外显子捕获结题报告

2010-11-22

内容

1 项目信息

PROJECT NAME			
CONTRACT NUMBER			
SAMPLE INFORMATION			
Species Information			
Genome Information			
Additional Information			
CUSTOMER INFORMATION			
PI			
Contact Person			
Company Name			
Contact Methods			
Name		Tel	
		E-mail	
Name		Tel	
		E-mail	
CONTACT INFORMATION (BGI)			
Sales Information			
Name		Tel	
		E-mail	
Name		Tel	
		E-mail	
Customer Service			
Name		Tel	
		E-mail	
Name		Tel	
		E-mail	
PROJECT DIRECTOR APPROVAL			
THE RESULTS HAVE BEEN APPROVED AND CAN BE SUBMITTED Signature: Date:			

2 工作流程介绍

采用 AglientSureSelect外显子靶向序列富集系统和 NimbleGenSeqCap EZ人全外显子捕获系统。这两个系统都采用液相系统进行高特异性和高覆盖率的外显子区域捕获。

2.1 Agilent液相捕获平台

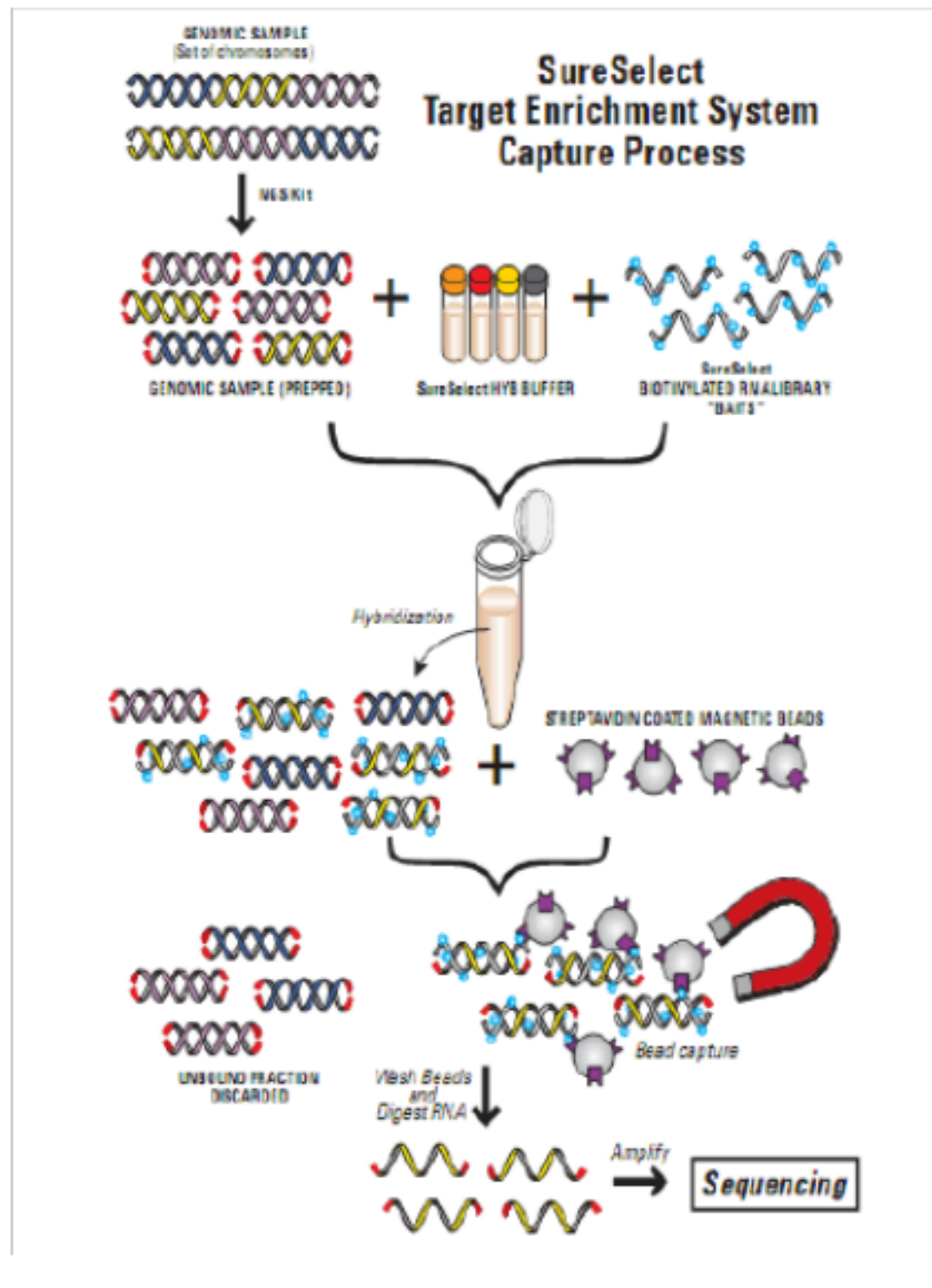


图 2.1Aglient 外显子捕获和测序流程

基本流程：首先将基因组 DNA 随机打断成 150-200bp 左右的片段，随后在片段两端分别连接上接头制备杂交文库。文库经纯化后经过 LM-PCR 的线性扩增与 SureSelect Biotinylated RNA Library (BAITS) 进行杂交富集，再经过 LM-PCR 的线性扩增，文库检测合格后即可上机测序 （Hiseq2000测序仪）。对每个捕获文库进行高通量测序并保证测序深度达到要求，原始图像文件经过 Illumina

