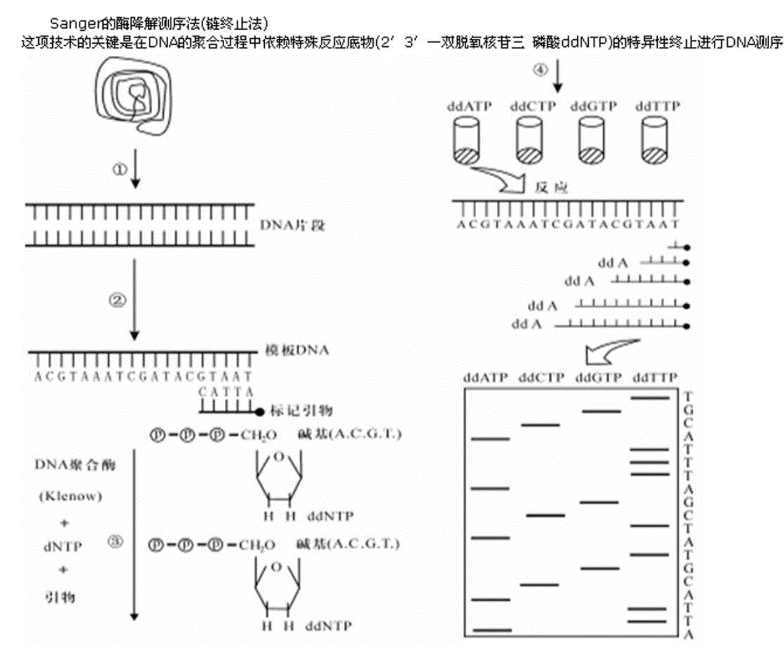
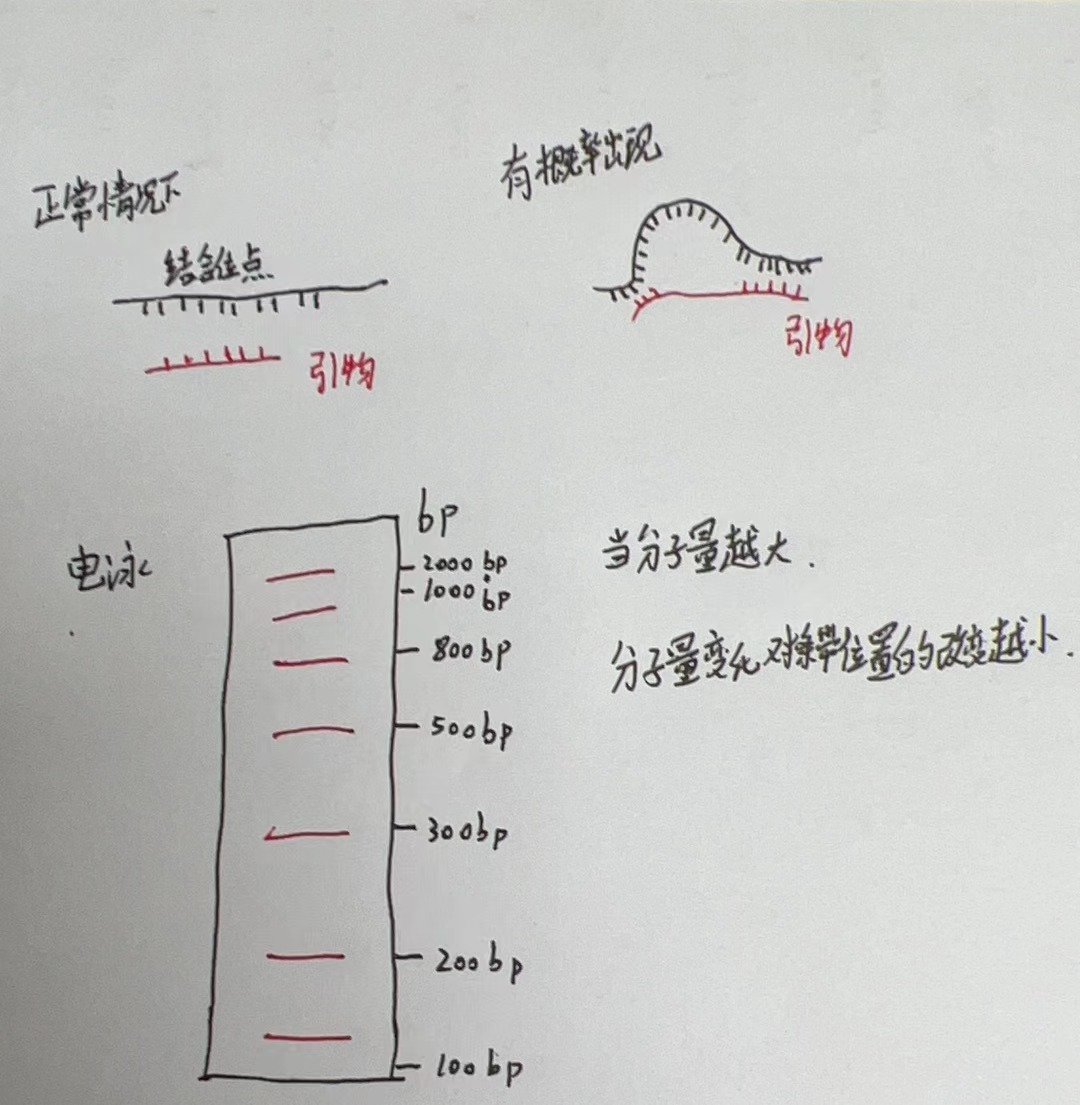
测序说白就是我们需要了解样本中的碱基顺序是什么，腺嘌呤（adenine, A）、胞嘧啶（thymine, T）、鸟嘌呤（guanine, G）、胸腺嘧啶（cytosine, C）其中之一，通过这样的方式去确定reads中的碱基（对）的排列，通过比对核酸数据库确定基因或者是其他产物

