**高通量测序常用名词汇总**

**技术支持**

Q20值是指的测序过程碱基识别（Base Calling）过程中,对所识别的碱基给出的错误概率.  
如果质量值是Q20,则错误识别的概率是1%,即错误率1%,或者正确率是99%；  
如果质量值是Q30,则错误识别的概率是0.1%,即错误率0.1%,或者正确率是99.9%；  
如果质量值是Q40,则错误识别的概率是0.01%,即错误率0.01%,或者正确率是99.99%；  
你发现规律没有,Q“N”0的质量值,就是正确率有N个9的百分比,这样就非常容易记忆了.

基因高通量测序中，每测一个碱基会给出一个相应的质量值，这个质量值是衡量测序准确度的。碱基的质量值13，错误率为5%，20的错误率为1%，30的错误率为0.1%。行业中Q20与Q30则表示质量值≧20或30的碱基所占百分比。例如一共测了1G的数据量，其中有0.9G的碱基质量值大于或等于20，那么Q20则为90%。Q20值是指的测序过程碱基识别（Base Calling）过程中，对所识别的碱基给出的错误概率。质量值是Q20，则错误识别的概率是1%，即错误率1%，或者正确率是99%；  
        质量值是Q30，则错误识别的概率是0.1%，即错误率0.1%，或者正确率是99.9%；  
        质量值是Q40，则错误识别的概率是0.01%，即错误率0.01%，或者正确率是99.99%；

**一代测序技术：**即传统的Sanger测序法，Sanger法是根据核苷酸在待定序列模板上的引物点开始，随机在某一个特定的碱基处终止，并且在每个碱基后面进行荧光标记，产生以A、T、C、G结束的四组不同长度的一系列核苷酸，每一次序列测定由一套四个单独的反应构成，每个反应含有所有四种脱氧核苷酸三磷酸(dNTP)，并混入限量的一种不同的双脱氧核苷三磷酸(ddNTP)。由于ddNTP缺乏延伸所需要的3-OH 基团，使延长的寡聚核苷酸选择性地在G、A、T或C处终止，使反应得到一组长几百至几千碱基的链终止产物。它们具有共同的起始点，但终止在不同的的核苷酸上，可通过高分辨率变性凝胶电泳分离大小不同的片段，通过检测得到DNA碱基序列。

**二代测序技术：**next generation sequencing（NGS）又称为高通量测序技术，与传统测序相比，二代测序技术可以一次对几十万到几百万条核酸分子同时进行序列测定，从而使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序（Deep sequencing）。NGS主要的平台有Roche（454 & 454+），Illumina（HiSeq 2000/2500、GA IIx、MiSeq），ABI SOLiD等。

**基因**：Gene，是遗传的物质基础，是DNA或RNA分子上具有遗传信息的特定核苷酸序列。基因通过复制把遗传信息传递给下一代，使后代出现与亲代相似的性状。

**DNA**：Deoxyribonucleic acid，脱氧核糖核酸，一个脱氧核苷酸分子由三部分组成：含氮碱基、脱氧核糖、磷酸。脱氧核糖核酸通过3',5'-磷酸二酯键按一定的顺序彼此相连构成长链，即DNA链，DNA链上特定的核苷酸序列包含有生物的遗传信息，是绝大部分生物遗传信息的载体。

**RNA**：Ribonucleic Acid，，核糖核酸，一个核糖核苷酸分子由碱基，核糖和磷酸构成。核糖核苷酸经磷酯键缩合而成长链状分子称之为RNA链。RNA是存在于生物细胞以及部分病毒、类病毒中的遗传信息载体。不同种类的RNA链长不同，行使各式各样的生物功能，如参与蛋白质生物合成的RNA有信使RNA、转移RNA和核糖体RNA等。