basecalling Software 1.7进行碱基读取，获得读长为　90bp 双末端序列（ reads)。

2.2 NimbleGen液相捕获平台

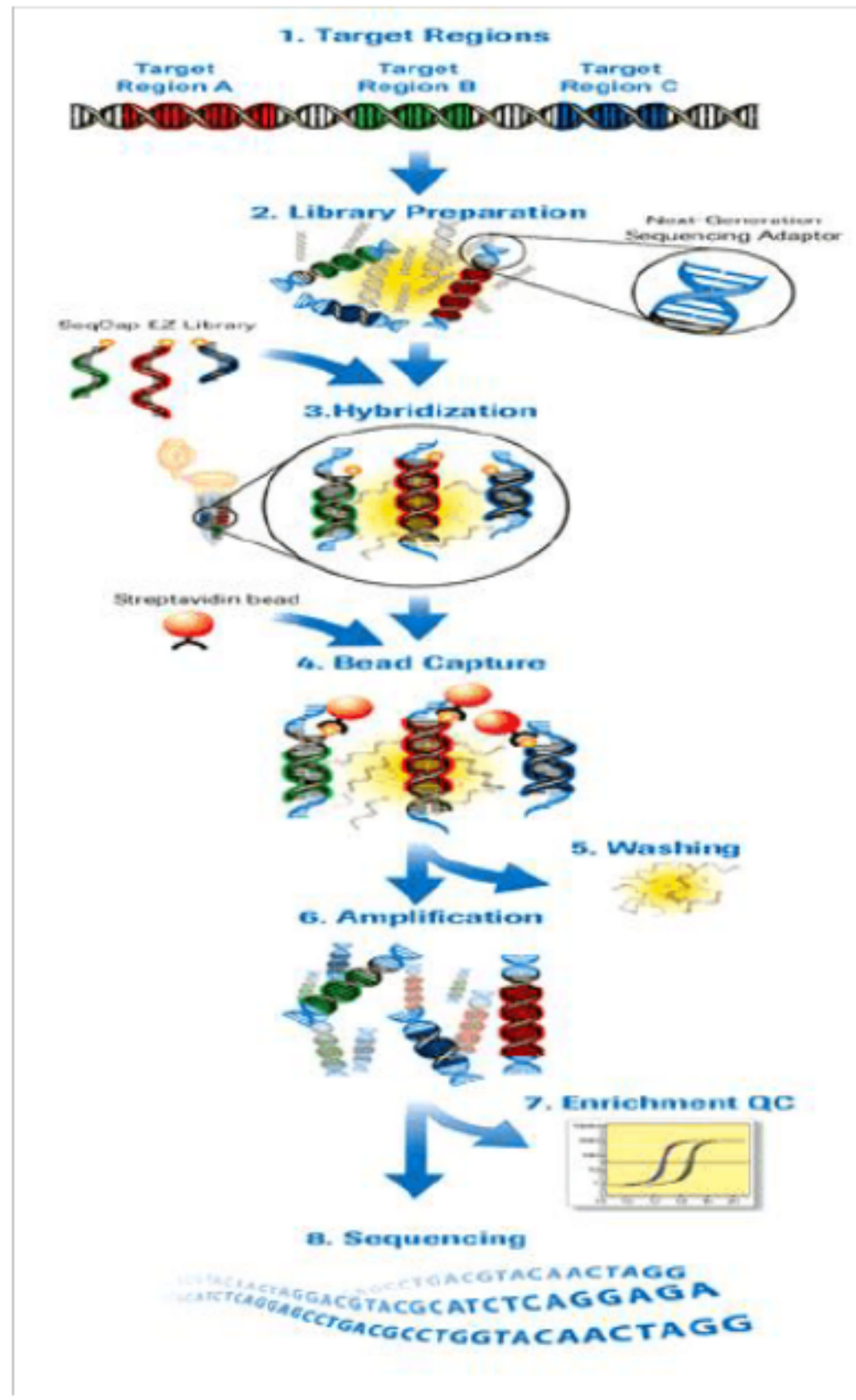


图 2.2NimbleGen 外显子捕获和测序流程

基本流程：首先将基因组　DNA 随机打断成　200-300bp 左右的片段，随后在片段两端分别连接上接头制备杂交文库。文库经纯化后经过　　LM-PCR 的线性扩增与 Biotinylated DNA Library 进行杂交富集，再经过　LM-PCR 的线性扩增，文库检测合格后即可上机测序　（Hiseq2000 测序仪）。对每个捕获文库进行高通量测序并保证测序深度达到要求，原始图像文件经过　　Illumina basecalling Software 1.7 进行碱基读取，获得读长为　90bp 双末端序列（ reads)。

2.3 生物信息分析流程

测序完成之后，下机数据为 fastq 文件格式，随后对数据进行信息分析，分析流程如下：

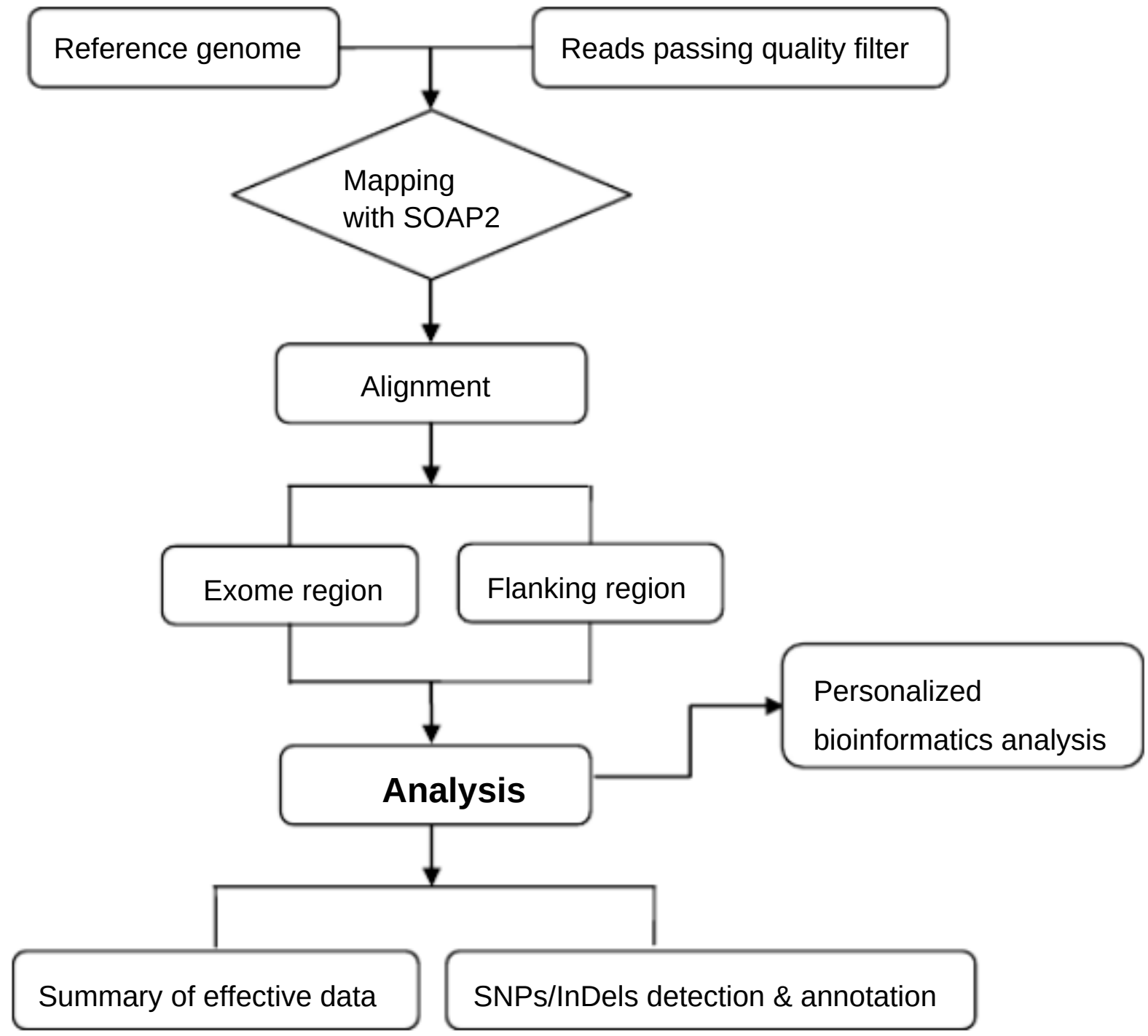


图 2.3 外显子测序信息分析流程

（1）SOAPaligner是华大自主研发的比对软件，用于将高质量的原始 reads 比对到参考基因组上，详细说明见信息分析软件及参数介绍部分，或者登录网站 http://soap.genomics.org.cn/ 仅比对到参考基因组的 reads 用于后续分析。

（2）计算得到的 Coverage和 Depth 是指目标区域的覆盖度和测序深度，计算时所用的数据是所有比对到参考基因组的 reads。

3 分析报告

结果

3.1 标准生物信息分析

3.1.1 数据产出统计

基本数据分析统计结果主要包括：测定的序列（ reads）长度、 reads 数量、数据产量、 reads 序列与参考基因组序列比对结果、目标外显子区域测序深度及覆盖度分析、目标外显子区域 SNP检测及注释等。具体统计结果参照表 3.1。

