

高通量测序基础知识简介

陆桂

什么是高通量测序？

高通量测序技术（High-throughput sequencing, HTS）是对传统 Sanger 测序（称为一代测序技术）革命性的改变,一次对几十万到几百万条核酸分子进行序列测定,因此在有些文献中称其为下一代测序技术(next generation sequencing, NGS)足见其划时代的改变,同时高通量测序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)。

什么是 Sanger 法测序（一代测序）

Sanger 法测序利用一种 DNA 聚合酶来延伸结合在待定序列模板上的引物。直到掺入一种链终止核苷酸为止。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。终止点由反应中相应的双脱氧而定。每一种 dNTPs 和 ddNTPs 的相对浓度可以调整,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,凝胶处理后可用 X-光胶片放射自显影或非同位素标记进行检测。

什么是基因组重测序（Genome Re-sequencing）

全基因组重测序是对基因组序列已知的个体进行基因组测序,并在个体或群体水平上进行差异性分析的方法。随着基因组测序成本的不断降低,人类疾病的致病突变研究由外显子区域扩大到全基因组范围。通过构建不同长度的插入片段文库和短序列、双末端测序相结合的策略进行高通量测序,实现在全基因组水平上检测疾病关联的常见、低频、甚至是罕见的突变位点,以及结构变异等,具有重大的科研和产业价值。

什么是 de novo 测序

de novo 测序也称为从头测序:其不需要任何现有的序列资料就可以对某个物种进行测序,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图谱。获得一个物种的全基因组序列是加快对此物种了解的重要捷径。随着新一代测序技术的飞速发展,基因组测序所需的成本和时间较传统技术都大大降低,大规模基因组测序渐入佳境,基因组学研究也迎来新的发展契机和革命性突破。利用新一代高通量、高效率测序技术以及强大的生物信息分析能力,可以高效、低成本地测定并分析所有生物的基因组序列。

什么是外显子测序（whole exon sequencing）

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,但无法研究基因组结构变异如染色体断裂重组等。

什么是 mRNA 测序（RNA-seq）

转录组学（transcriptomics）是在基因组学后新兴的一门学科，即研究特定细胞在某一功能状态下所能转录出来的所有 RNA（包括 mRNA 和非编码 RNA）的类型与拷贝数。Illumina 提供的 mRNA 测序技术可在整个 mRNA 领域进行各种相关研究和新的发现。mRNA 测序不对引物或探针进行设计，可自由提供关于转录的客观和权威信息。研究人员仅需要一次试验即可快速生成完整的 poly-A 尾的 RNA 完整序列信息，并分析基因表达、cSNP、全新的转录、全新异构体、剪接位点、等位基因特异性表达和罕见转录等最全面的转录组信息。简单的样品制备和数据分析软件支持在所有物种中的 mRNA 测序研究。

什么是 small RNA 测序

Small RNA（micro RNAs、siRNAs 和 pi RNAs）是生命活动重要的调控因子，在基因表达调控、生物个体发育、代谢及疾病的发生等生理过程中起着重要的作用。Illumina 能够对细胞或者组织中的全部 Small RNA 进行深度测序及定量分析等研究。实验时首先将 18-30 nt 范围的 Small RNA 从总 RNA 中分离出来，两端分别加上特定接头后体外反转录做成 cDNA 再做进一步处理后，利用测序仪对 DNA 片段进行单向末端直接测序。通过 Illumina 对 Small RNA 大规模测序分析，可以从中获得物种全基因组水平的 miRNA 图谱，实现包括新 miRNA 分子的挖掘，其作用靶基因的预测和鉴定、样品间差异表达分析、miRNAs 聚类 and 表达谱分析等科学应用。