**16S rDNA：**"S"是沉降系数，是反映生物大分子在离心场中向下沉降速度的一个指标，值越高，说明分子越大。rDNA（ribosome DNA）指的是原核生物基因组中编码核糖体RNA（rRNA）分子对应的DNA序列，16S rDNA是原核生物编码核糖体小亚基16S rRNA的基因。细菌rRNA（核糖体RNA）按沉降系数分为3种，分别为5S、16S和23S rRNA。16S rDNA是细菌染色体上编码16S rRNA相对应的DNA序列，存在于所有细菌染色体基因中。16S rRNA 普遍存在于原核生物中。16S rRNA 分子，其大小约1540bp，既含有高度保守的序列区域，又有中度保守和高度变化的序列区域，其可变区序列因细菌不同而异，恒定区序列基本保守，所以可利用恒定区序列设计引物，将16S rDNA片段扩增出来，通过高通量测序利用可变区序列的差异来对不同菌属、菌种的细菌进行分类鉴定。

**cDNA**：complementary DNA，互补脱氧核糖核酸，与RNA链互补的单链DNA，以RNA为模板，在反转录酶的作用下所合成的DNA。

**Small RNA**：生物体内一类高度保守的重要的功能分子，其大小在18-30nt，包括microRNA、siRNA、snRNA、snoRNA和piRNA（piwi-interacting RNA）等，它的主要功能是诱导基因沉默，调控细胞生长、发育、基因转录和翻译等生物学过程。以miRNA为例介绍它们的功能：miRNA与RNA诱导沉默复合体（RNA induced silencing complex, RISC）结合，并将此复合体与其互补的mRNA序列结合，根据靶序列与miRNA的互补程度，从而导致靶序列降解或干扰靶序列蛋白质的翻译过程。

**SD 区域：**Segment duplication，串联重复是由序列相近的一些 DNA 片段串联组成。串联重复在人类基因多样性的灵长类基因中发挥重要作用。

**Genotype and phenotype：**基因型与表型，基因型是指某一生物个体全部基因组合的总称；表型，又称性状，是基因型和环境共同作用的结果。

**基因组**：Genome，单倍体细胞核、细胞器（线粒体、叶绿体）或病毒粒子所含的全部DNA分子或RNA分子。

**全基因组*de novo*测序**：又称从头测序，它不依赖于任何现有的序列资料，而直接对某个物种的基因组进行测序，然后利用生物信息学分析手段对序列进行拼接、组装，从而获得该物种的基因组序列图谱。

**全基因组重测序**：对已有参考序列（Reference Sequence）物种的不同个体进行基因组测序，并以此为基础进行个体或群体水平的遗传差异性分析。全基因组重测序能够发现大量的单核苷酸多态性位点（SNP）、拷贝数变异（Copy Number Variation，CNV）、插入缺失（InDel，Insertion/Deletion）、结构变异（Structure Variation，SV）等变异类型，以准确快速的方法将单个参考基因组信息上升为群体遗传特征。

**转录组**：Transcriptome，是指特定生长阶段某组织或细胞内所有转录产物的集合；狭义上指所有mRNA的集合。

**转录组测序**：对某组织在某一功能状态下所能转录出来的所有RNA进行测序，获得特定状态下的该物种的几乎所有转录本序列信息。通常转录组测序是指对mRNA进行测序获得相关序列的过程。其根据所研究物种是否有参考基因组序列分为转录组*de novo*测序（无参考基因组序列）和转录组重测序（有参考基因组序列）。

**外显子组**：Exome，人类基因组全部外显子区域的集合称为外显子组，是基因中重要的编码蛋白的部分，并涵盖了与个体表型相关的大部分的功能性变异。

**外显子组测序**：是指利用序列捕获技术将全基因组外显子区域DNA捕捉并富集后进行高通量测序的基因组分析方法。外显子测序相对于基因组重测序成本较低，对研究已知基因的SNP、InDel 等具有较大的优势。

**目标区域测序：**应用相关试剂盒对基因组上感兴趣的目标区域进行捕获富集后进行大规模测序，一般需要根据目标区域专门定制捕获芯片。

**宏基因组：**Metagenome，指特定生活环境中全部微小生物遗传物质的总和。它包含了可培养的和未可培养的微生物的基因。目前主要指环境样品中的细菌和真菌的基因组总和。

**宏基因组16S rRNA测序**：可以对特定环境下的细菌和古细菌群体的微生物种类和风度进行有效的鉴定。对不同地点、不同条件下的多个样本16S rRNA的PCR产物平行测序，可以比较不同样本间的微生物组成及成分差异，进而阐明物种丰度、种群结果等生态学信息。

