Ion Torrent 总体介绍：

Ion Torrent 的核心技术是使用 IS-FET 半导体技术将生化反应与电流强度直接联系起来。半导体芯片的每一个微孔都装载一个表面带有约 100 万个 DNA 单链的微粒。其测序原理是依次向芯片泵入碱基（T、A、C 或 G）流（flow），如果泵入的碱基与微粒上的模板链互补配对，即可在 DNA 聚合酶的作用下发生聚合反应。每聚合一个碱基，就会释放一个氢离子。H^+的释放会引起周围环境 pH 发生变化，而 pH 的变化可被 IS-FET 场效应管感应并引起电流的变化，最终记录电流信号。如果 DNA 链含有两个相同的碱基，则记录电压信号是双倍的；如果碱基不匹配，则无 H^+释放，没有电压信号的变化。这种方法属于直接检测 DNA 的合成，少了 CCD 扫描，荧光激发等环节，大大缩短了运行时间。同时由于中间无需信号的转换，测序信号准确率也将会大大提升。同 454 一样，Ion Torrent 每次只能聚合一种碱基，4 次循环完成一轮反应（4 个 flow 为 1 个 cycle）（图 1）。Ion Torrent 的应用范围涵盖已有高通量测序技术的应用领域（图 3）。

根据 Ion Torrent 的测序原理可知，它无需光学检测、扫描系统、焦磷酸酶化学级联、标记荧光染料和化学发光的配套试剂。事实上，Ion Torrent 由测序仪器、One Touch EPCR 系统、测序试剂和耗材三部分构成。此外，Ion Torrent 有自己配套的分析软件 Torrent Suite，其功能丰富的插件系统可实现对测序数据的多种分析，包括序列拼接、突变分析、SNP 分析等。自 2010 年 12 月 Ion PGM™ Sequencer 发布后，Life Technology 公司又于 2012 年 9 月推出 Ion Proton™ Sequencer。

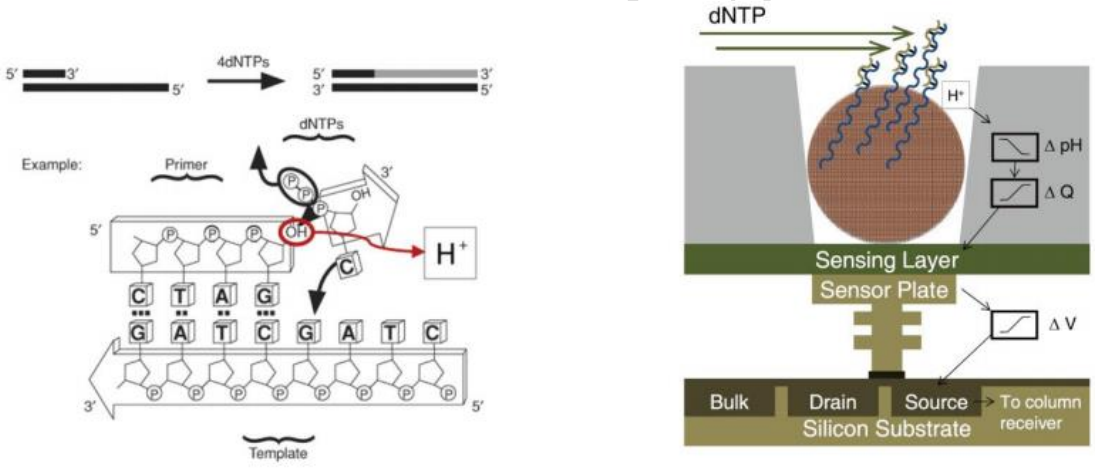

图 1 Ion Torrent 测序技术原理

微孔中有 Ion Sphere™ 微粒，微粒上含有 DNA 模板，每结合一个核苷，就释放一个质子，导致微孔中 pH 变化。感知层检测 pH 变化，并将化学信号转换成数字信号。

与其他测序技术相比而言，Ion Torrent 具有以下技术优势：

1. 系统构成更简单：系统无激光光源，无光学系统，无照相系统；
2. 更准确：使用无标记的天然核苷酸及酶进行测序，本底干扰低；
3. 快速：以 Ion Proton 为例，从 One Touch2 进行乳化 PCR 到测序完毕只需 8 小时，弥补了已有高通量测序方法“无快速测序模式”的缺陷；
4. 灵活：通量为 10M，100M，1G，20G，100G（尚未推出）测序芯片可以任意选择（表 1、图 2）。

	Ion PI Chip	Ion PII Chip
#of sensor	165M	660M
Expected output	10Gb	20x human genome
Filter Reads	60-80M	240-320M
Read Length(Max)	200bp	200bp
Sequencing Run time	~1day	~1day