什么是 miRNA 测序

成熟的 microRNA（miRNA）是 17~24nt 的单链非编码 RNA 分子，通过与 mRNA 相互作用影响目标 mRNA 的稳定性及翻译，最终诱导基因沉默，调控着基因表达、细胞生长、发育等生物学过程。基于第二代测序技术的 microRNA 测序，可以一次性获得数百万条 microRNA 序列，能够快速鉴定出不同组织、不同发育阶段、不同疾病状态下已知和未知的 microRNA 及其表达差异，为研究 microRNA 对细胞进程的作用及其生物学影响提供了有力工具。

什么是 ChIP-seq

染色质免疫共沉淀技术（Chromatin Immunoprecipitation, ChIP）也称结合位点分析法，是研究体内蛋白质与 DNA 相互作用的有力工具，通常用于转录因子结合位点或组蛋白特异性修饰位点的研究。将 ChIP 与第二代测序技术相结合的 ChIP-Seq 技术，能够高效地在全基因组范围内检测与组蛋白、转录因子等互作的 DNA 区段。

ChIP-Seq 的原理是：首先通过染色质免疫共沉淀技术（ChIP）特异性地富集目的蛋白结合的 DNA 片段，并对其进行纯化与文库构建；然后对富集得到的 DNA 片段进行高通量测序。研究人员通过将获得的数百万条序列标签精确定位到基因组上，从而获得全基因组范围内与组蛋白、转录因子等互作的 DNA 区段信息。

什么是 CHIRP-Seq

CHIRP-Seq(Chromatin Isolation by RNA Purification)是一种检测与 RNA 绑定的 DNA 和蛋白的高通量测序方法。方法是通过设计生物素或链霉亲和素探针，把目标 RNA 拉下来以后，与其共同作用的 DNA 染色体片段就会附在磁珠上，最后把染色体片段做高通量测序，这样会得到该 RNA 能够结合到在基因组的哪些区域，但由于蛋白测序技术不够成熟，无法知道与该 RNA 结合的蛋白。

什么是 RIP-seq

RNA Immunoprecipitation 是研究细胞内 RNA 与蛋白结合情况的技术，是了解转录后调控网络动态过程的有力工具，能帮助我们发现 miRNA 的调节靶点。这种技术运用针对目标蛋白的抗体把相应的 RNA-蛋白复合物沉淀下来，然后经过分离纯化就可以对结合在复合物上的 RNA 进行测序分析。

RIP 可以看成是普遍使用的染色质免疫沉淀 ChIP 技术的类似应用，但由于研究对象是 RNA-蛋白复合物而不是 DNA-蛋白复合物，RIP 实验的优化条件与 ChIP 实验不太相同（如复合物不需要固定，RIP 反应体系中的试剂和抗体绝对不能含有 RNA 酶，抗体需经 RIP 实验验证等等）。RIP 技术下游结合 microarray 技术被称为 RIP-Chip，帮助我们更高通量地了解癌症以及其它疾病整体水平的 RNA 变化。

什么是 CLIP-seq

CLIP-seq, 又称为 HITS-CLIP , 即紫外交联免疫沉淀结合高通量测序 (crosslinking-immunoprecipitation and high-throughput sequencing), 是一项在全基因组水平揭示 RNA 分子与 RNA 结合蛋白相互作用的革命性技术。其主要原理是基于 RNA 分子与 RNA 结合蛋白在紫外照射下发生耦联，以 RNA 结合蛋白的特异性抗体将 RNA-蛋白质复合体沉淀之后，回收其中的 RNA 片段，经添加接头、RT-PCR 等步骤，对这些分子进行高通量测序，再经生物信息学的分析和处理、总结，挖掘出其特定规律，从而深入揭示 RNA 结合蛋白与 RNA 分子的调控作用及其对生命的意义。

什么是 metagenomic (宏基因组):

Metagenomics 研究的对象是整个微生物群落。相对于传统单个细菌研究来说，它具有众多优势，其中很重要的两点：(1)微生物通常是以群落方式共生于某一小生境中，它们的很多特性是基于整个群落环境及个体间的相互影响的，因此做 Metagenomics 研究比做单个个体的研究更能发现其特性；(2) Metagenomics 研究无需分离单个细菌，可以研究那些不能被实验室分离培养的微生物。