**表观遗传学**：Epigenetics，是指在基因组DNA序列没有改变的情况下，基因的表达调控和性状发生了可遗传的变化。表观遗传的现象很多，已知的有DNA甲基化（DNA methylation），基因组印记（genomic impriting），母体效应（maternal effects），基因沉默（gene silencing），核仁显性，休眠转座子激活和RNA编辑（RNA editing）等。

**全基因组甲基化测序**：DNA 甲基化是指在 DNA 甲基化转移酶的作用下，在基因组 CpG 二核苷酸的胞嘧啶5'碳位共价键结合一个甲基基团。DNA 甲基化已经成为表观遗传学和表观基因组学的重要研究内容。甲基化是基因表达的主要调控方式之一，研究染色体DNA甲基化情况是了解基因调控的重要手段。对已经有参考基因组的物种的基因组DNA用标准亚硫酸氢盐（Bisulfite）处理后，未甲基化的胞嘧啶C会脱氨基形成尿嘧啶U，经PCR扩增，U替换为胸腺嘧啶T，而发生甲基化的胞嘧啶C保持不变。将处理组与参考基因组序列进行比对，可发现甲基化位点并对甲基化情况进行定量分析的方法叫做全基因组甲基化测序。

**ChIp-Seq**：Chromatin Immunoprecipitation sequencing，即染色质免疫共沉淀-测序技术，即通过染色质免疫共沉淀技术特异性地富集目的蛋白结合的DNA片段。对富集得到的DNA片段进行纯化与文库构建，然后进行高通量测序，从而得到全基因组范围内可以与目的蛋白相互作用的DNA片段的方法叫做ChIP-Seq。

**数字表达谱**：Digital Gene Expression Profile，利用新一代高通量测序技术和高性能计算分析技术，能够全面、经济、快速地检测某一物种特定组织在特定状态下的基因表达情况，即运用特定的酶对mRNA距polyA tail 21-25nt的位置进行酶切，所获得的带polyA尾的序列(Tag)通过高通量测序，该tag被测得的次数即是对应基因的表达值。数字基因表达谱已被广泛应用于基础科学研究、医学研究和药物研发等领域。特点是经济，但获得的数据量有限。若想获得转录本的更多信息的话，一般都采用转录组测序的方法来测序。

**SBS**：sequencing by synthesis，边合成边测序反应，是指在DNA聚合酶的作用下延伸碱基所进行的测序。

**Run**：指高通量测序平台单次上机测序反应。

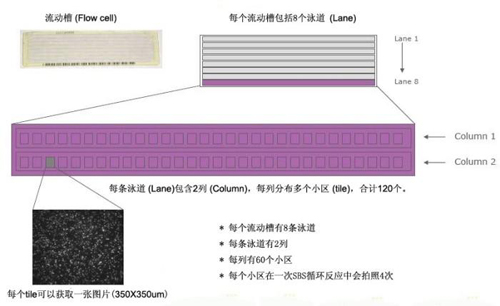


图1. Flow Cell结构示意图

**Lane**：也叫channel，单泳道，每条泳道包含2列（column），每列分布有多个小区（tile），如图1。不同的测序平台Flow Cell中所含的Lane不一样，如HiSeq 2000是2个flow cell，每个flow cell中含有8个lane；HiSeq 2500是包含2个mini flow cell（快速运行模式）和2个high output flow cell，两个模式不能同时运行，其中每个mini flow cell包含2个lane，每个high output flow cell中包含8个lane；Miseq系统的flow cell仅含有1个lane。