表 3.1 统计量详细说明

统计量	定义及计算方法
Target region (bp)	设计探针覆盖的区域，作为目标区域，用于捕获外显子
Raw reads	测序得到的原始 reads 个数
Raw data yield (Mb)	原始 reads 产量，即所有碱基个数 (以 Mb 为单位)
Reads mapped to genome	比对到参考基因组上的 reads 个数
Reads mapped to target region	比对到目标区域上的 reads 个数
Data mapped to target region (Mb)	比对到目标区域上的碱基个数 (以 Mb 为单位)
Mean depth of target region	目标区域的平均深度
Coverage of target region (%)	目标区域的覆盖度
Average read length (bp)	平均 read 长度
Rate of nucleotide mismatch (%)	碱基错配率
Fraction of target covered >= 4x	目标区域深度 >= 4x 的碱基覆盖度
Fraction of target covered >=10x	目标区域深度 >= 10x 的碱基覆盖度
Capture specificity (%)	唯一比对到参考基因组的 reads 中, 唯一比对到目标区域的 reads 所占的比例
Reads mapped to flanking region	比对到侧翼区 (每段目标区域两侧扩展 200bp) 的 reads 数
Mean depth of flanking region	侧翼区域的平均深度
Coverage of flanking region	侧翼区域的覆盖度
Fraction of flanking region covered >= 4x	侧翼区域深度 >= 4x 的碱基覆盖度
Fraction of flanking region covered >= 10x	侧翼区域深度 >= 10x 的碱基覆盖度
Fraction of unique mapped bases on or near target	唯一比对到目标区域和侧翼区域的碱基比例

Duplication rate	reads 重复率
Mean depth of chrX	X 染色体的平均深度
Mean depth of chrY	Y 染色体的平均深度
Sample gender	样本性别
Gender test result	性别测试结果