宏基因组是基因组学一个新兴的科学研究方向。宏基因组学（又称元基因组学，环境基因组学，生态基因组学等），是研究直接从环境样本中提取的基因组遗传物质的学科。传统的微生物研究依赖于实验室培养，元基因组的兴起填补了无法在传统实验室中培养的微生物研究的空白。过去几年中，DNA 测序技术的进步以及测序通量和分析方法的改进使得人们得以一窥这一未知的基因组科学领域。

什么是 SNP、SNV（单核苷酸位点变异）

单核苷酸多态性 **single nucleotide polymorphism**, **SNP** 或单核苷酸位点变异 **SNV**。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异的现象。有这种差异的基因座、**DNA** 序列等可作为基因组作图的标志。人基因组上平均约每 1000 个核苷酸即可能出现 1 个单核苷酸多态性的变化, 其中有些单核苷酸多态性可能与疾病有关, 但可能大多数与疾病无关。单核苷酸多态性是研究人类家族和动植物物种遗传变异的重要依据。在研究癌症基因组变异时, 相对于正常组织, 癌症中特异的单核苷酸变异是一种体细胞突变 (**somatic mutation**), 称做 **SNV**。

什么是 **INDEL** (基因组小片段插入)

基因组上小片段 (>50bp) 的插入或缺失, 形同 **SNP/SNV**。

什么是 **copy number variation (CNV)**: 基因组拷贝数变异

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。例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。

什么是 **structure variation (SV)**: 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的变异。主要包括染色体大片的插入和缺失 (引起 **CNV** 的变化), 染色体内部的某块区域发生翻转颠换, 两条染色体之间发生重组 (**inter-chromosome trans-location**) 等。一般 **SV** 的展示利用 **Circos** 软件。

什么是 **Segment duplication**

一般称为 **SD** 区域, 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。在人类染色体 Y 和 22 号染色体上, 有很大的 **SD** 序列。

什么是 **genotype and phenotype**

既基因型与表型; 一般指某些单核苷酸位点变异与表现形式间的关系。

什么是 Read?

高通量测序平台产生的序列标签就称为 reads。

什么是 soft-clipped reads

当基因组发生某一段的缺失，或转录组的剪接，在测序过程中，横跨缺失位点及剪接位点的 reads 回帖到基因组时，一条 reads 被切成两段，匹配到不同的区域，这样的 reads 叫做 soft-clipped reads，这些 reads 对于鉴定染色体结构变异及外源序列整合具有重要作用。

什么是 multi-hits reads

由于大部分测序得到的 reads 较短，一个 reads 能够匹配到基因组多个位置，无法区分其真实来源的位置。一些工具根据统计模型，如将这类 reads 分配给 reads 较多的区域。

什么是 Contig?

拼接软件基于 reads 之间的 overlap 区，拼接获得的序列称为 Contig（重叠群）。

什么是 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、6Kb、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。

什么是 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。

什么是 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。

什么是测序深度和覆盖度?

测序深度是指测序得到的总碱基数与待测基因组大小的比值。假设一个基因大小为 2M，测序深度为 10X，那么获得的总数据量为 20M。覆盖度是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，覆盖度是 98%，那么还有 2% 的序列区域是没有通过测序获得的。

什么是 RPKM、FPKM

RPKM, Reads Per Kilobase of exon model per Million mapped reads, is defined in this way [Mortazavi et al., 2008]:

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。

假如有 1 百万个 reads 映射到了人的基因组上，那么具体到每个外显子呢，有多少映射上了呢，而外显子的长度不一，那么每 1K 个碱基上又有多少 reads 映射上了呢，这大概就是这个 RPKM 的直观解释。

$$\text{RPKM (exon)} = 10^9 * \text{exon_tag_count} / (\text{total_tag_count} * \text{exon_size})$$
$$\text{RPKM (gene)} = 10^9 * \text{gene_tag_count} / (\text{total_tag_count} * \text{canonical_transcript_size})$$