表 1 芯片 Ion PI 和 Ion PII 数据信息

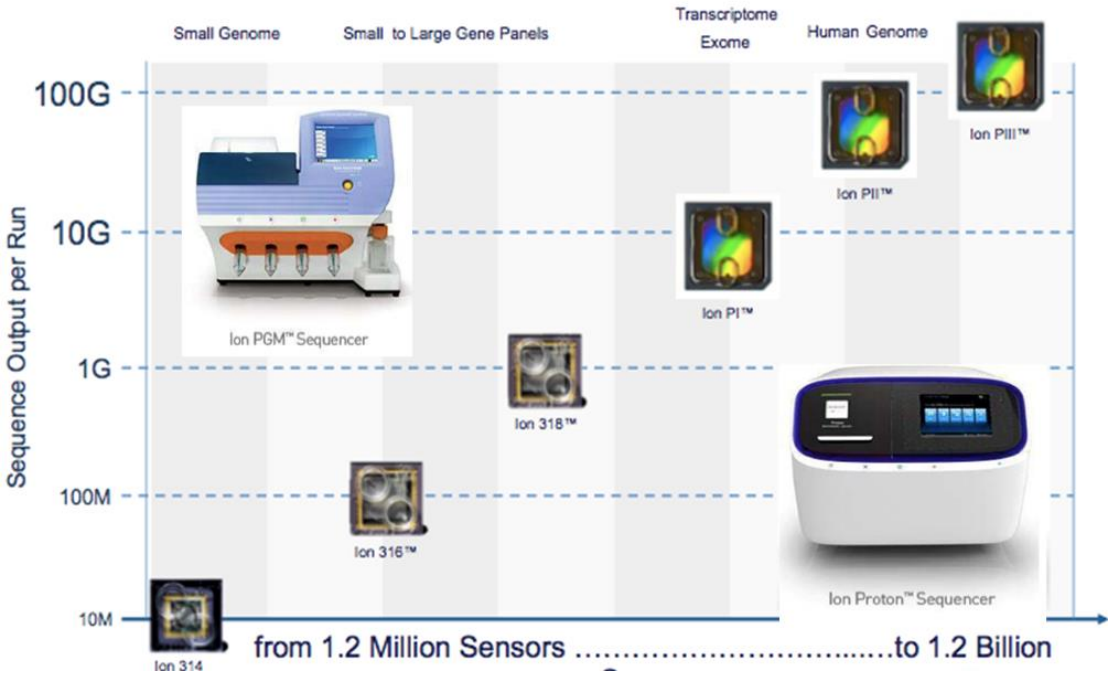

图 2 各种通量的 Ion Torrent 芯片

Ion Torrent 的主要应用（图 3）：

1. 小基因组测序（如：微生物和病毒的从头测序和重测序；线粒体测序等）
2. 扩增子重测序（如：16S 宏基因组测序）
3. 靶向重测序
4. 基因组/全外显子验证