**一代测序（sanger）：**核心原理就是‘’双脱氧终止法‘’，因为在PCR扩增中，DNA/RNA链（单链/双链）的合成需要DNA聚合酶催化，接着加入带荧光的脱氧核糖核苷三磷酸（dNTP）-----主要是用来区别ACTG。在DNA聚合酶作用下从5’端向3’端进行延伸反应，当ddNTP掺入时，由于它在3’位置没有羟基，故不与下一个dNTP结合，从而使链延伸终止。ddNTP在不同位置掺入，因而产生一系列不同长度的新的DNA链。

这时候有不同的长度的DNA链，此时我们需要去跑电泳，因为在不同长度下，100bp与200bp之间存在一定距离，小分子量的肯定跑在最前面，而且不同荧光下ATCG分别显示不同颜色，因此提供不同bp最后一位显色确定碱基，最后通过电泳下1bp。。。。。1000bp的最后一位碱基进行排列组合，得到一条序列。



但是，在sanger测序中，开头和结尾的碱基一般都是混乱的，因此需要切除；

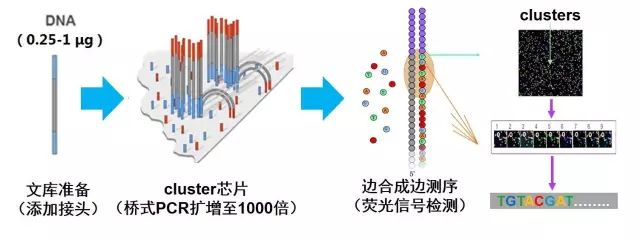


那么问题来了，如何我们要测序超过2000bp的怎么办？

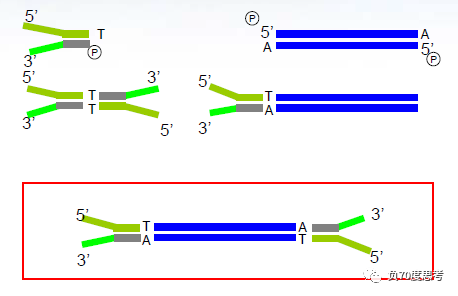
其实需要把序列分成两段进行测序，read1和read2，然后通过overlap重叠部分进行合并，重新组装成新一条序列就可以。当然5000bp怎么办，那就再把read1分成read1-1和read1-2这样下去。但是，这样的成本和工作量太大，一条序列我要5000条read，一个样品至少成千上万reads（但是准确）