**Tile：**小区，每条Lane中有2列tile，合计120个小区。每个小区上分布数目繁多的簇结合位点，如图1。

**Cluster**：簇，在Illumina测序平台中会采用桥式PCR方式生产DNA簇，每个DNA簇才能产生亮度达到CCD可以分辨的荧光点。

**Index：**标签，在Illumina平台的多重测序（Multiplexed Sequencing）过程中会使用Index来区分样品，并在常规测序完成后，针对Index部分额外进行7个循环的测序，通过Index的识别，可以在1条Lane中区分12种不同的样品。

**Barcode**：与Index同义，多指在Roche GS FLX 454测序平台的16S PCR产物的测序过程中接头序列所包含的的用来区分不同样本的序列。

**PF%：**PF%是指符合测序质量标准的簇的百分比，与测序的通量相关联。

**Fasta：**一种序列存储格式。一个序列文件若以FASTA格式存储，则每一条序列的第一行以“>”开头，而跟随“>”的是序列的ID号（即唯一的标识符）及对该序列的描述信息；第二行开始是序列内容，序列短于61nt的，则一行排列完；序列长于61nt的，则每行存储61nt，最后剩下小于61nt的，在最后一行排列完；第二条序列另起一行，仍然由“>”和序列的ID号开始，以此类推。

**Fastq：**Fastq是Solexa测序技术中一种反映测序序列的碱基质量的文件格式。第一行以“@”符号开头，后面紧跟一个序列的描述信息；第二行是该序列的内容；第三行以“+”符号开头，后面可以是该序列的描述信息，也可省略；而第四行是第二行中的序列内容每个碱基所对应的测序质量值。

**Read：**高通量测序平台产生的序列标签就称为 reads。

**基因组组装：**进行基因组或转录组*de novo*测序时，物种基因组经构建不同的文库测序所得的片段需经过生物信息学手段对其进行整理拼接，并通过一定的标准（如N50）对后续组装结果进行质量评估等，最终获得高准确度的基因组序列的过程。

**基因组测序深度：**测序得到的总碱基数与待测基因组大小的比值。如测一个物种的全基因组的重测序，基因组大小约为5G，测序获得100G的数据量，则测序深度为20×。

**基因组覆盖率：**指测序获得的序列占整个基因组的比例。由于基因组中的高GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖有所的区域，这部分没有获得的区域就称为Gap。例如一个细菌基因组测序，覆盖率是98%，那么还有2%的序列区域是没有通过测序获得的。

**Contig：**在*de novo*测序中拼接软件基于 reads 之间的 overlap 区，拼接获得的中间没有gap的序列称为Contig（重叠群）。

**Scaffold：**基因组 de novo 测序，通过 reads 拼接获得 Contigs 后，往往还需要构建 454 Paired-end 库或Illumina Mate-pair 库，以获得一定大小片段（如 3Kb、8Kb、10Kb、20Kb）两端的序 列。基于这些序列，可以确定一些 Contig 之间的顺序关系，这些先后顺序已知的 Contigs 组成 Scaffold。

**Contig N50：**Reads拼接后会获得一些不同长度的Contigs。将所有的Contig长度相加，能获得一个Contig总长度。然后将所有的Contigs按照从长到短进行排序，如获得Contig 1，Contig 2，Contig 3……Contig 25。将Contig按照这个顺序依次相加，当相加的长度达到Contig总长度的一半时，最后一个加上的Contig长度即为Contig N50。举例：Contig 1+Contig 2+ Contig 3 +Contig 4=Contig总长度\*1/2时，Contig 4的长度即为Contig N50。Contig N50可以作为基因组拼接的结果好坏的一个判断标准。

**Scaffold N50：**Scaffold N50与Contig N50的定义类似。Contigs拼接组装获得一些不同长度的Scaffolds。将所有的Scaffold长度相加，能获得一个Scaffold总长度。然后将所有的Scaffolds按照从长到短进行排序，如获得Scaffold 1，Scaffold 2，Scaffold 3……Scaffold 25。将Scaffold按照这个顺序依次相加，当相加的长度达到Scaffold总长度的一半时，最后一个加上的Scaffold长度即为Scaffold N50。举例：Scaffold 1+Scaffold 2+ Scaffold 3 +Scaffold 4 +Scaffold 5=Scaffold总长度\*1/2时，Scaffold 5的长度即为Scaffold N50。Scaffold N50可以作为基因组拼接的结果好坏的一个判断标准。

