASTA 格式）进行比对，以确定该群体在参考基因组的每个位置最可能的基因型组合。

调用命令：freebayes -f ref.fa aln.bam >var.vcf

（1）相关参数说明：

–min-alternate-count 默认值 2

要求在单个个体中至少有这个支持替代等位基因的观察数，以评估该位置，双倍体默认值为 2

–genotype-qualities

计算基因型的边际概率

–min-mapping-quality 默认值 1

基于匹配质量值进行筛选，低于 Q 值的筛掉

如果前期 reads 已做质控，且仅作 call 变异，那么使用默认参数即可（对于 2 倍体生物），后续有其他需求可再加入其他参数。

（2）提取 SNP 以及 INDEL

freebayes 产生的 VCF 文件中 INFO 一列中的 tag 来专门来注释 snp、ins（插入）、del（缺失）、mnp（连续两个 snp 位点，如 ref 为 AT， alt 为 CG）以及 complex（composite insertion and substitution events），我们可以基于文件进行 SNP 和 INDEL 的提取的命令行：

grep ‘TYPE=snp’ freebayes.vcf > freebayes_snp.vcf

变异检测结果文件格式解读

VCF 格式是用于描述 SNP、INDEL 和 SV 结果的文本文件。我们以本课程的结果文件做如下说明：

```
S=1;NUMALT=1;ODDS=488.422;PAIRED=0;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=13609;QR=0;RO=0;RPL=349;RPP=760.854;RPPR=0;RPR=
rs3737576       1024    .       T       A,C,G   0       .       AB=0,0,0;ABP=0,0,0;AC=0,0,0;AF=0,0,0;AN=2;AO=0,3,0;CIGA
rs3737576       1049    .       C       G       2.67606e-14     .       AB=0;ABP=0;AC=0;AF=0;AN=2;AO=46;CIGAR=1X;DP=386
rs1698647       1056    .       C       T       3440.97 .       AB=0.497093;ABP=3.03555;AC=1;AF=0.5;AN=2;AO=171;CIGAR=1
rs1343469       1038    .       A       G       18794.1 .       AB=0.497838;ABP=3.08542;AC=1;AF=0.5;AN=2;AO=921;CIGAR=1
rs11239930      1025    .       G       A,C,T   11072.4 .       AB=0.478571,0.00357143,0;ABP=7.47733,2400.44,0;AC=1,0,0
rs7554936       1026    .       C       A,G,T   64125.9 .       AB=0,0,0;ABP=0,0,0;AC=0,0,2;AF=0,0,1;AN=2;AO=0,0,2400;0
rs3829868       1035    .       C       A,G,T   4.24306e-14     .       AB=0,0,0;ABP=0,0,0;AC=0,0,0;AF=0,0,0;AN=2;AO=0,
rs2814778       1028    .       T       A,C,G   0       .       AB=0,0,0;ABP=0,0,0;AC=0,0,0;AF=0,0,0;AN=2;AO=0,8,0;CIGA
rs560681        1029    .       A       C,G,T   11044.6 .       AB=0,0.50565,0;ABP=0,3.30474,0;AC=0,1,0;AF=0,0.5,0;AN=2
rs10801520      1046    .       C       A,G,T   6529.65 .       AB=0,0,0.496885;ABP=0,0,3.06442;AC=0,0,1;AF=0,0,0.5;AN=
rs1106201       1052    .       C       A,G,T   12231   .       AB=0,0,0.640821;ABP=0,0,154.07;AC=0,0,1;AF=0,0,0.5;AN=2
rs2013162       1022    .       C       A,G,T   4.08875e-13     .       AB=0,0,0;ABP=0,0,0;AC=0,0,0;AF=0,0,0;AN=2;AO=0,
rs2292564       1051    .       G       A,C,T   7730.22 .       AB=0.441573,0,0;ABP=29.3998,0,0;AC=1,0,0;AF=0.5,0,0;AN=
rs1294331       1057    .       C       A,G,T   5827.6   .       AB=0,0,0.420757;ABP=0,0,41.8988;AC=0,0,1;AF=0,0,0.5;AN=
rs10495407      1032    .       G       A,C,T   0       .       AB=0,0,0;ABP=0,0,0;AC=0,0,0;AF=0,0,0;AN=2;AO=2,0,0;CIGA
rs891700        1040    .       CTTTTTTTTTGAAG  CTTTTTTTTGAAG     1207.91 .       AB=0.175187;ABP=1347.35;AC=1;AF=0.5;AN=
rs891700        1058    .       A       C,G,T   16750.1 .       AB=0,0.52992,0;ABP=0,14.7051,0;AC=0,1,0;AF=0,0.5,0;AN=2
rs1413212       1040    .       T       A,C,G   41238.9 .       AB=0,0,0;ABP=0,0,0;AC=0,2,0;AF=0,1,0;AN=2;AO=0,1546,0;0
rs876724        1050    .       C       A,G,T   56543.8 .       AB=0,0,0;ABP=0,0,0;AC=0,0,2;AF=0,0,1;AN=2;AO=0,0,2119;0
rs798443        1047    .       G       A,C,T   9899.74 .       AB=0,0,0;ABP=0,0,0;AC=2,0,0;AF=1,0,0;AN=2;AO=370,0,0;CI
rs1109037       1044    .       G       A,C,T   18307.1 .       AB=0,0,0;ABP=0,0,0;AC=2,0,0;AF=1,0,0;AN=2;AO=689,0,0;CI
rs1049500       1026    .       G       A,C,T   8792.48 .       AB=0.500597,0,0;ABP=3.01289,0,0;AC=1,0,0;AF=0.5,0,0;AN=
rs1876482       1023    .       G       A,C,T   12047   .       AB=0.501292,0.000861326,0;ABP=3.02713,2515.41,0;AC=1,0,
```