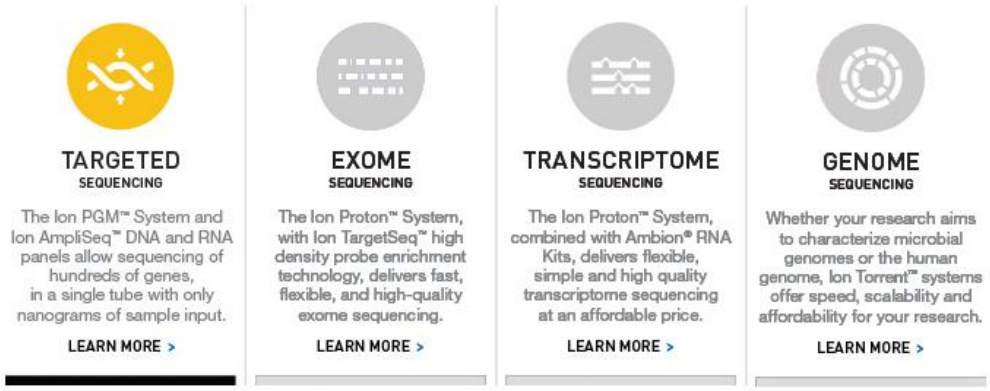

图 3 IonTorrent 的应用范围

以 Ion Proton™ Sequencer 为例介绍 Ion Torrent 实验流程（图 4）：

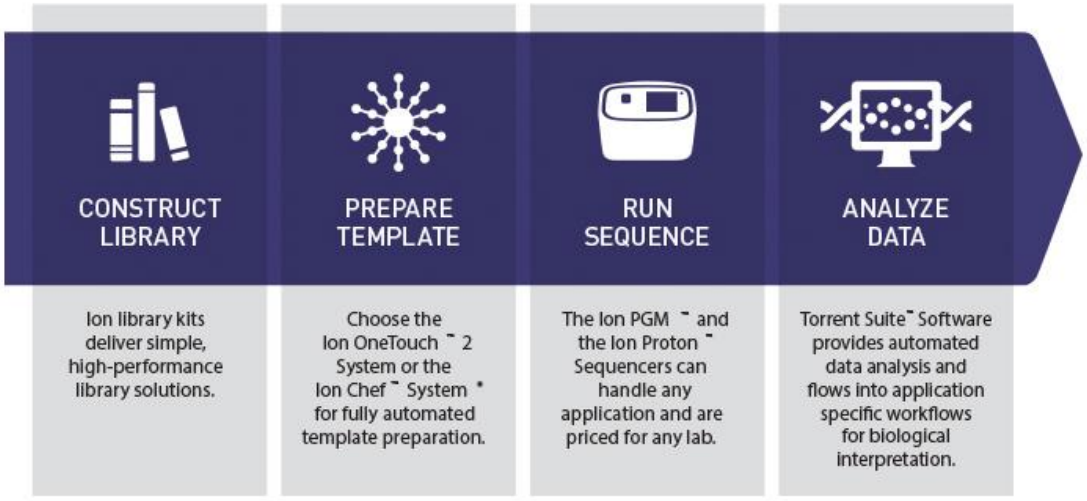

图 4：Ion Torrent 实验流程

1. 文库构建；
   基因组样品-->打断-->末端修复-->加接头-->片段选择-->PCR（PCRfree 或 5-7 cycle PCR）；
2. 模板珠制备（乳液 PCR）；
   1) ISP，即 Ion Sphere Particle，表面带有引物的珠子，乳液 PCR 时文库模板可在其表面进行扩增，也是最终用于上机测序的珠子；无磁性；
   2) 先将包含 ISP 的水相 PCRmix 以及油相准备好，混合成乳液体系后，进行 PCR 扩增，在 ISP 表面上延伸出文库模板，PCR 结束后将所有的珠子从乳液中回收出来，然后通过富集反应，回收带有模板的 ISP 除去空载的 ISP（图 5~7）；
3. 上机测序；
   1) 把保留下来的珠子装载到芯片上，并把芯片置于 Proton 左上角；仪器的下面则装上 ATCG 与 buffer，清洗液等试剂（图 8）；
   2) A “flow” is the event of exposing the chip to one particular dNTP (T, A, C, or G), followed by a washing step；A “cycle” is four consecutive dNTP flows: for instance, T-A-C-G = 1 cycle（图 9）；
   3) 芯片上每个微孔发生聚合反应时会放出氢离子，氢离子改变 pH 值，间接改变电位，测序仪通过电位变化信号来决定其该位点的碱基（图 10）；
4. 数据分析；