3.1.2 目标区域单碱基深度分布图

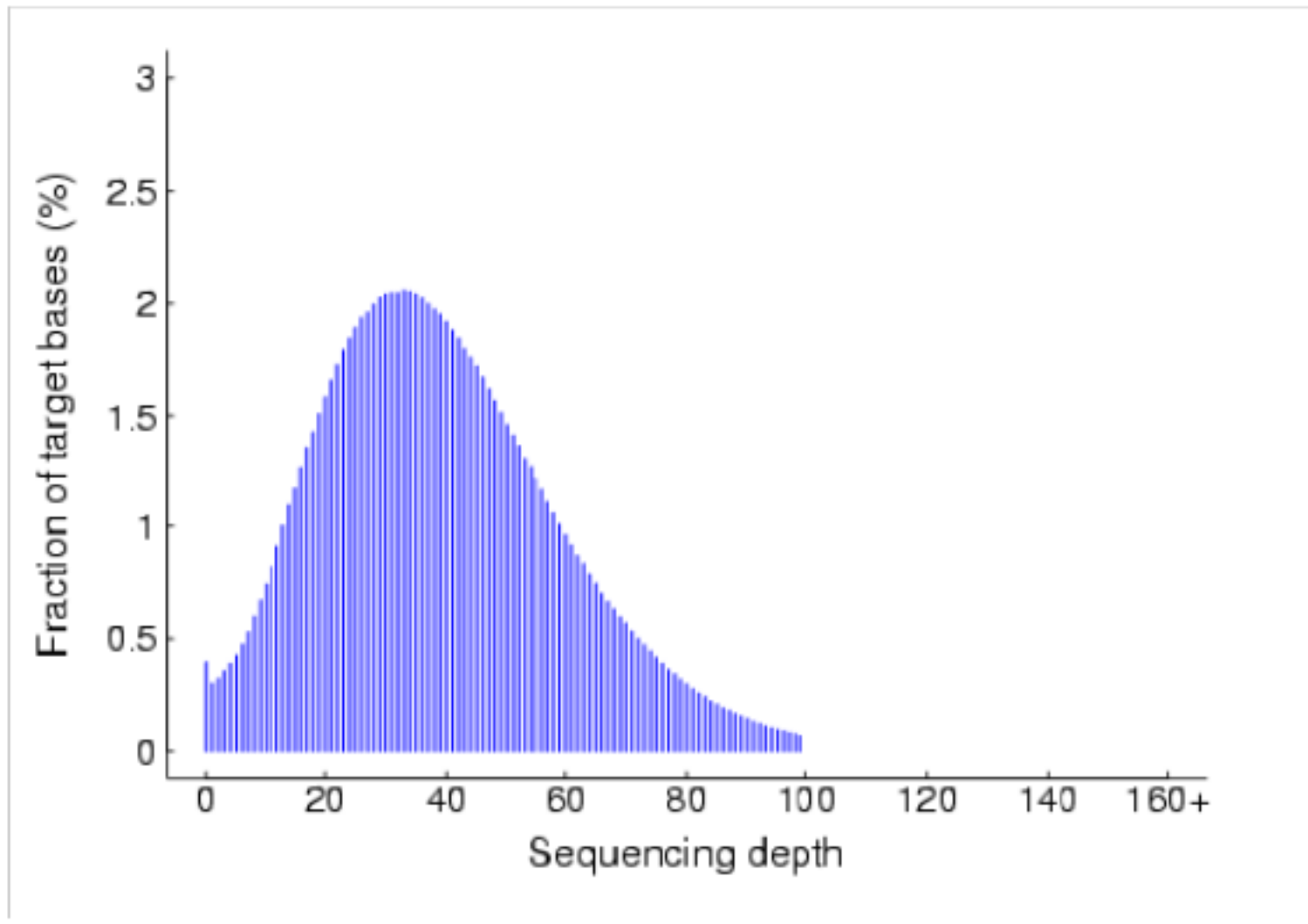


图 3.1 目标区域的单碱基深度分布图

横坐标代表测序深度， 纵坐标代表目标区域上对应深度的碱基数占总碱基数的百分比。目标区域的单碱基分布近似服从泊松分布。

3.1.3 外显子捕获测序的均一性

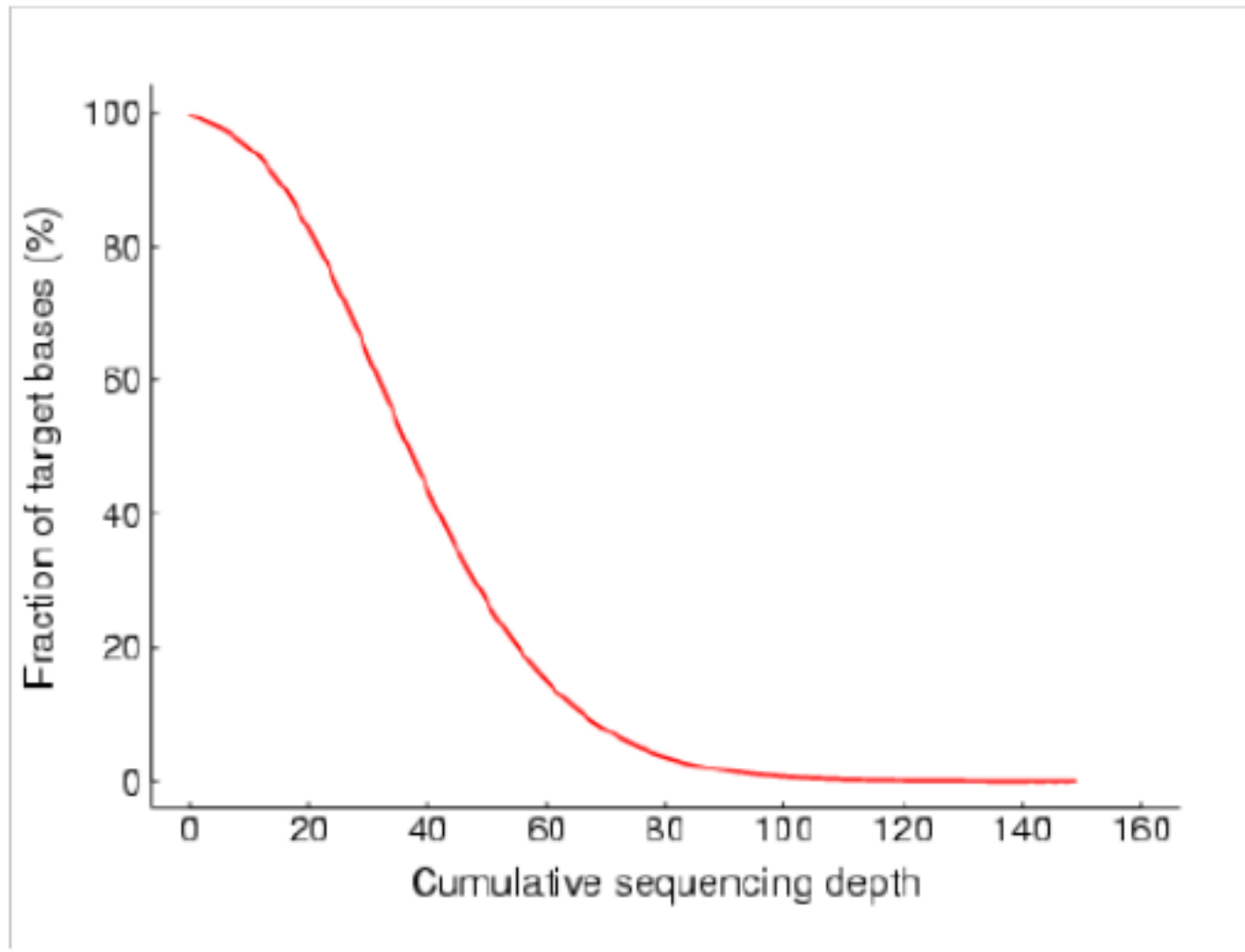


图 3.2 目标区域的累积深度分布图

横坐标代表累积测序深度， 纵坐标代表目标区域上大于或等于对应累积深度的碱基数占总碱基数的百分比。

根据表中基本数据的统计量及单碱基深度分布图和累积深度分布图， 除了可以得到通过外显子捕获的样本基本信息外， 还可以判断捕获的数据是否符合要求，即进行质控。

3.1.4 一致序列组装和 SNP检测

对于 soap 比对之后的结果，我们采用 SOAPsnp软件进行一致序列组装，得到每个位点的基因型，进而进行 SNP检测。

生成文件如下：

CNS文件 (*.cns)：包含位点的基因型等详细信息。

SNP文件 (*.snp、*.snp.filter) ：其中 *.snp 包含 *.cns 中所有的可能 SNP位点，即基因型与参考序列基因型不一致的位点； *.snp.filter 包含最终的 SNP集合，即对 *.snp 中所有 SNP位点按一定标准（如质量值、深度等）进行过滤后所得到的高置信度的 SNP结果。

3.1.5 SNP注释

对最终检测出的 SNP结果, 即*.snp.filter 中所有 SNP进行注释分类, 每个 SNP 的详细信息见 gff 文件, gff 文件的详细说明见数据文件格式说明部分。

对 SNP的统计信息见表 3.2。

表 3.2 SNP 统计

Categories	SampleID
Number of genomic positions for calling SNPs (1)	87,444,832
Number of high-confidence genotypes (2)	63,608,643
Number of high-confidence genotypes in target regions	33,006,340
Number of known dbSNP sites in target region	192,415
Coverage of known dbSNP sites (3)	178963 (93.01%)
Number of detected SNPs on target	
Number of detected SNPs near target	
Total number of SNPs	45,671
Synonymous-coding	8,036
Missense	6,817
Nonsense	51
Readthrough	9
Splice site (4)	347
Intron	27,151
5' UTRs (5)	1,381
3' UTRs	1,548
Intergenic	331

注：

(1) Number of genomic positions for calling SNPs ：指*.cns 文件中的所有位点，即包括捕获的目标区域和前后 200bp 的侧翼区域。

(2) Number of high-confidence genotypes ：*.cns 文件中质量值不低于 20 的碱基数

(3) Number of high-confidence genotypes in target regions ：*.cns 文件中，目标区域内质量值不低于 20 的碱基数

(4) Number of known dbSNP sites in target region ：目标区域内所有在 dbSNP 数据库中已知SNP位点数。

(5) Coverage of known dbSNP sites ：在目标区域内，我们所定义的高可信度的位点 (即*.cns 文件中碱基质量值不低于 20 的位点)所覆盖到的已知 SNP位点数（ dbSNP）的比例。

(6) Total number of SNPs :最终得到的高可信度 (采用一定的过滤标准过滤之后的结果)的 SNP 位点数。

(7) Splice site：外显子与内含子交界处 4bp 的内含子 SNP位点？

(8) 5' UTRs：指初始密码子上游 200bp ；3' UTRs则指终止密码子下游 200bp ；

3.1.6 插入/缺失(indels)检测

通过对获得的测序 reads 重新组装，可发现外显子区的插入与缺失 (InDels)。重新组装是运用 SOAPdenovo (Li et al. Genome Res, 2010)软件，随后，通过 LASTZ 软件将组装的一致性序列比对到参考基因组上。将比对结果输入到 axtBest (Schwartz et al. Genome Res, 2003), 以将 orthologous 比对与 paralogous 比对分离。最后，检测到比对的断裂点（ breakpoints），以及进行后续的 Indels 的注释。

3.1.7 插入/缺失(indels)注释

对检测出的 indels 结果进行统计，举例统计信息见下表：

表 3.3 InDels 统计

SampleID	SH002	SH003	SH005	SH029	SH048	SH050
Total number of InDels	640	466	436	629	579	635
Ins-coding (1)	82	57	55	62	70	74
Del-coding (2)	79	55	56	78	74	73
5' UTRs	13	12	4	10	10	13
3' UTRs	23	20	17	22	16	17
Intergenic	593	311	341	533	563	513
Total insertion	345	240	220	347	299	331
Total deletion	295	226	216	282	280	304
Heterozygous InDels	442	254	226	440	383	447
Homozygous InDels	198	212	210	189	196	188

(1) 指编码区的插入（ insertion ）

(2) 指编码区的缺失（ deletion ）

3.2 个性化分析

3.2.1 氨基酸替换预测

在遗传学中，遗传变异对表型的影响具有很重要的意义。引起蛋白序列中单氨基酸替换的遗传变异类型为非同义的 SNP(non-synonymous single nucleotide polymorphism, nsSNP)，非同义的 SNP很可能影响蛋白质的功能，从而影响表型。

我们可采用 SIFT (Sorting Intolerant From Tolerant) 软件和 PolyPhen(Polymorphism Phenotyping)软件进行预测，预测单氨基酸替换对蛋白质

功能的影响。

SIFT简介

SIFT(**S**orting **I**ntolerant **F**rom **T**oleran)是一个用于预测氨基酸替换对蛋白质功能影响的软件，它可以判断出这个氨基酸置换在蛋白质功能上是无害的（functionally neutral）的还是有害的（deleterious），研究者可以由这个结果推断是否要对这种替换做进一步的研究。详细信息见 http://sift.jcvi.org/ 。SIFT预测结果举例如下：