第一列 ID：variant 的 rsID 号，对应 dbSNP 里的 rs 编号

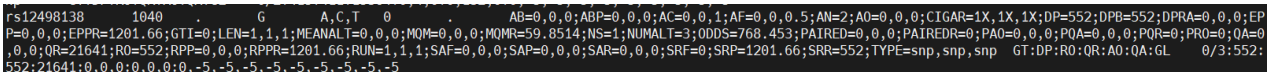
第二列 POS：变异位点相对于参考基因组所在的位置

第四列和第五列 **REF 和 ALT**：参考基因组中所对应的碱基和研究对象基因组（Variant）中所对应的碱基

第七列 INFO：variant 的详细信息，其中一下两种信息比较重要

GT： 表示这个样本的基因型，对于一个二倍体生物，GT 值表示的是这个样本在这个位点所携带的两个等位基因。0 表示跟 REF 一样；1 表示表示跟 ALT 一样；2 表示第二个 ALT。当只有一个 ALT 等位基因的时候，0/0 表示纯和且跟 REF 一致；0/1 表示杂合，两个 allele 一个是 ALT 一个是 REF；1/1 表示纯和且都为 ALT

DP： 覆盖到这个位点的 reads 数量，相当于这个位点的深度（并不是所有的 reads 数量，而是达到一定质量值要求的 reads 数）。



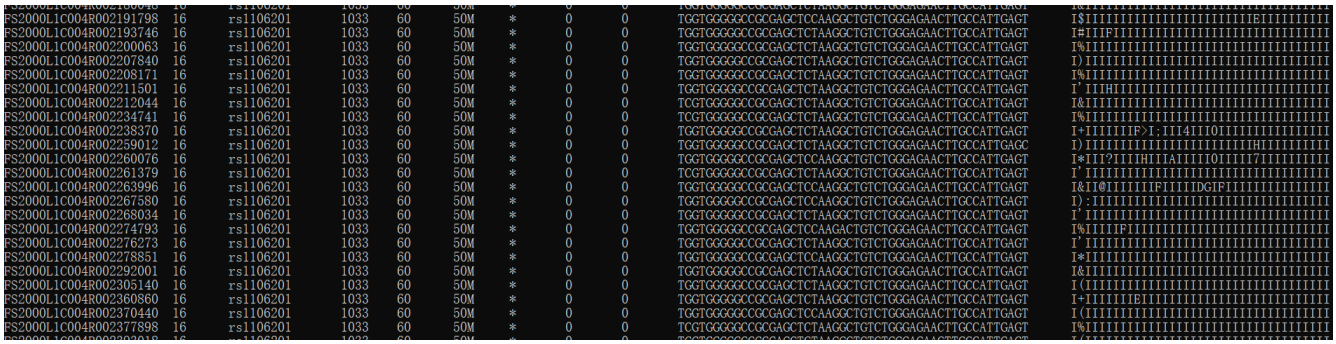
基因型结果文件解读

SampleName	SNP_Marker		SNP_Genotype		AlleleRatio		Type	TotalDepth	A	T	C	G	QC_Info
liyuqi	ID.rs1490413	GA	0.7873	Hete	1128	630	1	1	496	-			
liyuqi	ID.rs5745448	CC	0	Homo	348	0	0	348	0	-			
liyuqi	Popu.rs3737576	TT	0	WT	754	0	751	3	0	-			
liyuqi	ID.rs11239930	GA	0.9241	Hete	1120	536	0	4	580	-			
liyuqi	Popu.rs7554936	TT	0	Homo	2404	0	2400	4	0	-			
liyuqi	ID.rs3829868	CC	0	WT	387	0	1	386	0	-			
liyuqi	Popu.rs2814778	TT	0	WT	2950	0	2942	8	0	-			
liyuqi	ID.rs560681	AG	0.9777	Hete	1062	525	0	0	537	-			
liyuqi	ID.rs10801520	CT	0.9876	Hete	642	0	319	323	0	-			
liyuqi	ID.rs1106201	CT	0.5605	Hete	877	0	562	315	0	-			
liyuqi	ID.rs2013162	CC	0	WT	1049	0	2	1047	0	-			
liyuqi	ID.rs2292564	GA	0.7907	Hete	890	393	0	0	497	-			