Mortazavi et al. (2008) Nature Methods

如果对应特定基因的话，那么就是每 1000000 mapped 到该基因上的 reads 中每 kb 有多少是 mapped 到该基因上的 exon 的 read

Total exon reads: This is the number in the column with header Total exonreads in the row for the gene. This is the number of reads that have been mapped to a region in which an exon is annotated for the gene or across the boundaries of two exons or an intron and an exon for an annotated transcript of the gene. For eukaryotes, exons and their internal relationships are defined by annotations of type mRNA. 映射到外显子上总的 reads 个数。这个是映射到某个区域上的 reads 个数，这个区域或者是已知注释的基因或者跨两个外显子的边界或者是某个基因已经注释的转录本的内含子、外显子。对于真核生物来说，外显子和它们自己内部的关系由某类型的 mRNA 来注释。

Exon length: This is the number in the column with the header Exon length in the row for the gene, divided by 1000. This is calculated as the sum of the lengths of all exons annotated for the gene. Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region. 外显子的长度。计算时，计算所有某个基因已注释的所有外显子长度的总和。即使某个基因以多种注释的转录本呈现，这个外显子在求和时只被包含一次。即使部分重叠的外显子共享相同的区域，重叠的外显子以其总长来计算。

Mapped reads: The sum of all the numbers in the column with header Total gene reads. The Total gene reads for a gene is the total number of reads that after mapping have been mapped to the region of the gene. Thus this includes all the reads uniquely mapped to the region of the gene as well as those of the reads which match in more places (below the limit set in the dialog in figure 18.110) that have been allocated to this gene's region. A gene's region is that comprised of the flanking regions (if it was specified in figure 18.110), the exons, the introns and across exon-exon boundaries of all transcripts annotated for the gene. Thus, the sum of the total gene reads numbers is the number of mapped reads for the sample (you can find the number in the RNA-Seq report). map 的 reads 总和。映射到某个基因上的所有 reads 总数。因此这包含所有的唯一映射到这个区域上的 reads。

举例：比如对应到该基因的 read 有 1000 个，总 reads 个数有 100 万，而该基因的外显子总长为 5kb，那么它的 RPKM 为： $10^9 \times 1000(\text{reads 个数}) / 10^6(\text{总 reads 个数}) \times 5000(\text{外显子长度}) = 200$ 或者： $1000(\text{reads 个数}) / 1(\text{百万}) \times 5(K) = 200$ 这个值反映基因的表达水平。

FPKM(fragments per kilobase of exon per million fragments mapped). FPKM 与 RPKM 计算方法基本一致。不同点就是 FPKM 计算的是 fragments，而 RPKM 计算的是 reads。Fragment 比 read 的含义更广，因此 FPKM 包含的意义也更广，可以是 pair-end 的一个 fragment，也可以是一个 read。

什么是转录本重构

用测序的数据组装成转录本。有两种组装方式：1，de-novo 构建；2，有参考基因组重构。其中 de-novo 组装是指在不依赖参考基因组的情况下，将有 overlap 的 reads 连接成一个更长的序列，经过不断的延伸，拼成一个个的 contig 及 scaffold。常用工具包括 velvet, trans-ABYSS, Trinity 等。有参考基因组重构，是指先将 read 贴回到基因组上，然后在基因组通过 reads 覆盖度，junction 位点的信息等得到转录本，常用工具包括 scripture、cufflinks。

什么是 genefusion

将基因组位置不同的两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。

什么是表达谱

基因表达谱(gene expression profile)：指通过构建处于某一特定状态下的细胞或组织的非偏性 cDNA 文库,大规模 cDNA 测序,收集 cDNA 序列片段、定性、定量分析其 mRNA 群体组成,从而描绘该特定细胞或组织在特定状态下的基因表达种类和丰度信息,这样编制成的数据表就称为基因表达谱