表 3.4 SIFT 预测结果举例

Coordinates	Codons	Substitution	SNP Type	Prediction	Score[1]	Median Info [2]	Gene Name
10,17125881,1, C/G	AGG-AGc	R1260S	Nonsynonymous	DAMAGING *Warning! Low confidence.	0	3.38	CUBN
10,22062710,1, C/T	ACC-AtC	T835I	Nonsynonymous	TOLERATED	0.06	3.4	MLLT10
11,116138821, 1,G/A	CGT-tGT	R232C	Nonsynonymous	DAMAGING	0.02	3.05	BUD13
1,111830738,1, G/A	ACC-ACt	T147T	Synonymous	N/A	N/A	N/A	ADORA3
15,29004656,1, C/T	CCG-CtG	P736L	Nonsynonymous	DAMAGING	0.01	3.05	MTMR15
19,12624007,1, G/A	CCG-CtG	P669L	Nonsynonymous	TOLERATED	0.75	3.02	MAN2B1
19,15137764,1, C/T	CGG-CaG	R1834Q	Nonsynonymous	TOLERATED	1	3.03	NOTCH3
2,10103771,1, G/A	CGG-CaG	R29Q	Nonsynonymous	DAMAGING	0.03	3.03	KLF11
2,31426431,1, C/T	GCA-aCA	A932T	Nonsynonymous	DAMAGING	0	3.05	XDH
3,128822344,1, G/A	ATG-A Ta	M793I	Nonsynonymous	TOLERATED	0.13	3.05	MCM2
4,69830873,1, T/A	AGA-AGt	R428S	Nonsynonymous	DAMAGING	0.01	2.95	UGT2A3
9,138364025,1, G/A	GTG-aTG	V459M	Nonsynonymous	TOLERATED	0.15	3.36	GPSM1
X,48432692,1, C/T	CCT-tCT	P460S	Nonsynonymous	TOLERATED	0.24	4.32	WAS
7,102503456,1, G/T	-	NA	NA	Not scored	NA	NA	

注：

[1] Coordinates：突变发生的染色体编号及坐标位置

[2] Codons：密码子的变化情况

[3] Substitution ：氨基酸的替换信息

[4] SNP Type：SNP的类型

[5] Prediction ：预测结果 (damaging/tolerated)

[6] Score：SIFT对于一个氨基酸置换的预测结果被计算为一个标准化的分值，变化范围从 0 到 1，当这个值大于 0.05 的时候表示这个突变是可以容忍的，即对蛋白质功能没有影响或影响很小；小于等于 0.05 的时候则说明这个突变是有害的，即对蛋白质功能有较大影响。

[7] Median Info ：中值信息。用来衡量用于比对的蛋白质序列的多样性情况，变化范围从 0 到 4.32，理论上应该在 2.75 到 3.5 之间。如果这个值大于 3.25，系统将会发出警告信息，因为这说明本次预测分析是基于一系列紧密联系的蛋白质序列的，结果可信度可能不高。

[8] Gene Name：发生替换所在的基因名称

PolyPhen简介

PolyPhen(Polymorphism Phenotyping)也是一种预测氨基酸置换对蛋白质结构和功能影响的工具。详细信息见 http://genetics.bwh.harvard.edu/pph/

PolyPhen 预测结果主要包括三部分， Query、Prediction、Details。Query 部分包含查询信息，与输入文件类似。 Prediction 部分显示了预测的结果。 Details 部分显示了 PolyPhen 预测的详细信息，包括所有的数据信息。我们着重关注的为预测结果， 如“ This variant is predicted to be probably damaging”。详细说明见：http://genetics.bwh.harvard.edu/pph/pph_help_text.html#OutputQueryAccession

举例如下：

表 3.5 PolyPhen 预测结果举例

Query				
Acc number	Position	AA 1	AA 2	Description
21040341	176	C	Y	.1|hemochromatosis protein isoform 3 precursor, hereditary haemochromatosis protein[Homo sapiens]
Prediction				
This variant is predicted to be probably damaging				
Prediction	Available data	Prediction basis	Substitution effect	Prediction data(1)
Probably damaging	FT alignment	alignment	N/A	PSIC score difference:2.943
Details				

PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS					
Score1[2]	Score2[3]	|Score1-Score2|	Observations[4]	Diagnostics[5]	Multiple alignment around substitution position
+2.415	-0.528	2.943	9	precomputed	Secquences: Flanks:
MAPPING OF THE SUBSTITUTION SITE TO KNOWN PROTEIN 3D STRUCTURES					
Database	Initial number of structures			Number of structure	
PQS	709			0	

3.2.2 群体 SNP检测和等位基因频率估计

在群体分析中， 不同于单个样本的分析研究， 它不考虑单个个体基因型的可信度，而是在群体的层面上得到位点的基因型信息， 通常可以有较低的测序深度。群体分析时， 对于每一个位点， 通过贝叶斯算法估计每个可能基因型的概率、 为SNP的概率以及群体等位基因频率。由于较大的数据量，这样与单个样本的 SNP检测相比能够更有力地检测变异信息， 其结果更具有说服力， 并且能发现很多低频罕见变异。 这种方法成功应用于 50 个藏族人 (Yi et al. Science, 2010)和 200 个丹麦人 (Li et al. Nature Genetics, 2010)的外显子分析。