**Isotig：**指在转录组*de novo*测序时，用454平台测序完成后组装出的结果，一个isotig可视为一个转录本。

**Isogroup：**指转录组*de novo*测序中，用454平台测序完成后组装出的结果获得的可聚类到同一个基因的转录本群。

**GC%：**GC含量，全基因组范围内或在特定基因组序列内的4种碱基中，鸟嘌呤和胞嘧啶所占的比率。

**SNP：**single nucleotide polymorphism，单核苷酸多态性，个体间基因组DNA序列同一位置单个核苷酸变异(替代、插入或缺失)所引起的多态性；不同物种个体基因组 DNA 序列同一位置上的单个核苷酸存在差别的现象。有这种差别的基因座、DNA序列等可作为基因组作图的标志。SNP 在CG序列上出现最为频繁，而且多是C转换为T ，原因是CG中的C 常为甲基化的，自发地脱氨后即成为胸腺嘧啶。一般而言，SNP 是指变异频率大于1 %的单核苷酸变异，主要用于高危群体的发现、疾病相关基因的鉴定、药物的设计和测试以及生物学的基础研究等。

**InDel：**Insertion/Deletion，插入/缺失，在基因组重测序进行mapping时，进行容Gap的比对并检测可信的Short InDel，如基因组上小片段>50bp的插入或缺失。在检测过程中，Gap的长度为1~5个碱基。

**CNV：**copy number variation，基因组拷贝数变异，是基因组变异的一种形式，通常使基因组中大片段的DNA形成非正常的拷贝数量。如人类正常染色体拷贝数是2，有些染色体区域拷贝数变成1或3，这样，该区域发生拷贝数缺失或增加，位于该区域内的基因表达量也会受到影响。如果把一条染色体分成A-B-C-D四个区域，则A-B-C-C-D/A-C-B-C-D/A-C-C-B-C-D/A-B-D 分别发生了C区域的扩增及缺失，扩增的位置可以是连续扩增如 A-B-C-C-D 也可以是在其他位置的扩增，如A-C-B-C-D。

**SV：**structure variation，基因组结构变异，染色体结构变异是指在染色体上发生了大片段的变异。主要包括染色体大片段的插入和缺失（引起 CNV 的变化），染色体内部的某块区域发生重复复制、翻转颠换、易位、两条染色体之间发生重组（inter-chromosome trans-location）等。

**基因表达差异：**是指某一物种或特定细胞在特定时期/功能状态下，多样本间不同基因在mRNA水平上表达量的差异，可通过RPKM/FPKM值来体现。

**RPKM：**Reads Per Kilobase per Million mapped reads [[Mortazavi etal., 2008]](http://www.clcbio.com/index.php?id=1330&manual=Bibliography.html)，是指每 1 百万个map 上 的reads中 map 到外显子的每1K 个碱基上的reads 个数。计算公式四RPKM=106C/NL/103，其中C为唯一比对到目的基因的reads数；N为唯一比对到参考基因的总reads数，L是目的基因编码区的碱基数。RPKM法可以消除基因长度、数据量之间的差异进行计算基因表达量。

**可变剪切：**alternative splicing大多数真核基因转录产生的mRNA前体是按一种方式剪接产生出一种mRNA，因而只产生一种蛋白质。但有些基因产生的mRNA前体可按不同的方式剪接，产生出两种或更多种mRNA，即可变剪接。

**基因融合：**Gene fusion，将基因组位置不同的两个或多个基因中的一部分或全部整合到一起，形成新的基因，称作融合基因或嵌合体基因，该基因有可能翻译出融合或嵌合体蛋白。

**基因家族分析：**通过进行BLASTN/ HMM比对等查找基因归属的基因家族并添加相关功能注释。

**基因组注释：**Genome annotation是利用生物信息学方法和工具,对基因组所有基因的生物学功能进行高通量注释,是当前功能基因组学研究的一个热点。基因组注释的研究内容包括基因识别和基因功能注释两个方面。基因识别的核心是确定全基因组序列中所有基因的确切位置。常见的基因组注释有GO注释、pathway分析。