liyuqi	ID.rs1294331	CT	0.7264	Hete	713	0	300	413	0	-			
liyuqi	ID.rs10495407	GG	0	WT	1798	2	0	0	1796	-			
liyuqi	ID.rs891700	AG	0.8871	Hete	1504	707	0	0	797	-			
liyuqi	ID.rs1413212	CC	0	Homo	1547	0	1	1546	0	-			
liyuqi	ID.rs876724	TT	0	Homo	2119	0	2119	0	0	-			
liyuqi	Popu.rs798443	AA	0	Homo	371	370	0	0	1	-			
liyuqi	ID.rs1109037	AA	0	Homo	690	689	0	0	1	-			
liyuqi	ID.rs1049500	GA	0.9976	Hete	837	419	0	0	418	-			
liyuqi	Popu.rs1876482	GA	0.9931	Hete	1161	582	0	1	578	-			
liyuqi	Popu.rs1834619	AA	0	Homo	847	845	0	0	2	-			
liyuqi	ID.rs3899750	GG	0	Homo	612	0	0	0	612	-			
liyuqi	ID.rs11123823	AG	0.9898	Hete	390	194	0	0	196	-			
liyuqi	Popu.rs3827760	NA	-	-	0	0	0	0	0	Not Detected			
liyuqi	Popu.rs260690	CC	0	WT	1537	0	0	1537	0	-			
liyuqi	ID.rs993934	GG	0	Homo	693	0	0	0	693	-			
liyuqi	Traits.rs4988235		GG	0	WT	1060	0	0	1	1059	-		
liyuqi	Traits.rs182549	CC	0	WT	641	0	0	641	0	-			
liyuqi	Popu.rs6754311	CC	0	Homo	1680	0	0	1680	0	-			
liyuqi	Popu.rs10497191	CC	0	Homo	450	0	1	449	0	-			
liyuqi	ID.rs12997453	GG	0	Homo	1357	0	0	0	1357	-			
liyuqi	ID.rs907100	GG	0	WT	1340	1	0	0	1339	-			

这一步是基于 VCF 文件的 GT 和 DP 两列信息进行的提取，找出每个变异的基因型以列表的形式表示。

四、实验步骤

1. 输入文件：序列比对结果的 bam 文件



2. 为 bam 文件建立索引

samtool index 基于坐标排序后 bam 或者 cram 的文件创建索引，生成以.bai 或者.crai 为后缀的索引文件，这样可以快速的访问 bam 文件