整体实验流程时间

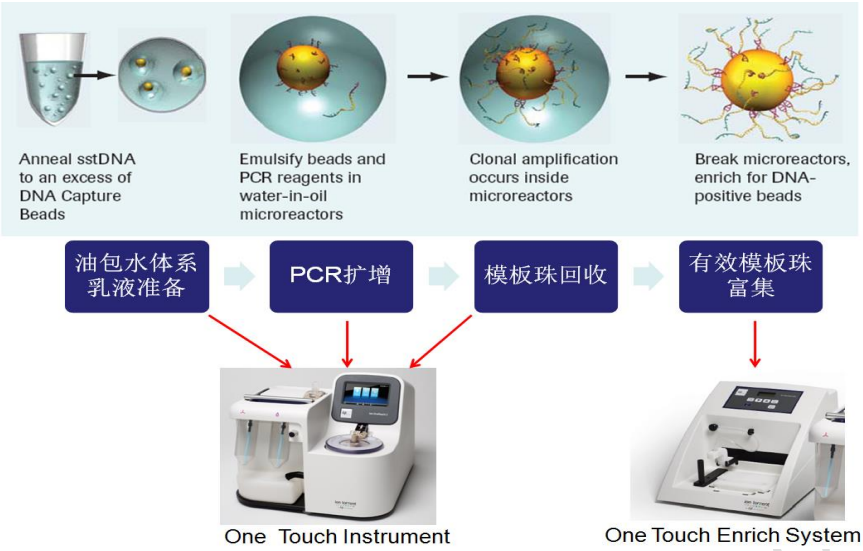

图 5 模板珠制备

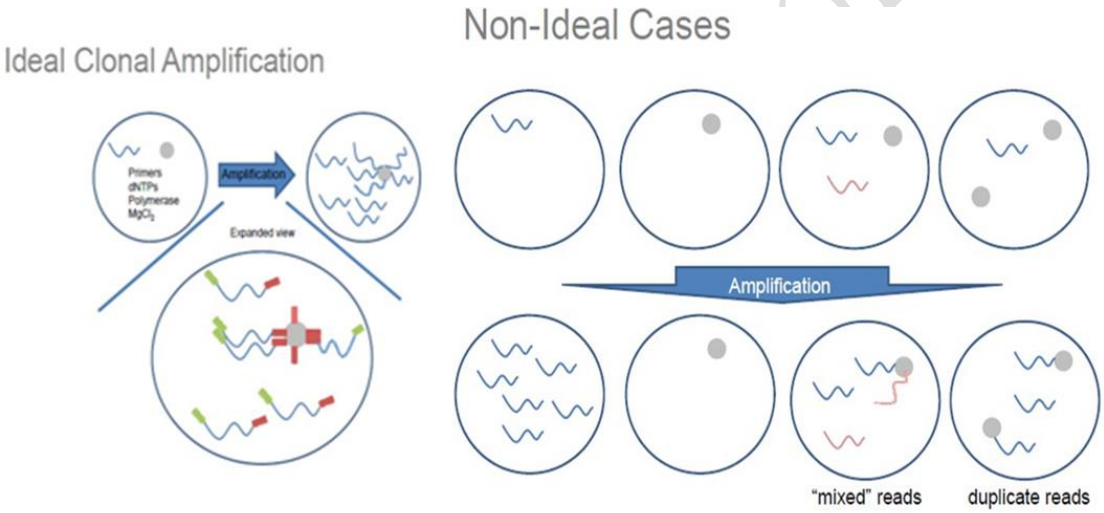

图 6 油包水体系可能出现的情况

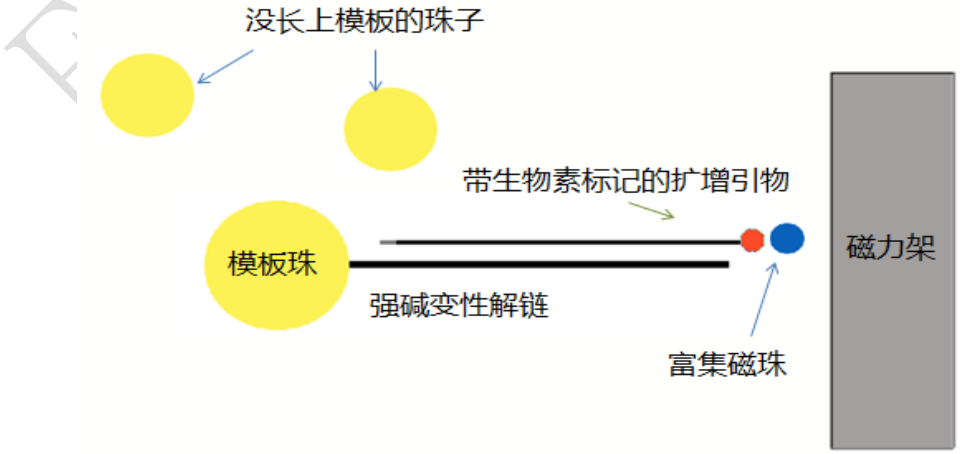

图 7 富集有模板的 ISP

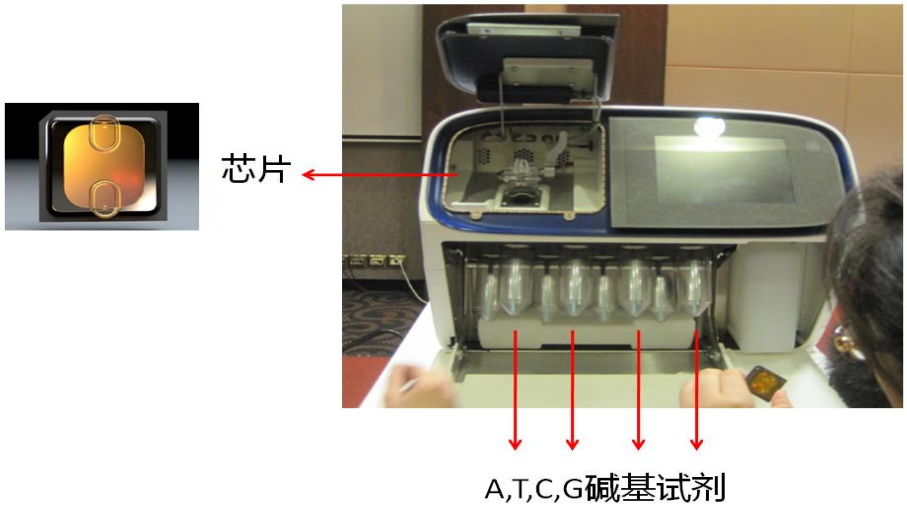

图 8 上机测序

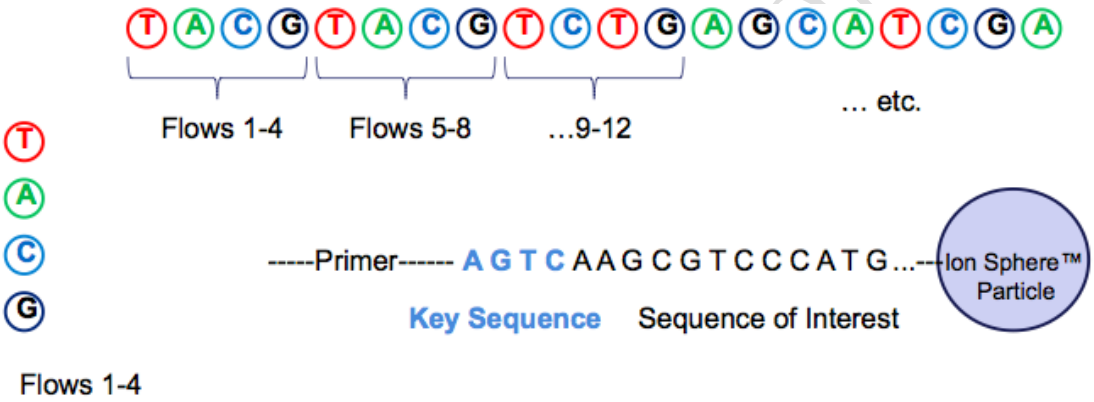

图 9 flow 和 cycle

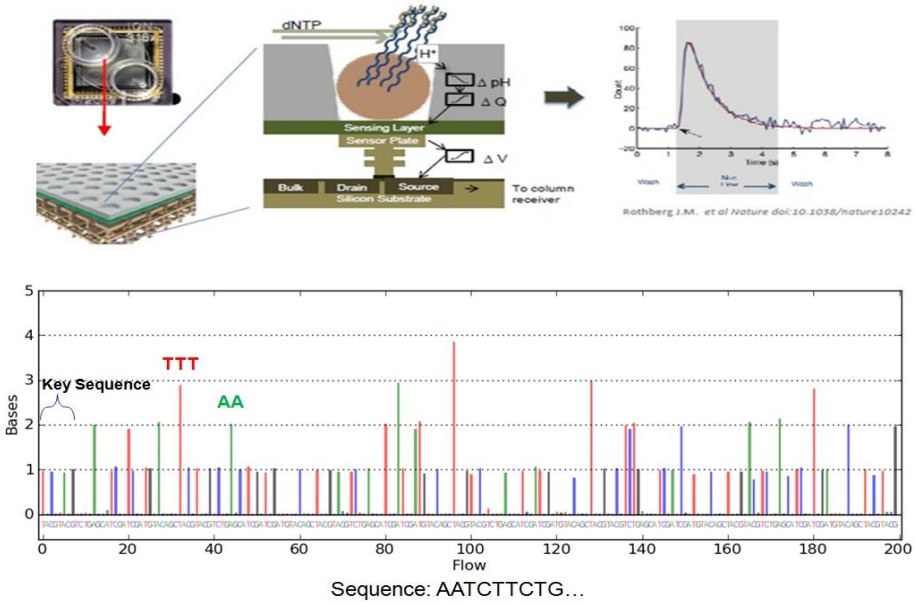

图 10 Base calling

Proton 既可以进行单端测序，也可以进行双端测序，现在已经有 Kit，但目前华大还没有做 Proton 的 PE（图 11）。

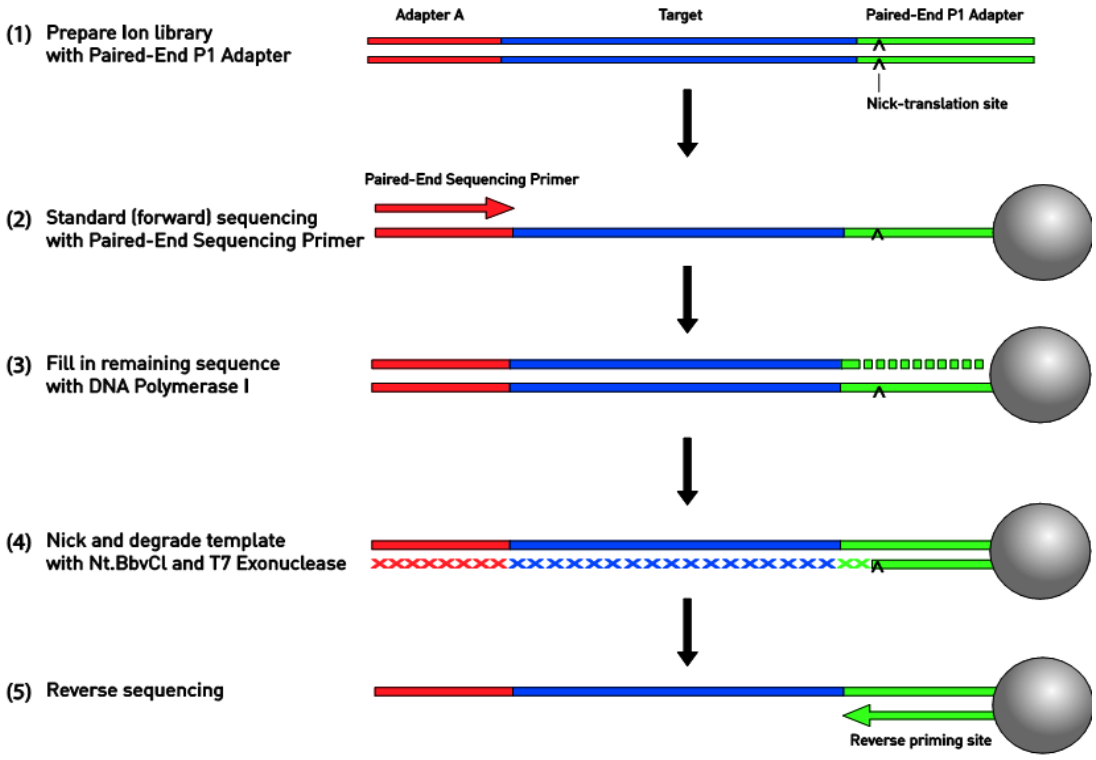

图 11 Proton 双端测序过程

Proton 实验数据报告：Proton 提供“快速数据报告”及“整体报告”。其中快速数据报告，随机挑取几十 M 的数据进行分析，评估整个 run 的数据情况，在测序反应结束后刚进入数据分析阶段十多分钟后即可获得；整体报告，需要在所有数据分析完毕后才能查看（图 12）。