**GO注释：**gene ontology是指对基因功能的注解。GO强调基因产物在细胞中的功能。GO不能反映此基因的表达情况，即是否在特定细胞中、特定组织中、特定发育阶段或与某种疾病相关，但GO支持其他的OBO(open biology ontologies)成员成立其他类型的本体论数据库（如发育本体学、蛋白组本体学、基因芯片本体学等）

**Pathway注释：**是指对功能基因参与的信号通路等进行分析注释。

**甲基化率：**是指在甲基化测序中，发生甲基化的胞嘧啶占所有胞嘧啶的比率。

**CpG岛：**CpG island 是指DNA上一个区域，此区域含有大量相联的胞嘧啶（C）、鸟嘌呤（G），以及使两者相连的磷酸酯键（p）。基因组中长度为300～3000 bp的富含CpG二核苷酸的一些区域，主要存在于基因的5’区域。启动子区中CpG岛的未甲基化状态是基因转录所必需的，而CpG序列中的C的甲基化可导致基因转录被抑制。

## 全基因组测序 - 研究经过

1986年， Renato Dulbecco是最早提出人类基因组定序的科学家之一。他认为如果能够知道所有人类基因的序列，对[癌症](http://www.baike.com/sowiki/%E7%99%8C%E7%97%87?prd=content_doc_search)的研究将会很有帮助。[美国能源部](http://www.baike.com/sowiki/%E7%BE%8E%E5%9B%BD%E8%83%BD%E6%BA%90%E9%83%A8?prd=content_doc_search)（DOE）与美国国家卫生研究院（NIH），分别在1986年与1987年加入[人类基因组计划](http://www.baike.com/sowiki/%E4%BA%BA%E7%B1%BB%E5%9F%BA%E5%9B%A0%E7%BB%84%E8%AE%A1%E5%88%92?prd=content_doc_search)。除了美国之外，日本在1981年就已经开始研究相关问题，但是并没有美国那样积极。到了1988年，詹姆士·华生（DNA双螺旋结构发现者之一）成为NIH的基因组部门主管。1990年开始国际合作。1996年，多个国家招开百慕达会议，以2005年完成定序为目标，分配了各国负责的工作，并且宣布研究结果将会即时公布，并完全免费。

1998年，克莱格·凡特的塞雷拉基因组公司成立，而且宣布将在2001年完成定序工作。随後国际团队也将完成工作的期限提前。2000年6月26日，塞雷拉公司的代表凡特，以及国际合作团队的代表弗朗西斯·柯林斯（Francis Collins），在美国总统柯林顿的陪同下发表演说，宣布人类基因组的概要已经完成。2001年2月，国际团队与塞雷拉公司，分别将研究成果发表于[《自然》](http://www.baike.com/sowiki/%E3%80%8A%E8%87%AA%E7%84%B6%E3%80%8B?prd=content_doc_search)与[《科学》](http://www.baike.com/sowiki/%E3%80%8A%E7%A7%91%E5%AD%A6%E3%80%8B?prd=content_doc_search)两份期刊。在基因组计划的研究过程中，塞雷拉基因组使用的是[霰弹枪定序法](http://www.baike.com/sowiki/%E9%9C%B0%E5%BC%B9%E6%9E%AA%E5%AE%9A%E5%BA%8F%E6%B3%95?prd=content_doc_search)（shotgun sequencing），这种方法较为迅速 ，但是仍需以传统定序来分析细节。目前，全基因组测序技术主要包括第二代测序技术（NGS）和第三代测序技术。第二代测序技术已经能够快速、低成本的进行全基因组测序，其设备供应商主要是Solexa （现被Illumina公司合并），454（罗氏公司）和SOLiD（AB公司）。第三代测序技术于2011年4月正式推广，其单分子实时（SMRT）测序技术完全不同与第二代测序，它的序列读长高达3000 bp（Pacific Biosciences 公司研发）。