分析结果举例如下：

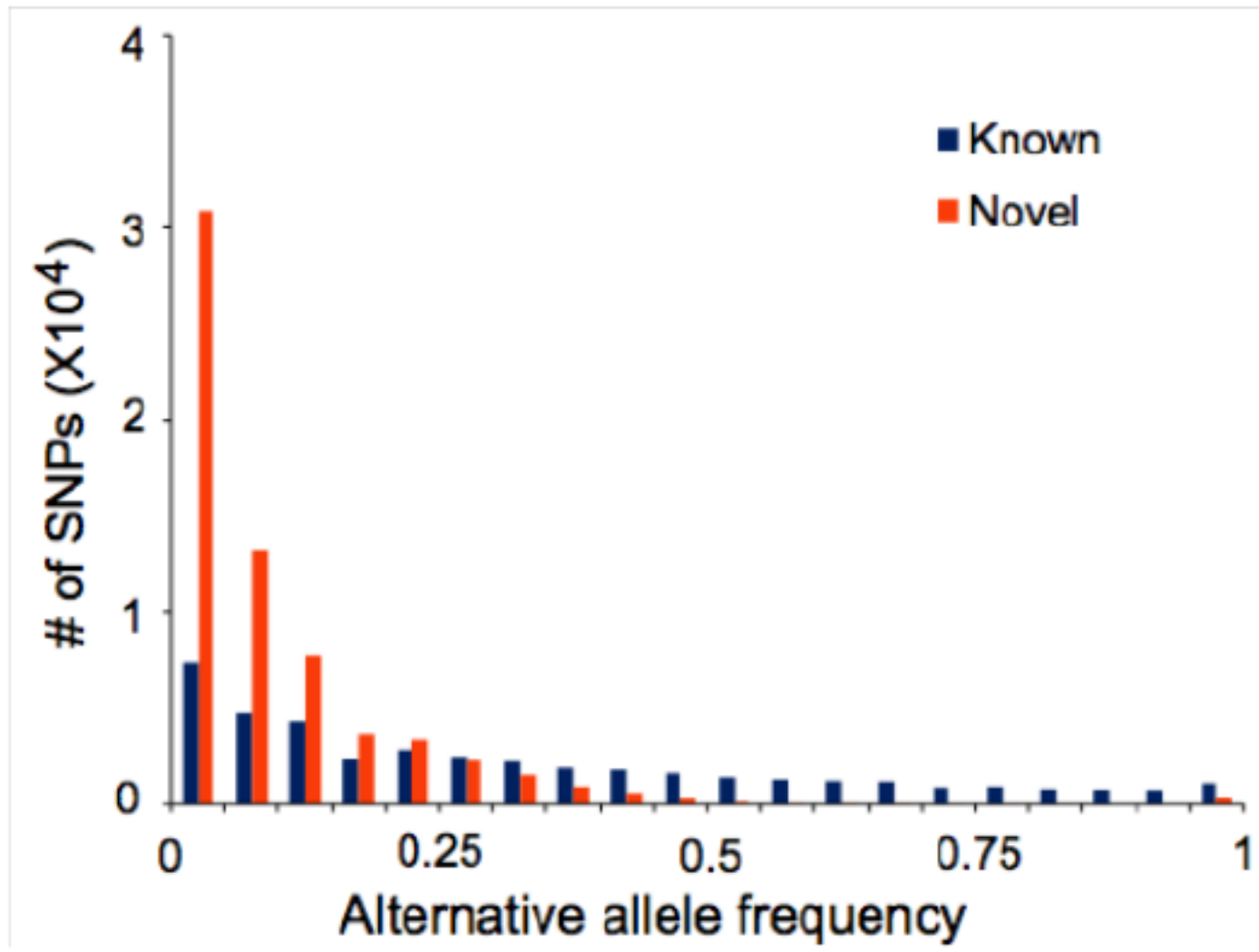


图 3.3 群体外显子分析的可变位点频谱（ SFS）

图 3.3 为群体分析中，外显子区域的可变等位基因的频谱。横坐标表示可变

等位基因频率 0~1，纵坐标表示对应频率的 SNP数目，图中红色表示新的 SNP数目，蓝色表示数据库 (dbSNP v129)中已知的 SNP数目，由图可以看出，在低频范围内，可以找出更多的新的 SNP，这些低频 SNP很可能与罕见疾病变异密切相关。

3.2.3 孟德尔遗传病分析

孟德尔遗传病通常指单基因遗传病， 简称单基因病（monogenic disease/single gene disorder)，是指单一基因突变引起的疾病， 符合孟德尔遗传方式， 所以也称为孟德尔式遗传病。 对变异结果进行注释后， 我们致力于寻找候选基因， 从而进一步确定致病基因。

筛选候选基因的方法如下：首先，将每个病例中已知的 SNPs进行过滤，采用的筛选数据库主要包括 dbSNP129 千人基因组数据库、 hapmap 外显子数据库，以及正常样本的数据。 其次， 假定候选变异都是非同义突变或者在剪接位点， 因此我们可以去除其它不改变蛋白产物的变异。 最后，我们得到在所有或大部分病例中存在的变异。这样就大大减少了候选变异的数量，缩小了寻找范围。

改为：

筛选候选基因的方法如下： 首先，过滤每个病例中已知的 SNPs，筛选用到的数据库包括 dbSNP129 千人基因组数据库， Hapmap 外显子数据库以及正常对照的 SNP 数据。其次，假定疾病是由非同义突变或者剪接位点突变导致，则去除其它不改变蛋白产物的变异。 最后，我们筛选出在所有或大部分病例中存在的变异，以减少候选变异的数量，从而缩小寻找范围。

举例如下：

表 3.6 不同范畴内的 SNPs 统计

Filter	Sample A (Whole/ Locus)	Sample B (Whole/ Locus)	Sample C (Whole/ Locus)	Sample D (Whole/ Locus)	Sample (A+B) (Whole/ Locus)	Sample (A+B+C) (Whole/ Locus)	Sample (A+B+C+D) (Whole/Locus)	2 affected (Whole/Locus)
NS/SS/Indel	5796/ 34	5649/ 40	5780/ 40	5842/ 37	3964/ 26	3099/26	2443/20	3736-3964/26 -30
Notin dbSNP 129	869/6	734/9	931/8	891/8	288/3	134/3	68/2	207-288/3-5

Not in dbSNP 129, nor in eight HapMap exomes	616/6	520/6	674/7	661/7	155/3	43/3	15/2	87-155/3-4
Not in dbSNP 129,eightHapMap exomes, nor in dbSNP1000 genomes	309/4	262/3	341/6	384/5	75/1	1-May	1-Jan	48-101/1
Predicted to be damaging	211/1	203/1	214/1	212/1	48/1	3/1	1/1	36-52/1

注： Whole/Locus ：Whole 表示整个外显子区域, Locus 表示特定的区域； NS/SS/Indel：表示非同义突变位点、剪接位点以及 Indel 的个数总和； 2 affected ：表示在两两不同组合病人中所检测到的相应信息的数量范围。

3.2.4 NGS-GWAS分析

基于芯片的 GWAS分析不能检测出稀有突变（即次等位基因频率 MAF 小于 0.05 的突变），外显子测序技术能够获得 MAF≥0.02 的等位基因频谱 (200 Danish exome, Li et al. Nature Genetics,2010)，这些有助于我们进行基于新一代测序技术的 GWAS分析。

3.2.5 正向选择信号的检测

通常更多的研究指向正向选择的基因， 我们可通过大量的数据集对每一个基因进行检测， 看其固定替换的比例是否显著偏移全基因组范围的期望， 通常采用 HKA test (Hudson-Kreitman-Aguad 检验)方法进行检验。 最近一项研究表明这种检验方法在检测正向选择上具有很大的效力 (Zhai et al. MolBiolEvol, 2009)。

采用之前的研究结果进行举例说明，显示结果如下：

表 3.7 HKA 检验

Gene Symbol	Description	F	P	F/P	Score
KIR3DP1	killer-cell Ig-like receptor	82	10	8.20	>7
LILRA1	leukocyte immunoglobulin-like receptor,	60	7	8.57	7
TPTE	transmembrane phosphatase with tensin homology	86	16	5.38	7
KIR2DL1	killer cell immunoglobulin-like receptor, two	40	3	13.33	6.05
VPS13D	vacuolar protein sorting 13D isoform 1	39	4	9.75	5.19

FLG	filaggrin	99	28	3.54	5.03
CES2	carboxylesterase 2 isoform 1	22	0	∞	4.95
TPRX1	tetra-peptide repeat homeobox	22	0	∞	4.95
HMCN1	hemicentin 1	62	15	4.13	4.12
TRPM2	transient receptor potential cation channel,	32	4	8.00	3.92
KIR2DL3	killer cell immunoglobulin-like receptor, two	34	5	6.80	3.76
KIAA1199	KIAA1199	21	1	21.00	3.75
SORBS2	sorbin and SH3 domain containing 2 isoform 2	24	2	12.00	3.62
TTC26	tetratricopeptide repeat domain 26 isoform 1	16	0	∞	3.60
SULT1C3	sulfotransferase family, cytosolic, 1C, member	33	5	6.60	3.59
HERC2	hect domain and RLD 2	43	9	4.78	3.50
SGTA	small glutamine-rich tetratricopeptide	15	0	∞	3.37
DYNC1H1	cytoplasmic dynein 1 heavy chain 1	47	11	4.27	3.37
CBWD2	COBW domain-containing protein 2	19	1	19.00	3.33
CSHL1	chorionic somatomammotropin hormone-like 1	22	2	11.00	3.24

注：

P：观察到的多态替换数； F：观察到的固定替换数； F/P：固定替换和多态替换的比值； Score：HKA 检验的得分。

4 数据分析方法说明

4.1 信息分析软件及常用参数介绍

1. SOAPaligner(soap2.21)用于将 reads与参考序列进行比对

参数设置如下： -a -b -D -o -u -p -2 -m -x -s 40 -l 35 -v 3

-a 查询文件，包含 single-end 比对的所有 reads 文件或者包含 pair-end 比对的其中一端的所有 reads 的文件

-b 查询文件，包含 pair-end 比对的另一端的 reads

-D 参考序列索引的前缀 [*.index]