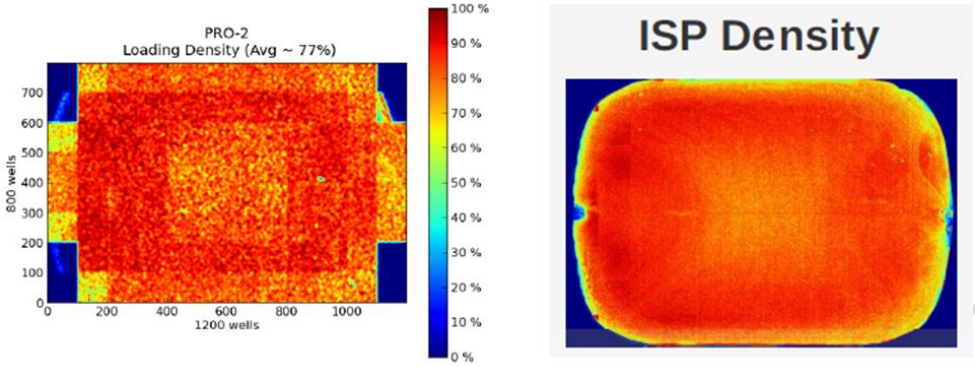

图 12 Proton 快速报告与整体报告的密度图

Proton 数据质量：Slide（图 13） from Chad Nusbaum (Ph.D.,MIT Broad Institute, Co-Director, Genome Sequencing and Analysis Program) illustrated the jump in quality from TS (Torrent Server software) 1.5 to 3.0 by improving the error rate of mismatch and indel. Chad makes the comment that he feels good that 'solving this problem is a software problem'. His Proton data comparison was from one run on a 600Kb region that had a known set of mutations.