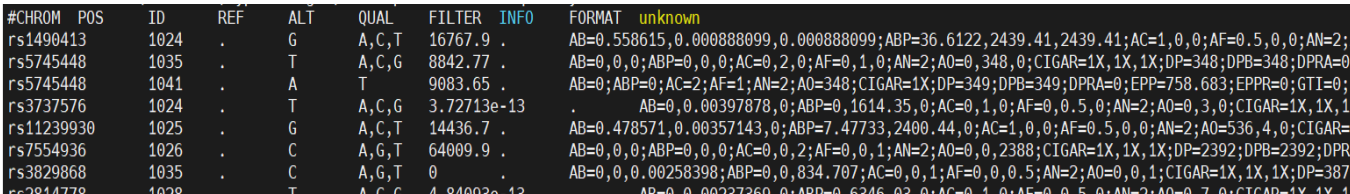
samtools index ./02_align/test.clean.sort.uniq.bam

3. 选用 freebayes 进行变异检测

/Pipeline/FIS.Traits/tools/freebayes -m 30 -q 20 -f ./00_ref/MGI358.SNP.fa -@ ./00_ref/alleles_all.vcf.gz -t ./00_ref/target.358.SE50.subSNP.bed --report-all-haplotype-alleles ./02_align/test.clean.sort.uniq.bam > ./03_SNPCalling/test.clean.SNP.vcf

4. 查看变异检测结果文件

less -S ./03_SNPCalling/test.clean.SNP.vcf



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
rs1490413	1024	.	G	A,C,T	16767.9	.		AB=0.558615,0.000888099,0.000888099;ABP=36.6122,2439.41,2439.41;AC=1,0,0;AF=0.5,0,0;AN=2;	
rs5745448	1035	.	T	A,C,G	8842.77	.		AB=0,0,0;ABP=0,0,0;AC=0,2,0;AF=0,1,0;AN=2;AO=0,348,0;CIGAR=1X,1X,1X;DP=348;DPB=348;DPRA=0	
rs5745448	1041	.	A	T	9083.65	.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=348;CIGAR=1X;DP=349;DPB=349;DPRA=0;EPP=758.683;EPPR=0;GTI=0;	
rs3737576	1024	.	T	A,C,G	3.72713e-13	.		AB=0,0.00397878,0;ABP=0,1614.35,0;AC=0,1,0;AF=0,0.5,0;AN=2;AO=0,3,0;CIGAR=1X,1X,1	
rs11239930	1025	.	G	A,C,T	14436.7	.		AB=0.478571,0.00357143,0;ABP=7.47733,2400.44,0;AC=1,0,0;AF=0.5,0,0;AN=2;AO=536,4,0;CIGAR=	
rs7554936	1026	.	C	A,G,T	64009.9	.		AB=0,0,0;ABP=0,0,0;AC=0,0,2;AF=0,0,1;AN=2;AO=0,0,2388;CIGAR=1X,1X,1X;DP=2392;DPB=2392;DPR	
rs3829868	1035	.	C	A,G,T	0	.		AB=0,0,0.00258398;ABP=0,0,834.707;AC=0,0,1;AF=0,0,0.5;AN=2;AO=0,0,1;CIGAR=1X,1X,1X;DP=387	
rs2814778	1028	.	T	A,C,G	4.84093e-13	.		AB=0,0.00237369,0;ABP=0,6346.03,0;AC=0,1,0;AF=0,0.5,0;AN=2;AO=0,7,0;CIGAR=1X,1X,1	
rs560681	1029	.	A	C,G,T	14155.8	.		AB=0,0.502841,0;ABP=0,3.08433,0;AC=0,1,0;AF=0,0.5,0;AN=2;AO=0,531,0;CIGAR=1X,1X,1X;DP=105	
rs10801520	1046	.	C	A,G,T	8463.98	.		AB=0,0,0.496885;ABP=0,0,3.06442;AC=0,0,1;AF=0,0,0.5;AN=2;AO=0,0,319;CIGAR=1X,1X,1X;DP=642	
rs1106201	1052	.	C	A,G,T	14678.6	.		AB=0,0,0.639175;ABP=0,0,149.887;AC=0,0,1;AF=0,0,0.5;AN=2;AO=0,0,558;CIGAR=1X,1X,1X;DP=873	
rs2013162	1022	.	C	A,G,T	9.31541e-13	.		AB=0,0,0.00190658;ABP=0,0,2263.55;AC=0,0,1;AF=0,0,0.5;AN=2;AO=0,0,2;CIGAR=1X,1X,1	
rs2292564	1051	.	G	A,C,T	10360.8	.		AB=0.440315,0,0;ABP=30.4863,0,0;AC=1,0,0;AF=0.5,0,0;AN=2;AO=391,0,0;CIGAR=1X,1X,1X;DP=888	
rs1294331	1057	.	C	A,G,T	7927.59	.		AB=0,0,0.419944;ABP=0,0,42.6458;AC=0,0,1;AF=0,0,0.5;AN=2;AO=0,0,299;CIGAR=1X,1X,1X;DP=712	
rs10495407	1032	.	G	A,C,T	0	.		AB=0.00111235,0,0;ABP=3889.97,0,0;AC=1,0,0;AF=0.5,0,0;AN=2;AO=2,0,0;CIGAR=1X,1X,1X;DP=179	
rs891700	1058	.	A	C,G,T	21269	.		AB=0,0.529294,0;ABP=0,14.206,0;AC=0,1,0;AF=0,0.5,0;AN=2;AO=0,795,0;CIGAR=1X,1X,1X;DP=1502	