什么是功能基因组学

功能基因组学(Functional genomics)又往往被称为后基因组学(Postgenomics)，它利用结构基因组所提供的信息和产物，发展和应用新的实验手段，通过在基因组或系统水平上全面分析基因的功能，使得生物学研究从对单一基因或蛋白质得研究转向多个基因或蛋白质同时进行系统的研究。这是在基因组静态的碱基序列弄清楚之后转入对基因组动态的生物学功能学研究。研究内容包括基因功能发现、基因表达分析及突变检测。基因的功能包括：生物学功能，如作为蛋白质激酶对特异蛋白质进行磷酸化修饰；细胞学功能，如参与细胞间和细胞内信号传递途径；发育上功能，如参与形态建成等。采用的手段包括经典的减法杂交，差示筛选，cDNA 代表差异分析以及 mRNA 差异显示等，但这些技术不能对基因进行全面系统的

分析,新的技术应运而生,包括基因表达的系统分析(serial analysis of gene expression,SAGE), cDNA 微阵列(cDNA microarray), DNA 芯片(DNA chip)和序列标志片段显示(sequence tagged fragmentsdisplay)。

什么是比较基因组学

比较基因组学(Comparative Genomics)是基于基因组图谱和测序基础上,对已知的基因和基因组结构进行比较,来了解基因的功能、表达机理和物种进化的学科。利用模式生物基因组与人类基因组之间编码顺序上和结构上的同源性,克隆人类疾病基因,揭示基因功能和疾病分子机制,阐明物种进化关系,及基因组的内在结构。

什么是表观遗传学

表观遗传学是研究基因的核苷酸序列不发生改变的情况下,基因表达了可遗传的变化的一门遗传学分支学科。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 imprinting),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。

什么是计算生物学

计算生物学是指开发和应用数据分析及理论的方法、数学建模、计算机仿真技术等。当前,生物学数据量和复杂性不断增长,每 14 个月基因研究产生的数据就会翻一番,单单依靠观察和实验已难以应付。因此,必须依靠大规模计算模拟技术,从海量信息中提取最有用的数据。

什么是基因组印记

基因组印记(又称遗传印记)是指基因根据亲代的不同而有不同的表达。印记基因的存在能导致细胞中两个等位基因的一个表达而另一个不表达。基因组印记是一正常过程,此现象在一些低等动物和植物中已发现多年。印记的基因只占人类基因组中的少数,可能不超过 5%,但在胎儿的生长和行为发育中起着至关重要的作用。基因组印记病主要表现为过度生长、生长迟缓、智力障碍、行为异常。目前在肿瘤的研究中认为印记缺失是引起肿瘤最常见的遗传学因素之一。

什么是基因组学

基因组学(英文 genomics),研究生物基因组和如何利用基因的一门学问。用于概括涉及基因作图、测序和整个基因组功能分析的遗传学分支。该学科提供基因组信息以及相关数据系统利用,试图解决生物,医学,和工业领域的重大问题。

什么是 DNA 甲基化

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。正常情况下，人类基因组“垃圾”序列的 CpG 二核苷酸相对稀少，并且总是处于甲基化状态，与之相反，人类基因组中大小为 100—1000 bp 左右且富含 CpG 二核苷酸的 CpG 岛则总是处于未甲基化状态，并且与 56% 的人类基因组编码基因相关。人类基因组序列草图分析结果表明，人类基因组 CpG 岛约为 28890 个，大部分染色体每 1 Mb 就有 5—15 个 CpG 岛，平均值为每 Mb 含 10.5 个 CpG 岛，CpG 岛的数目与基因密度有良好的对应关系[9]。由于 DNA 甲基化与人类发育和肿瘤疾病的密切关系，特别是 CpG 岛甲基化所致抑癌基因转录失活问题，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。