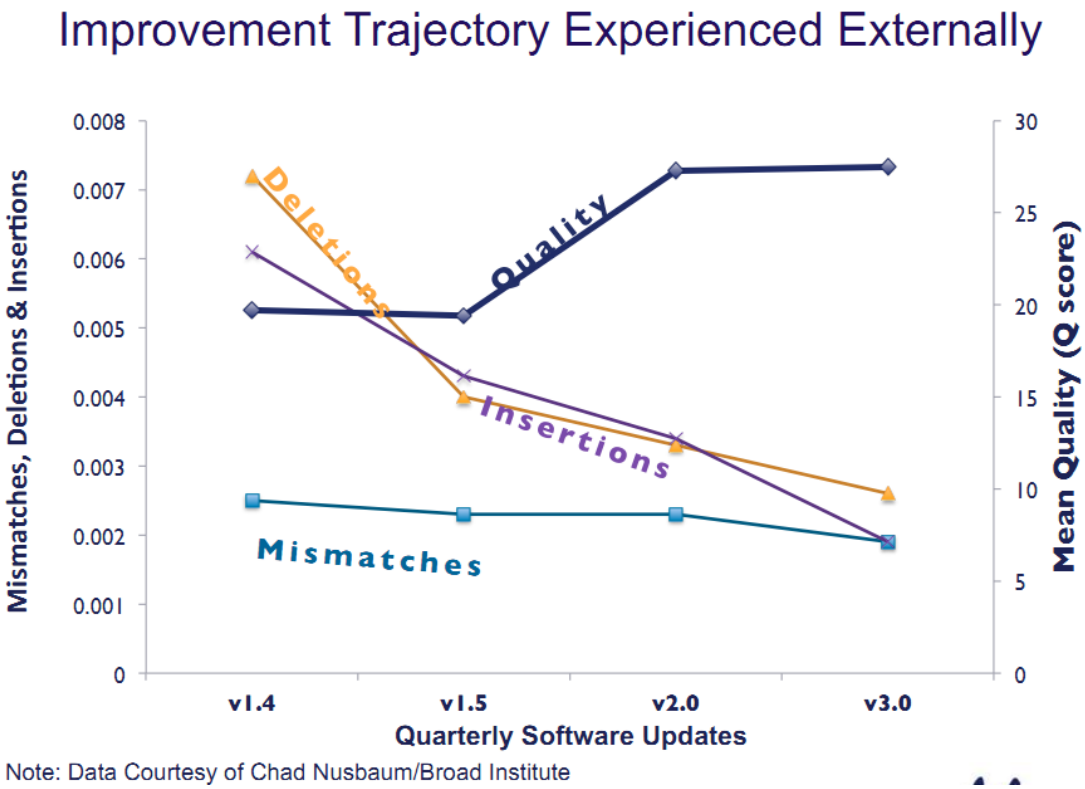

图 13 Proton 数据质量情况

现存 Proton 实验过程中的问题：

1) 乳化 PCR 对文库长度的要求比较严格，例如 100bpKit 的试剂要求文库长度高达 190bp，如果更长的话会影响数据质量；
2) 虽然有 One Touch 2 这样的自动化仪器，但是模板珠制备过程仍存在人手操作部分，时间虽然相对短，但要求高，主要体现在某些试剂容易出气泡但不能吹打出气泡，实验中多处需要经验判断液体剩余体积，ISP 肉眼不可见，对操作要求高
3) 前期测试，无论是 100bp Kit 还是 200bp Kit，平均读长都在 80bp 左右，尚需要进一步研发稳定，提高读长；
4) 仪器软件尚待升级，使用外置电脑进行 run setup 时可能出现失败，而直接在 Proton 界面上进行 setup 必须选择进行比对分析（这样必须先导入参考序列并且无法保留原数据进行后续分析）；
5) 数据质量相对其它平台低；
6) 与其他测序平台相比，目前 Proton 的通量较小（ProtonII 芯片的通量只有 60G 左右）；