## 全基因组测序 - 技术路线

提取基因组DNA，然后随机打断，电泳回收所需长度的DNA片段（0.2~5kb），加上接头,进行基因簇cluster制备或电子扩增E-PCR，最后利用Paired-End（Solexa）或者Mate-Pair（SOLiD）的方法对插入片段进行测序。然后对测得的序列组装成Contig，通过Paired-End的距离可进一步组装成Scaffold，进而可组装成染色体等。组装效果与[测序深度](http://www.baike.com/sowiki/%E6%B5%8B%E5%BA%8F%E6%B7%B1%E5%BA%A6?prd=content_doc_search)与覆盖度、测序质量等有关。目前常用的组装有：SOAPdenovo、Trimity、Abyss等。

## 全基因组测序 - 原理

**双末端（Paired-End）测序原理**

测序深度（Sequencing Depth）：测序得到的碱基总量（bp）与基因组大小（Genome）的比值，它是评价测序量的指标之一。测序深度与基因组覆盖度之间是一个正相关的关系，测序带来的错误率或假阳性结果会随着测序深度的提升而下降。重测序的个体，如果采用的是双末端或Mate-Pair方案，当测序深度在10~15X以上时，基因组覆盖度和测序错误率控制均得以保证。

**测序深度对基因组覆盖度和测序错误率的影响**

（HOM：纯合体 HET：杂合体）

[全基因组重测序](http://www.baike.com/sowiki/%E5%85%A8%E5%9F%BA%E5%9B%A0%E7%BB%84%E9%87%8D%E6%B5%8B%E5%BA%8F?prd=content_doc_search)的个体，通过序列比对，可以找到大量的[单核苷酸](http://www.baike.com/sowiki/%E5%8D%95%E6%A0%B8%E8%8B%B7%E9%85%B8?prd=content_doc_search)多态性（SNP），插入缺失（[InDel](http://www.baike.com/sowiki/InDel?prd=content_doc_search)，Insertion/Deletion）和结构变异（SV，Structure Variation）位点。SBC可以协助客户，通过生物信息手段，分析不同个体基因组间的结构差异， 同时完成SNP及基因组结构注释。

## 全基因组测序 - 分析流程

**1．数据量产出**

总碱基数量、Total Mapping Reads、Uniquely Mapping Reads统计，测序深度分析。

**2．一致性序列组装**

与参考基因组序列（Reference genome sequence）的比对分析，利用[贝叶斯统计模型](javascript:linkredwin('贝叶斯统计模型');)检测出每个碱基位点的最大可能性基因型，并组装出该个体基因组的[一致序列](http://www.baike.com/sowiki/%E4%B8%80%E8%87%B4%E5%BA%8F%E5%88%97?prd=content_doc_search)。

**3．SNP检测及在基因组中的分布**

提取全基因组中所有多态性位点，结合质量值、测序深度、重复性等因素作进一步的过滤筛选，最终得到可信度高的SNP数据集。并根据参考基因组信息对检测到的变异进行注释。

**4．InDel检测及在基因组的分布**

在进行mapping的过程中，进行容gap的比对并检测可信的short InDel。在检测过程中，gap的长度为1~5个[碱基](http://www.baike.com/sowiki/%E7%A2%B1%E5%9F%BA?prd=content_doc_search)。对于每个InDel的检测，至少需要3个Paired-End序列的支持。

**5．Structure Variation检测及在基因组中的分布**

SBC能够检测到的结构变异类型主要有：插入、缺失、复制、倒位、易位等。根据测序个体序列与参考基因组序列比对分析结果，检测全基因组水平的结构变异并对检测到的变异进行注释。