什么是基因组注释？

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。

什么是 Q30？

Q30 是指一个碱基的识别可靠性等于 99.9%，或者说出错可能性是 0.1%。Q20 则是指碱基识别的可靠性等于 99%。

Q30 数据量是指一批数据中，质量高于等于 Q30 的数据的量的总和。


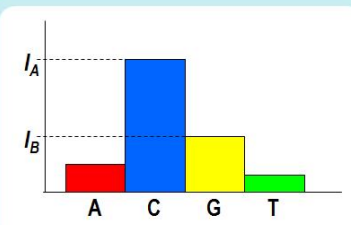
测序数据的 PF data/PF reads 是什么意思？

PF 是 pass filter 的意思。也就是质量合格的意思。Illumina 的测序仪会自动地对一个 read(序列)的质量可靠性进行打分。

对于前 25 个碱基中的是否有两个碱基的识别可靠性低于 0.6，是 PF 的判断标准。这句话翻译成比较容易理解的话：就是前 25 个碱基中，如果低质量的数据有 2 个或更多，则这条 read 被判定为不合格，PF 就不通过。反之，则质检通过。

Quality Filtering (PF)

- Filter clusters by signal purity
- Trade-off between lower error and higher yield
- Second-lowest CHASTITY at least 0.6 in first 25 cycles
 - Allows one “bad” cycle

$$C = \frac{I_A}{I_A + I_B}$$

CHASTITY Formula

PF 是国际公认的质检标准。

你们给的数据是什么质量的？

对于哺乳动物基因组重测序、外显子测序，我们保证数据质量是 Q30 的比例高于 80%。对于 mRNA 测序，smRNA 测序，我们保证对照 Lane 的数据质是 Q30 的比例高于 80%。

一般情况下：

哺乳动物基因组重测序、外显子测序，GC 比例在 40%左右，Q30 的比例是 80~95%

RNA-seq，GC 比例在 50%左右，Q30 的比例是~80%。如果 Poly(A)特别多的情况下，Q30 会更低一些

SmRNA-seq，因为有许多 read 读通之后，只剩下一串的 A，质量会更低，我们的实验结果%Q30 在 70~75%

测序中的 Duplication 是什么，如何避免，一般会有多少 Duplication？

所谓 Duplication 是指起始与终止位置完全一致的片段。

引起 Duplication 的主要原因是因为在测序中有 PCR 过程，来源于同一个 DNA 片段 PCR 的产物被重复测序，就会是 Duplication。次要原因是正巧两个片段的头和尾的位置完全一致。

一般通过控制 PCR 的循环数来控制 Duplication。我们一般控制 PCR 的循环次数在 10~12 个循环。

在药明康德外显子测序中，如果用 illumina 的捕获试剂盒 Duplication 的比例约为 10%，如果用 Nimblegen 的捕获试剂盒 Duplication 的比例波动较大，在 5~50%范围，平均为 30%。

在 RNA-seq 中，Duplication 的比例约为 40%。RNA-seq 中，因为高丰度的 mRNA 集中在几个基因上，集中度很高，所以 Duplication 的比例也就高。

测序的插入片段一般是多长？

测序的插入片段一般是 100bp 到 600bp。

因为 Hiseq 测序过程中有一个桥式 PCR 的过程。如果插入片段过长，测桥式 PCR 产生的 Cluster 就会太大，而且光强也会减弱。所以插入片段的长度是有限制的。

PhiX 文库有什么用？

PhiX 文库是一种用病毒基因组做的文库。其基因序列已精确知晓，GC 比例约为 40%，与人类、哺乳类的基因组的 GC 比例接近。其基因序列又与人类的基因序列相去甚远，在与哺乳类基因组一些测序时，可以轻松地通过基因序列比对而将之去除。

在测四种碱基不平衡（A、G、C、T 四种碱基的含量远远偏离 25%）的样本时，可以加入大量的 PhiX 文库，以部分抵消样本的不平衡性。例如 ChIPed DNA 测序，或者亚硫酸氢盐处理过的 DNA 文库，或者扩增子测序（PCR 样测序），都可以加入 PhiX，以部分弥补碱基不平衡性。