Ion Torrent PGM 数据属性：

1. 参考文献：
   1) Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer
   2) Evalutation and optimisation of indel detection workflows for ion torrent sequencing of the *BRCA1* and *BRCA2* genes
   3) Characterizing and measuring bias in sequence data
2. PGM 数据属性及 Indel 挖掘优化：
   1) 文献 3 利用三种微生物基因组（*Plasmodium falciparum*：mean 19% GC；*Escherichia coli*：mean 51% GC； *Rhodobacter sphaeroides* 69% GC）比较了 MiSeq、Pacific Biosciences RS 和 PGM 三个平台的测序数据，结果发现 PGM 的数据在 GC 含量适中的 *E.coli* 基因组覆盖度与 Miseq、Pac Bio 差不多，但是在 *P. falciparum* 和 *R, sphaeroides* 基因组中的出现的偏好非常大（图 14）。因此，测 GC 含量极端的微生物基因组时建议不要采用 PGM 测序。
   2) 文献 3 利用人类基因组比较了 HiSeq、CG 和 PGM 三个平台的测序数据，结果发现三种测序平台在 30%~70% GCper100base 的区域几乎没有 GC bias。但 Hiseq 没有 GC bias 的范围最广（目测 25%~78% GCper100base），而且往两个 GC 极端，bias 产生的变化相对其他两种平台更加缓慢；CG 和 PGM 没有 GC bias 的范围相差无几，但 PGM 在两端的变化更剧烈,可见 PGM 数据在人类基因组上的覆盖度表现比 HiSeq 以及 CG 差（图 15）。
   3) 文献 3 以%GC 100-base windows 做横坐标比较各测序平台的错误率（mismatch、insertion 和 deletion），结果显示 PGM 数据的三种错误率随%GC 变化的波动较 Illumina 和 CG 小，但是三种错误率均比其他两个测序平台高（图 16）。
   4) 文献 3 以同聚物长度做横坐标比较各测序平台的错误率（mismatch、insertion 和 deletion），结果显示 PGM 数据的三种错误率随同聚物长度的增长有上升趋势，而且三种错误率，特别是 insertion 和 deletion 的错误率，比其他测序平台明显高（图 17）。
   5) 文献 1 认为 the PGM produces high frequencies of homopolymer sequencing errors，而且 these errors tend to increase in genomic regions where the occurrence of true polymorphisms is also higher，所以对于 PGM 数据而言，要在 homopolymers 区域挖掘变异（SNVs 和 Indels）困难比较大（区分真的 variant 和测序错误）。
   6) 文献 2 的研究结果表示 Ion Torrent 配套的软件 Torrent Suite 在 variant calling 的灵敏度远低于其他开源 variant caller。作者认为用于 BWA 或 Torrent Suite3.4 或 BWA 做为 mapper，SAMtools 作为 variant caller 可以大大提高寻找 Indel 的灵敏度。但使用这种 mapper 和 caller 的组合会降低特异性（specificity），于是作者提出 4 个参数，分别是 BAF、QUAL、QD 和 VARW。其中 BAF（B-Allele Frequency）定义是 It represents the proportion of reads with the non-reference allele；QUAL 定义是 The QUALity scores of called variants were generated by the variant callers and were provided in their output VCF files；QD（Quality by depth）定义是 It was computed through the division of QUAL by read depth；VARW 定义是 Variation of the width of gaps and inserts；ideally，a true indel would be signaled by reads containing gaps or inserts of uniform width, and any deviation from this criterion suggests a potential error。通过分析，作者认为对于 *BRCA1* 和 *BRCA2* 这两个基因而言，选择 QD=2.5，VARW=0（BWA 做 mapper，GATK 做 caller）或 QD=1，VARW=0（BWA 做 mapper，SAMtools 做 caller）可以有效提高 pipeline 的特异性（即降低假阳性）。

7) 文献 2 的研究结果显示 Torrent Suite 对于 SNVs 的 calling 与其他开源工具差不多（从 sensitivity 和 specificity 两方面而言）。

8) 文献 2 中得到的 QD 和 VARW 阈值值不能直接运用到其他数据集。文章作者建议采用“prediction after training”策略。

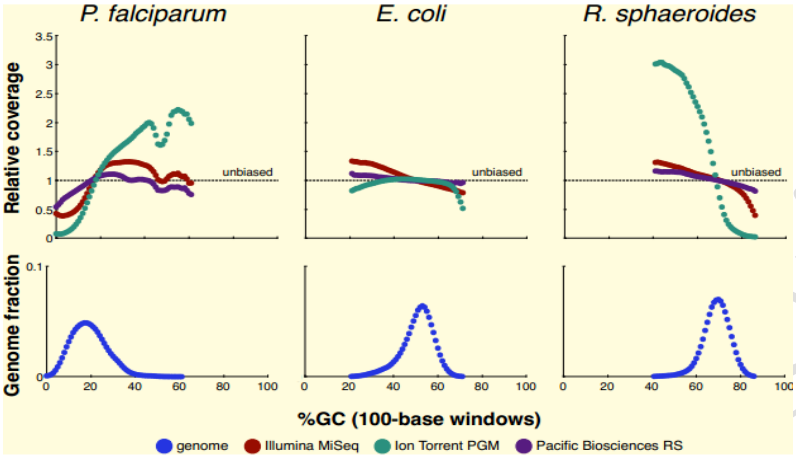

Figure 2 GC-bias plots for three microbial genomes. Top: plots showing the relative coverage GC-bias for Illumina MiSeq, Ion Torrent PGM and Pacific Biosciences RS on the *P. falciparum* (19% GC), *E. coli* (51%), and *R. sphaeroides* (69%) genomes (Table 2, data sets 1 to 9). Unbiased coverage would be represented by a horizontal line at a relative coverage = 1 (black dashed line). Relative coverage is only plotted for GC percentages for which there are at least 1,000 100-base windows in the genome. Bottom: the GC composition distribution of each genome.