-o 比对结果的输出文件

-u 包含没有比对上的 reads 输出文件

-p 使用的线程数

-2 包含 pair-end 比对中只有一端比对上的所有 reads 的文件

-m pair-end 比对最小插入片段长度

-x pair-end 比对最大插入片段长度

-s 最小的比对长度，我们设置的参数一般为 40bp

-l 对于 3'端具有较高的错误率而无法比对整个长度的长 reads，则先比对 5'端设置的长度序列作为种子序列，默认值为 256，表示使用 reads 的全长。？

-v 一条 reads 中允许的最大错配数

2. SOAPsnp 主要用于一致序列的组装

参数设置如下： -i -d -o -r 0.0005 -e 0.001 -u -L 150 -T -s -2

-i 将排序后的 SOAP比对结果作为输入文件

-d FASTA格式的 DNA 参考序列

-o 输出文件 (CNS文件)

-r 新的纯合 SNP的先验概率，默认值为 0.0005

-e 新的杂合 SNP的先验概率，默认值为 0.001

-u 秩和检验，检验可能杂合子的两个等位基因是否具有相同的测序质量

-L 最大 read 长度

-T 进行一致序列组装的目标区域

-s 已知 SNP 的信息文件

-2 通过已知的 SNP信息对 SNP进行修正

关于这两个软件的详细信息，请登录网站 http://soap.genomics.org.cn/

4.2 参考数据库

1. dbSNP数据库 ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606
2. Human reference genome(人类参考基因组)：UCSC(NCBI build36.3)

   http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/

   注：我们分析中所用的染色体坐标参照 UCSC Santa Cruz hg18, build 36.3
3. Target regions(目标区域)：使用的外显子芯片探针所覆盖到的区域

   http://www.nimblegen.com/seqcap/

   https://earray.chem.agilent.com/earray/
4. CCDS数据库

   ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_human/
5. RefSeq gene数据库
6. Ensembl 数据库

   http://www.ensembl.org/

4.3 数据文件格式

1. ***.fq[12].gz** —**fastq** 文件

   @A201GMABXX:5:1:14057:2058#GATCAG/1

   GCTATCCAGTGAGTCCTGCAAGACTTCAGGCTCTACTACCTCCAGCAG

   +

   Fefffaffecfffffffefffceefffcddffeecfcaddddd

   格式说明：

   每一条 reads 信息由四行组成，第一行以 '@'开头，其后接着序列的标志信息；第二行为序列的碱基组成；第三行以 '+'开头，其后可接与第一行相同的序列标志信息（可选）；第四行为第二行序列碱基的对应质量值，为一一对应关系，以 ASCII码表示。

2. ***.soap.gz—SOAP alignment of HiSeq 2000reads (**含有比对上参考序列的所有 **reads** 信息 **)**

   234
   GCTATCCAGTGAGTCCTGCAAGACTTCAGGCTCTACTACCTCCAGCAG
   fefffaffecfffffffefffceefffcddffeecfcaddddd
   1 a 48 + chr1 146653692 1 T->0G2 90M 0T89

   格式说明 **(**共 **13** 列）：

   1. Read 的 ID 号
   2. Read 序列的碱基组成。当第 7 列为 - 时（即比对到负链），此序列为原序列的反向互补序列。
   3. Read 序列的质量值，和第二列的序列成一一对应关系。
      计算方法为：质量值 = 相应的 ASCII 值 - 64 ，质量值范围一般为 0~40。
   4. best hit 数。没有 hit 的 reads 被忽略掉。
   5. Read 来源于哪个文件 (a/b)，对于 pair-end，包含 -a –b 两个参数， 即含有两个文件，对于 single-end，此列仅为 "a"。
   6. Read 长度。
   7. 比对参考序列的正负链。 +为正链， -为负链。
   8. 染色体 ID 号
   9. Read 的起始碱基在参考序列上的坐标
   10. Read 的碱基错配数
   11. Read 的错配信息

       例：T->0G2 T 为参考序列上的碱基类型， G 为 reads 上的碱基类型， 0 为其在 reads 上的位置， 2 为对应质量值。

   12. 匹配上的碱基数
   13. reads 的错配情况

例： 6T1A64 T 和 A 为错配的碱基，即在参考序列上对应位置是 T 和 A，但测得的 reads 上（第七和第九个位置）和参考序列不一致。

详细信息请登录： http://soap.genomics.org.cn/

3. ***.cns.gz**—**CNS**文件，由 **SOAPsnp**软件生成，包含识别出的外显子区域中一致序列基因型。

chrY 140161 G G 1 G 0 0 0 T 0 0 0 0 1.00000 255.000 0

格式说明：

1. 染色体 ID 号
2. 染色体上的坐标号
3. 参考序列上的基因型 (hg18, Mar. 2006)
4. 样本的一致序列二倍体基因型, .这里的基因型都是与参考序列的正链相关。
5. 一致基因型的质量得分
6. 最佳碱基，即根据贝叶斯先验概率，样本在此位置最可能的等位基因型。
7. 最佳碱基的质量得分
8. 唯一匹配上的最佳碱基数
9. 所有匹配上的最佳碱基数
10. 次佳碱基，即根据贝叶斯先验概率，样本在此位置次可能的等位基因型。
11. 次佳碱基的质量得分
12. 唯一匹配上的次佳碱基数
13. 所有匹配上的次佳碱基数
14. 此位点的测序深度
15. 秩和检验的 P 值
16. 附近区域的平均拷贝数
17. 此位点是否为 dbSNP