5. 转换为基因型格式

#准备vcf列表文件

echo -e "test\t /home/LY/03_SNPCalling/test.clean.SNP.vcf

\t1" > ./03_SNPCalling/vcf.list (注意：将LY换成自己的用户名)(注意：

是一行命令，准备vcf列表文件，其中-e参数支持echo命令输出转义

字符，“\t”表示制表符Tab)

#输出基因型结果文件

#将SNP映射至基因型

#使用脚本将SNP转换为基因，并查看

/Pipeline/FIS.Traits/bin/SNP/vcf2geno_free -in ./03_SNPCalling/vcf.list -dir ./03_SNPCalling/ -std ./00_ref/alleles_all.vcf.gz -summary （注意：是一行命令）

#使用less命令查看生成的输出文件

less ./03_SNPCalling/test/test.genotype

SampleName	SNP_Marker	SNP_Genotype	AlleleRatio	Type	TotalDepth	A	T	C	G
test	ID.rs1490413	GA	0.0000	Hete	495			495	-
test	ID.rs5745448	CC	0	Homo	0		0		-
test	Popu.rs3737576	TC	0.0000	Hete	751		751		-
test	ID.rs11239930	GA	0.0000	Hete	580			580	-
test	Popu.rs7554936	TT	0	Homo	4		4		-
test	ID.rs3829868	CT	0.0000	Hete	386		386		-
test	Popu.rs2814778	TC	0.0000	Hete	2942		2942		-
test	ID.rs560681	AG	0.0000	Hete	525	525			-
test	ID.rs10801520	CT	0.0000	Hete	323		323		-
test	ID.rs1106201	CT	0.0000	Hete	315		315		-
test	ID.rs2013162	CT	0.0000	Hete	1047		1047		-
test	ID.rs2292564	GA	0.0000	Hete	497			497	-
test	ID.rs1294331	CT	0.0000	Hete	413		413		-
test	ID.rs10495407	GA	0.0000	Hete	1796			1796	-
test	ID.rs891700	AG	0.0000	Hete	707	707			-
test	ID.rs1413212	CC	0	Homo	1		1		-
test	ID.rs876724	TT	0	Homo	0		0		-
test	Popu.rs798443	AA	0	Homo	1			1	-
test	ID.rs1109037	AA	0	Homo	1			1	-
test	ID.rs1049500	GA	0.0000	Hete	418			418	-
test	Popu.rs1876482	GA	0.0000	Hete	578			578	-

五、预期实验结果

以上vcf文件即为本次实验结果，记录个体相比于参考序列的变异位点信息，可直接用于后续复杂数据分析。