图 14

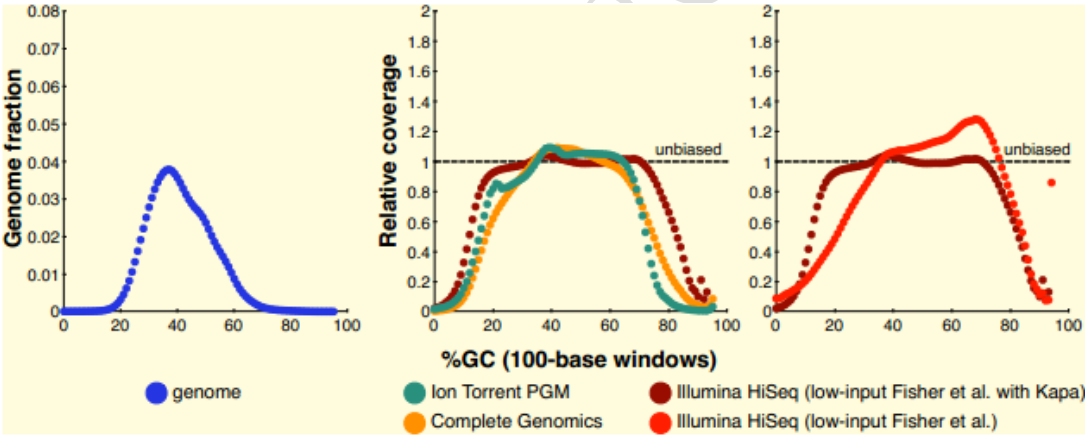

Figure 3 GC-bias plots for the human genome. Left: the GC composition distribution of the human genome (HG19, GRCh37). Center and right: GC-bias plots for several data sets from human NA12878. Unbiased coverage would be represented by a horizontal line at relative coverage = 1. Center: HiSeq v3 with sample-preparation reagents from Kapa Biosystems (Table 2, data set 14), Ion Torrent PGM (data set 15), and Complete Genomics data (data set 16). Right: HiSeq v3 with sample-preparation reagents from Kapa Biosystems (data set 14, as in center panel) and HiSeq v3 with the standard Fisher *et al.* [31] reagents (data set 13). Note that Illumina relative coverage exceeded the y-axis above 93% GC content. Relative coverage is only plotted for GC percentages for which there are at least 1,000 100-base windows in the genome.

图 15

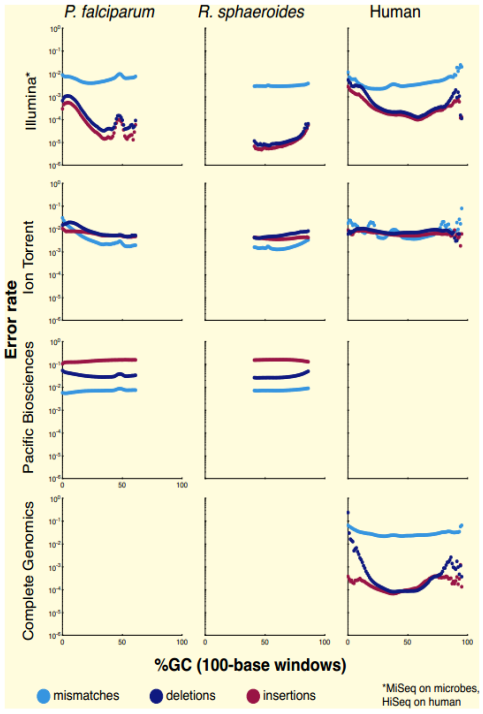

Figure 4 Error rates as a function of GC composition. Each graph shows mismatch (light blue), deletion (dark blue), and insertion (maroon) rates (y-axis) as a function of GC composition (x-axis). Data are shown for the Ion Torrent PGM from three organisms (*P. falciparum*, *R. sphaeroides*, and human), for the Illumina MiSeq on the two microbes, for the Illumina HiSeq on human, for Pacific Biosciences from the two microbes and from Complete Genomics for human (Table 2, data sets 1 to 3, 7 to 9, and 14 to 16). For human we note that *bona fide* differences between the sample and the reference sequence were recorded as errors. Error rates are only plotted for GC percentages for which there are at least 1,000 100-base windows in the genome.

图 16

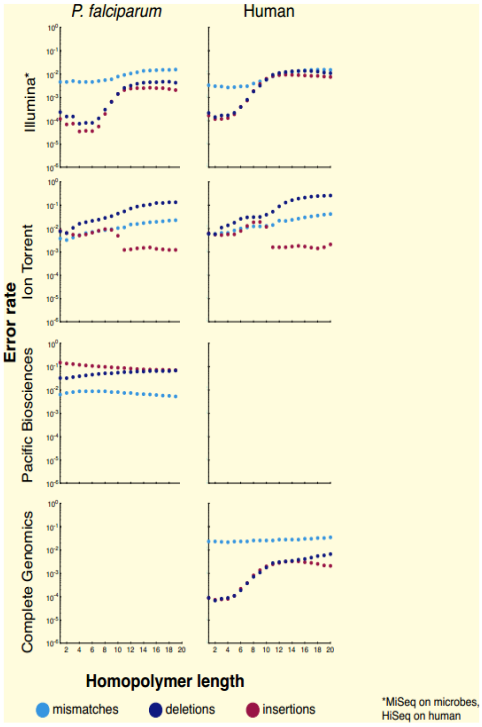

Figure 5 Error rates as a function of homopolymer length. Each graph shows mismatch (light blue), deletion (dark blue), and insertion (maroon) rates (y-axis) within homopolymers of various lengths (x-axis). Data are plotted from *P. falciparum* and human as available (Table 2, data sets 1 to 3 and 14 to 16). For human we note that *bona fide* differences between the sample and the reference sequence were recorded as errors.

图 17

关于 Ion Torrent 数据的想法：

1. Ion Torrent 测序仪因其速度快应该会被广泛应用到临床，但它现在的通量（只有 20x 人类基因组）和 homopolymer 区域的质量是它亟需解决的问题；
2. 如果用 Proton 的人类基因组数据来做成文章的话，我认为可以：

1) 虽然文献 3（Characterizing and measuring bias in sequence data）对 PGM 数据在不同的 motif 的覆盖度进行了评估，但我认为这些评估标准对 PGM 数据不公平，应该采用其他 motif（参见 Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies），计算 Proton 数据在这些 motif 的 coverage bias；（但文献 3 中依据%CG100bp windows 和同聚物长度变化统计错误率可以参考）
2) 因为 coverage 对 variant calling 影响很大，所以把第一步中那些存在明显 coverage bias 的 motif 拿出来，对这些 motif 做文献 2 中提出的 4 个指标，然后确定每种 motif 应该用什么 threshold 去除假阳性；
3) 此外，这些 motif 也可用于寻找适合 Proton 数据的 variant caller，思路参照文献 2。