4. ***.snp**—**SNP**文件，包含样本中所有可能的 **SNP**位点，即一致序列基因型与参考序列基因型不同的位点。

chrY 2782506 A G 1 A 0 0 0 T 0 0 0 0 1.00000 255.000 1 3782506

格式说明：

在 CNS文件中增加一列，前 17 列格式说明与 CNS文件相同。第 18 列指这个 SNP位点与其最相邻的 SNP 位点的距离，即相隔碱基数,

5. ***.snp.filter** —**SNP**文件，在 ***.snp** 基础上按一定标准过滤之后所得到的最终 **SNP** 集合。

格式说明：

与*.snp 说明一致。此文件中产生的 SNP均为高置信度的 SNP。

过滤标准：

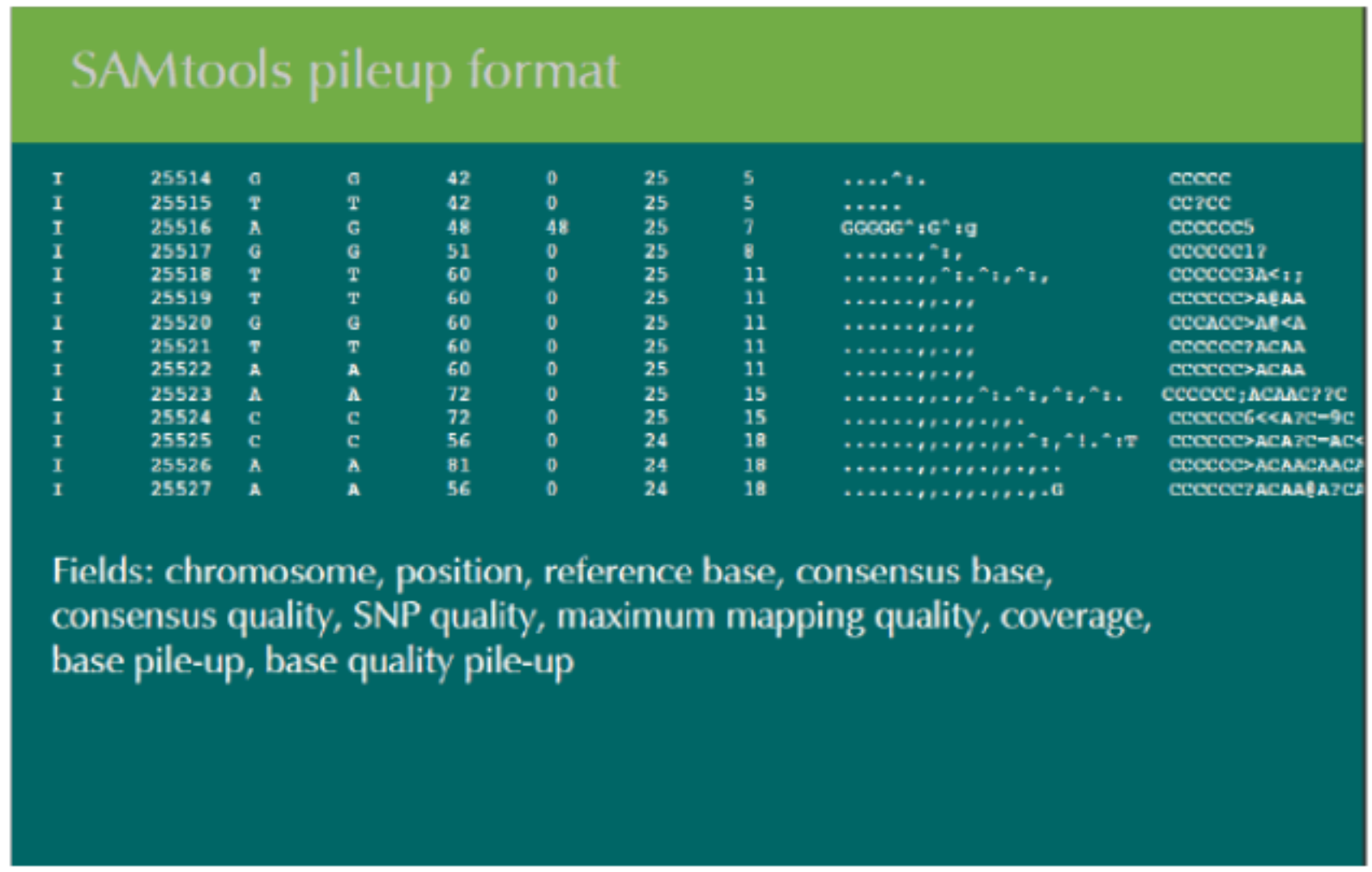
1. 位点质量值不低于 20，即过滤出 *.snp 文件中第 5 列的值大于 19 的所有位点
2. 位点的测序总深度不低于 4X，对于杂合位点，第一碱基的最佳碱基数和第二碱基的最佳碱基数分别大于 4X*0.5

6. ***.snp.filter.gff**—**gff** 注释文件，对结果 **SNPs**的详细注释

chr8 SOAPsnp SNP 1804774 1804774 99 + 2
ID=rs7003969;status=dbSNP;ref=G;alleles=A/G;support=7/5;name=ARHGEF10;geneID=9639;mutType=Het-one;transcriptID=CCDS34794.1;mRNAtoChr='+';exonNum=6;mRNA_pos=630;codonNum=210;codonChange='GAG=>GAA';residueChange='Glu=>Glu';function=synonymous-coding;

格式说明：

1. 染色体 ID 号
2. 名称，这里均为 SOAPsnp
3. 位点特征类型，这里均为 SNP
4. SNP的起始位置
5. SNP的终止位置
6. SNP的质量值
7. SNP在染色体上的链， SOAPsnp的结果通常为 ＋｜一致序列通常由参考序列正链表示
8. SNP的密码子相位， 用 0, 1, 2 来表示密码子的突变位置， 即密码子的第 1, 2, 3 个碱基， "."则表示其他的功能 SNP
9. SNP注释的详细信息，各字段详细说明如下：
   i. ID：唯一标志的 ID 号，如果此 SNP为新的 SNP，则使用 "snp+number" 进行标志；如果为 dbSNP，则用 dbSNP数据库中的 ID 号进行标志。
   ii. Ref： SNP位点的参考等位基因型，通常采用参考序列正链的等位基因。
   iii. Alleles：样本的 SNP位点的二倍体基因型， 如果位点为杂合， 则以 "best base/second best base"格式表示，如果位点为纯合，则以 "best base/best base 格式表示，其中 best base 表示最佳匹配碱基类型， second best base 表示次佳匹配碱基类型。
   iv. Support：唯一匹配的碱基数， 对于杂合位点， 通常以 "number of best base/number of second best base" 格式表示， 类似的， 对于纯合位点， 仅给出最佳匹配碱基数。
   v. Name： SNP所在的基因名
   vi. GeneID： SNP所在基因的基因 ID 号
   vii. MutType：变异类型 （Hom/ Het，即纯合或杂合），当两个等位基因都和参考序列不相同时，则变异类型定义为 Het-two
   viii. TranscriptID： SNP所在的转录本 ID，一个 SNP唯一对应一个转录本
   ix. mRNAtoChr：转录本位于参考序列的哪条链（ +/-）
   x. ExonNum：变异位于转录本的第几个外显子， 外显子的顺序由在参考序列上的位置坐标而定，从低到高排序。对于剪接位点 SNP和内含子 SNP，此字段命名为 "adj_exonNum"，如果 SNP在剪接位点，则 adj_exonNum 给出最近的外显子号，如果 SNP在内含子中，则 adj_exonNum 给出 SNP位点邻近的两个外显子号
   xi. mRNA_pos： SNP在转录本上的位置，从 5′端到 3 端计算
   xii. CodonNum：变异的密码子的排序位置
   xiii. CodonChange：密码子的变化情况
   xiv. ResidueChange：氨基酸的变化情况
   xv. Function：变异的功能分类， 非同义突变含 missense, nonsense ,readthrough 突变。


SAMtools pileup format
I 25514 G G 42 0 25 5 ....^:. CCCCC
I 25515 T T 42 0 25 5 ..... CC?CC
I 25516 A G 48 48 25 7 GGGGG^:G^:g CCCCCC5
I 25517 G G 51 0 25 8 ....^:, CCCCCC1?
I 25518 T T 60 0 25 11 ....,,^:.^:,^:, CCCCCC3A<:;
I 25519 T T 60 0 25 11 .....,,.,, CCCCCC>A@AA
I 25520 G G 60 0 25 11 .....,,.,, CCCACC>A@<A
I 25521 T T 60 0 25 11 .....,,.,, CCCCCC?ACAA
I 25522 A A 60 0 25 11 .....,,.,, CCCCCC>ACAA
I 25523 A A 72 0 25 15 .....,,.,,^:.^:,^:,^:. CCCCCC;ACAAC??C
I 25524 C C 72 0 25 15 .....,,.,,,,. CCCCCC6<<A?C-9C
I 25525 C C 56 0 24 18 .....,,.,,,,.^:,^!.^!T CCCCCC>ACA?C-AC<
I 25526 A A 81 0 24 18 .....,,.,,,,,,.. CCCCCC>ACAACAAC
I 25527 A A 56 0 24 18 .....,,.,,,,,.G CCCCCC?ACAA@A?C
Fields: chromosome, position, reference base, consensus base,
consensus quality, SNP quality, maximum mapping quality, coverage,
base pile-up, base quality pile-up