所以我们现在一般用的测序都是低成本高通量的二代测序。

**二代测序（下一代测序NGS）：**边合成边测序策略。（Illumina）

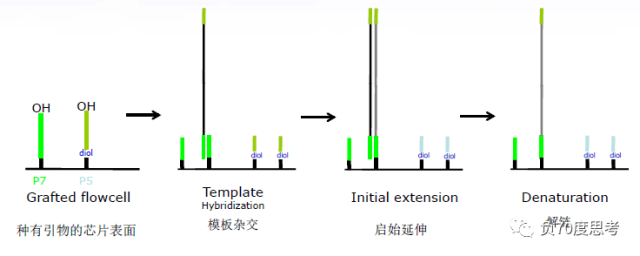


第一步，将样本的DNA进行fragment之后，打断后DNA片段基本都是100-200bp之间，属于小分子量，然后在每个DNA片段的两头添加接头adapter，构成DNA文库。adapter:是DNA片段末端的特定序列，将DNA片段固定在测序芯片上，以便被测序仪器读取。barcode/index：当多个样本并行测序的时候，会很多来自不同样本的reads，分别添加barcode/index用来区分不同序列。Index与adapter之间的位置关系一般为adapter1-Index-fragment-adapter2。

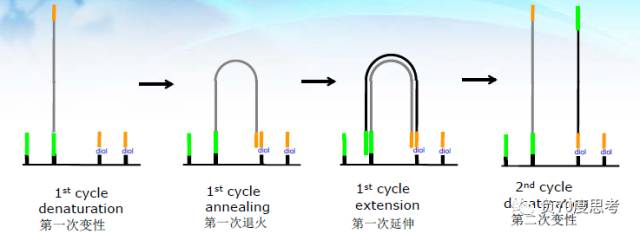


第二步就是在flowcell芯片上进行桥式PCR。首先flowcell是用于吸附流动DNA片段的芯片，分为8个通道，每个channel上面有很多与支持面共价锚定的oligo序列，简称P5/P7序列，可与接头两端的P5/P7序列形成碱基互补配对，进行桥式PCR反应。实质上就是把DNA文库种到芯片上了。

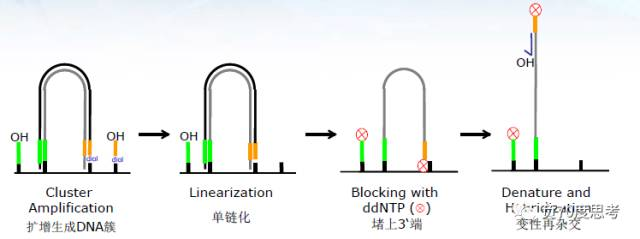




桥式PCR：当退火的时候第一链的P5端接头序列与Flowcell上的P5端簇碱基互补，然后以第一链为模板，以P5端的簇为引物进行延伸。再次变性后，就生成了2条固定在Flowcell上的单链。经过不断的扩增和变性循环，最终每个DNA片段都将在各自的位置上集中成簇，每一个束都含有单个DNA模板的很多拷贝。



桥式PCR完成后，需要把合成的双链变成可以测序的单链，这时把P5端引物上的一个特定基团给切掉，再用碱溶液洗芯片，DNA会解链并且被切下来的单链DNA会被冲走，这样的话芯片上长的链都是统一方向的，P5端在上、P7端在下。

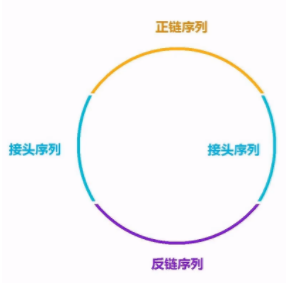


测序：加dNTP（荧光标记、3‘端带叠氮基，导致不能延长），加DNA聚合酶，这样每次合成只能延长一个碱基。用水冲掉多余反应物，进行激光扫描，即可读出一个碱基，即合成一节看一个。（跟sanger测序一样）

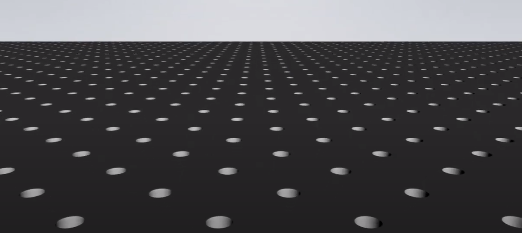
**三代测序：PacBio公司的Single Molecule Real-Time（SMRT）Sequencing**