也可以少量地加入样本，以作为 control library 来验证测序质量。

Hiseq 和 Miseq 有什么差别？

Hiseq 2000 的测序数据产量很高，一条 Lane 一次可以产生 35G 的 Q30 数据，一张 Flowcell 可以产生约 300G 的 Q30 数据。但是测一次序要 9~11 天的时间。所以较慢。

Hiseq 2500 的一张 PE 200 Flowcell 可以给出 60G 的 Q30 数据，测序本身是一天时间，可以快速地以较高的通量给出高质量的测序数据。

Miseq 的测序数据产量低，一次可以产生 1G~4G 的数据。但是测长可以做到较长，目前可以测 250*2。而且测序的速度非常快，一般一天就可以测完一张 Flowcell。

Hiseq 2000 和 Hiseq 2500 有什么差别？

仪器升级：

Hiseq 2500 是 Hiseq 2000 的升级版。

其主要的改进点是：Hiseq 2500 可以在快速、高通量两种模式之间切换。高通量模式就是原来的 Hiseq 2000 的每张 Flowcell 有 8 个 Lane 的模式。

Hiseq 2500 的快速模式，核心的改进是用 2 个 Lane 的 Flowcell 来测序，而且这种快速 Flowcell 的 Lane 比 Hiseq 2000 的 Lane 要短，数据产量也略低于高通量模式的 2 条 Lane。

Hiseq 2500 快速模式的试剂也有所改进。

速度提升：

Hiseq 高通量模式，PE100，双 Flowcell，11 天完成测序。数据量每 Flowcell 在 270G PF data 以上。

Hiseq 快速模式，PE100，双 Flowcel，27 小时完成测序。数据量每 Flowcell 在 60G PF data 以上。

数据质量提升：

在快速模式下，Hiseq 机器可以更快地拍完一个 cycle 的所有照片，也就是每个 cycle 的用时更少。SR50 可以在 1 天内走完，PE100 可以在 2 天内走完。这明显比原来的 3 天(SR50)、11 天(PE100)要快得多。

在速度加快的同时，还带来质量的提升。因为 Hiseq 测序过程中两个主要的物质：酶和荧光剂都是不稳定的，或者说是在融化后（原来是冰冻的）随时间延长而不断降解的。为此 Hiseq 还为试剂准备了 4 度冰格，以减慢其降解。原来的 Hiseq 2000 要走 11 天，现在 2 天完成，这带来了明显的测序质量提升。

实测哺乳类动物的基因组 DNA 文库，Q30 比例可达 85%以上，而且其中绝大部分是 90%以上。

测序长度提升：

而且因为测序质量的提升，也带动测序长度的提升，目前 Illumina 官方支持的 Hiseq 2500 的测长是 PE 2*150。

特别需要注意的，Illumina 目前不直接提供 PE150 的试剂，客户要用 1*PE Cluster kit + 1*PE100 SBS kit + 2*SR50 SBS kit 合起来，才能测 PE150。

直接兼容更多文库：

Hiseq 2500 的快速模式试剂直接支持双 Index 测序模式：

双 Index 是指两个接头各有一个 Index。这样两套 Index 排列组合，一个 Lane 里可以放更多的文库。目前 Illumina 官方试剂是支持 96 个排列组合（ $12 \times 8 = 96$ ），这对充分利用 Hiseq 平台巨大的测序数据产量有很大的帮助。原来的单 Index 是支持单侧 24 种 Index。

这与 Hiseq PE100 高通量模式标准 PE100 试剂只能测单 Index。当然，Hiseq2000b 也可以测双 Index，但是用 4 个 50 cycles SBS kit（每 Kit 保证 58 个 cycles）拼起来（ $58 \times 4 = 232$ ），才可以保证有足够的 SBS 试剂量，因为双 Index 会实际需要 216 cycles，这超过了 200 cycle SBS 试剂可以保证的 cycle 数。

仪器操作更方便：

Hiseq 2500 快速模式可以直接在 Hiseq 仪上进行 Cluster 生成，这大大节约了先要在 cBOT 上生成 Cluster，再要将 Flowcell 从 cBOT 上移到 Hiseq 的麻烦。

但是请注意，如果直接在 Hiseq 2500 上生成 cluster，两条 Lane 就只能上一种预混合文库，而不能象原来的 Hiseq 2000 上那样，两条 Lane 物理分开。也就是说预混合文库中的 Index 一定是要分得开的才行。

当然，快速模式也可以还用 cBOT 生成 cluster，但是那要另外买一个编号为 CT-402-4001（全名：TruSeq® Rapid Duo cBot™ Sample Loading Kit）的试剂盒，这个试剂盒要好几百美元。

试剂操作更方便：

Hiseq 2500 快速模式的试剂是做成 Master Mix 的，也就是酶、Buffer、荧光 dNTP 等都预先混合好了，一大管，拿来一化冻就可以用，很方便。这与高通量模式试剂把酶、荧光 dNTP 分几管的模式是不一样的，高通量模式的试剂因为是分管的，所以使用之前还要人工再混合，这样会多占用一点人工。

Hiseq 2500 的两个机位同时只能运行一种模式：

Hiseq 2500 在一台机器的两个机位同时只能跑同一种模式，也就是要么都跑快速模式，要么都跑高通量模式，而不能一个机位跑快速模式，另一个机位同时跑高通量模式。

Illumina、Roche 454、Life Ion Torrent、SOLID 和 PacBio 的高通量测序仪的优缺点是什么？

Illumina 的测序仪的数据产量高，数据质量也是最高的。因为采用带终止基团的荧光 dNTP，所以在测 Homopolymer（碱基同聚物，例如一串 4 个 T：TTTT）等的时候，不会产生移码错读。

Roche 454 采用的是 pyrosequencing 的测序原理，通过水解 DNA 全成过程中所产生的焦磷，放出光，通过测这光来读出序列。优点是读长最长。但是数据产量是最低的。

Ion Torrent，包括 PGM 和 Proton，采用测量 DNA 合成过程中所释放的氢离子引起的 PH 值的变化，来得到序列。优点是速度最快，上机前约 3~4 天的时间，上机只要 2~4 个小时。

SOLID 采用的是杂交，连接反应，再测荧光的方法。因为杂交，所以速度慢，测长

较短。现在事实上已被淘汰。

PacBio 是三代测序，也就是单分子测序。目前的情况是测序长度可以在 1 个 KB 以上，而且可以测出 DNA 序列的修饰情况。但是其缺点在于测序的准确度很低，目前的测序准确度只有每个碱基 80~90%。另一方面通量较小，一次读 7 万条 reads。

Illumina 测序过程中，Multiplex index 之间会有多少交叉的污染？

我们曾经专门做过实验，用 4 个亲缘关系很远的物种的 DNA，用 4 个 index 标记，进行测序。测序之后进行基因组比对，发现每种 index 之内会有 0.02~0.03% 的 reads 是别的物种的。也就是说因为 Multiplex index 引入的交叉污染，会以 0.02% 上下的比例存在。

这主要是由化学合成 index oligo 过程中的误差引起的。根据我司的引物合成专家的经验，即使经过 HPLC 的纯化，oligo 中还是会有 0.5~1% 甚至更高的错的引物。现在的 0.02% 的污染率，已经是很低了。

Hiseq 和 Miseq 都可以做双 index 测序吗？

Miseq 是天生就可以做双 index 测序的。

Hiseq 要升级到 2500 之后，才可以做双 index 测序。而且，在测的时候要加一个试剂盒：Truseq Dual Index Sequencing Primer Box(下称 Dual Index Box)。

这个试剂盒只能用于一整个 Hiseq 2000 的 Flowcell，也就是说无论一张 Flowcell 中有几条 Lane 是双 index 的，只要其中有一条 Lane 是双 index 的，就需要用一个 Dual Index Box。

我们对一个 Dual Index Box，收取 1000 元人民币的费用。

Dual Index Box 中主要是新加的测第 2 条 Index 的引物。