**同样测序原理跟二代是一样的，都是边合成边测序。但是三代测序非常简单粗暴，通过纳米孔进行测序，而且三代测序是全长测序，并不需要进行PCR扩增，那么data integration只需要考虑batch effect。**

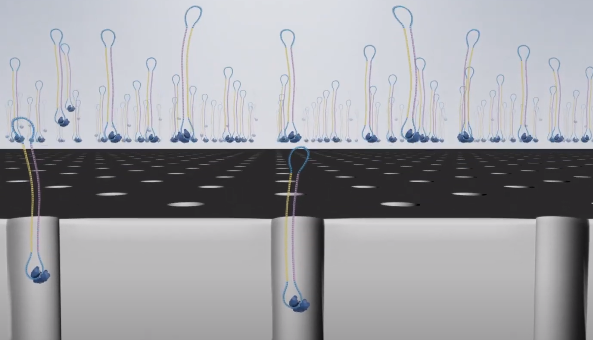
第一步：提取样本中的DNA或者RNA，然后接上adapter，构建文库。

****

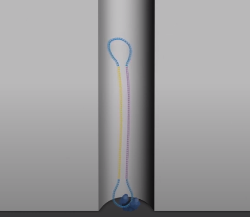
第二步测序：测序仪芯片（SMRT Cell），上面共有15万个直径为70纳米的测序微孔。



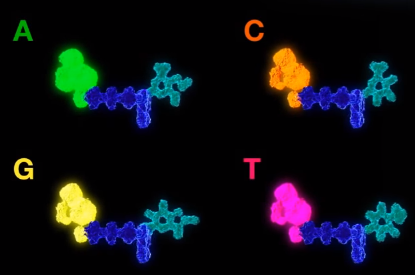
**复合物撒入测序小孔**



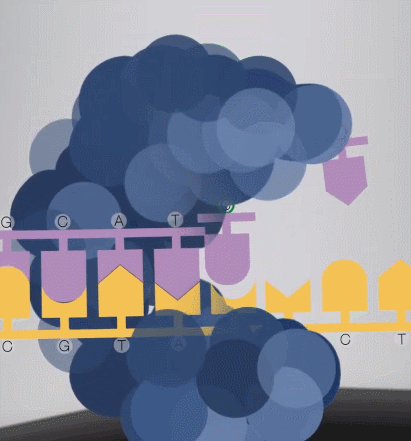
**固定测序复合物**



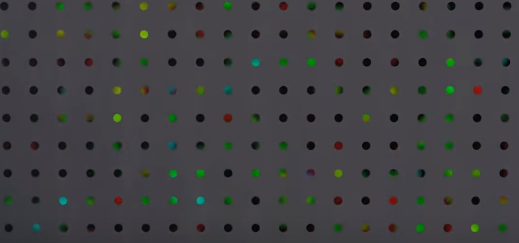
**构建带有荧光基团的dNTP：**ATGC四种碱基的dNTP，在磷酸基团上分别带有四种颜色的荧光基团。



**边合成边测序：**在合成时，游离的dNTP被固定在底板上的酶捕获，激发光会从玻璃板底部发出。由于测序小孔直径很小，激发光的穿透能力会逐渐衰减，只能在小孔中传输很短的距离，所以只有当dNTP足够靠近底部，荧光基团才会被激发光照到，发出荧光。



在发生测序的小孔有各自的DNA片段和测序复合物，同一时间发出不同颜色的激发光，机器会检测到如下的光信号，实际同时会得到多达几万个光点。



但是，如果现在下机，那么得到的raw data肯定是不正确的，**因为PacBio测序是基于荧光标记的原理（Nanopore测序原理是：核酸分子通过纳米孔时，记录电流发生的微小变化，借助机器学习方法将其转换为碱基序列）**，出现碱基读取不准的情况（三代测序的最大缺点：高错误率：12.5%），但是这种错误是随机出现的。

那么，通过上面的环状结构进行循环测序（多次循环读取），这种方法叫做Circular Consensus Sequencing (CCS)，也是SMRT最常用的方法。

现在通过多次测序后得到的下机数据通过SMRT link软件对错误碱基进行纠正得到CCS reads才能用于后续分析。

此时Pacbio-SMRT要优于Nanopore-ONT



