本文档主要根据以下 3 篇文献进行撰写：

1. Characterizing and measuring bias in sequencing data
2. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies
3. Performance comparison of whole-genome sequencing platforms

其中：

1) 文献 1 在做比较时用的是 relative coverage (coverage of a given reference base in a genome / mean coverage of all reference bases)，而文献 2、3 用的 coverage 即是有多少个 read 覆盖某特定位点。
2) 文献 1 用公式说明，如果要比较准确地识别出覆盖度低的碱基，需要较大的测序深度（deep sequencing is required to accurately identify bases having low relative coverage）。文献 1 三种平台的测序深度不一（Hiseq：120x；Ion Torrent PGM：1.1x；CG：79x），文献 2 在比较时用的测序深度是 30x（Hiseq，CG，SOLiD 4 和 5500xl SOLiD），文献 3 的测序深度是 76x（HIseq 和 CG）。我们在做平台比较时，需要注意这个问题。
3) 文献 1 只考虑常染色体和基因组上非 N 区域（For human data, reads were aligned to the complete reference sequence, but only autosomal contigs were considered in bias calculations. Plasmid, mitochondrial, and sex chromosomes were not included because they are not expected to be equimolar with the rest of the genome）；文献 2 考虑常染色体、X 染色体和非 N 区域（Reference genome regions composed of undefined bses (Ns) as well as chr Y were not considered in our analysis.）；文献 3 没有提供这方面的信息。我认为采用文献 1 的方法来处理炎黄数据。
4) 文献 1 对于比对上多个 locus 的 read 采取的策略是随机分配，而文献 2 只考虑 unique mapped reads，文献 3 没有提及。文献 1 是这样解释它的做法：It is impossible to know whether specific locations are evenly represented, but we can nonetheless expect to accurately assess the coverage of classes of bases as defined by some local sequence context (for example, involving GC content, and so on)。因此，我认为计算 per-base bias 时，只考虑 unique mapped reads，而计算 motif bias 时用随机分配策略。
5) 文献 1 认为当基因组达到一定测序深度后，才能比较每个碱基的覆盖度，而

motif 的覆盖度则不需要这个条件（Per-base bias measurements, which rely on deep-coverage sequencing, are hypothesis-free and ideal for discovering new types of bias. Motif bias measurements, which require only shallow-coverage sequencing, are ideal for comparisons across experimental conditions.），原因是 1 种 motif 在基因组上有很多 loci，当基因组的测序深度很浅时，所有 loci 的平均覆盖度也会比较大，足以用来说明问题（Because motifs are typically represented by may loci in a genome, the number of reads incident upon a motif is much larger than the number of reads incident upon a single base, and hence the relative coverage of a motif (that is, the mean of the relative coverages of its constituent bases) can be accurately measured even with low sequencing coverage）。由于 Ion Torrent PGM 的测序深度太小（1.1x），文献没有比较人基因组每个碱基的覆盖度（per-base bias），但比较了三种微生物基因组（100x）每个碱基的覆盖度。

6) 文献 2 中对一个碱基是否被测到的评判标准是 3 个 reads 覆盖这个碱基（a base was considered not covered if it was supported by less than three reads. The rationale behind this cutoff is that we argue 3 reads are the absolute minimum required to call a heterozygous variant – two reads with a non-reference vase and one with the reference base.）文献 3 没有做出规定。

7) 文献 2 比较 4 种测序平台(Hiseq，CG，SOLiD 4 和 5500xl SOLiD)，发现 CG 测到的基因组范围更广(覆盖基因组碱基数目最多)，CG 30x 时只有 1.61%的碱基没有被测到，50x 时只有 0.79%没有被测到，Hiseq2000 30x 时有 1.45%碱基没有被测到。

8) 文献 3 比较 CG 和 Illumina 两种测序平台的测序深度和测序覆盖的广度(the depth and breadth of genomic coverage by each platform)，结果显示 CG 数据每个碱基的覆盖度均一性较 Illumina 差(即有些碱基的测序深度很大，而有些碱基的测序深度很小)，文献 2 也得到相同的结论(4 种测序平台中，Hiseq 的测序深度范围最窄，CG 最宽)。如果需要获得较为均一的覆盖度，CG 需要更深的测序深度(to achieve a certain level of coverage for most of the genome, CG requires more overall sequencing than Illumina)。

9) 文献 1 和文献 2 对不同 genomic regions (文献 1 定义 motif)的覆盖度进行计算，比较不同测序平台在这些区域的 bias。

10) 文献 1 用到的 motif 见表 1。这些 motif 1~3 取 200~100bp 作为 motif 长度是基于 Illumina 数据读长一般在 100~200 之间，用这样的 motif 作为衡量标准可能会对比较结果产生偏好，所以我认为最好不要选用 motif 1~3 作为衡量标准。文章选用 motif 4~5 的依据是“anecdotal evidence that contig breaks in assemblies are frequently associated with these motif”，我认为可以把两个 motif 作为衡量标准。Motif 6 是根据 Illumina 数据定义出来的 bad promoters，我认为这个 motif 也会对比较结果产生偏好，所以不建议用做衡量标准。Motif 7 是%GCper100bp，我认为这个可以作为衡量标准。文献 1 考虑这些 motif 的相对 coverage，文中用表格形式作为结果展示（见图 1a）。我认为由于文章定义的 motif 是根据 Illumina 数据属性设计，其结果带有很大的偏向性，（人基因组）结果显示 Hiseq 在 motif 1~3 和 motif 6 都比 CG 和 Ion Torrent PGM 的结果好，而在 motif 4 和 motif 5 的结果显示 CG 和 Illumina 相差很小，其中 motif 4 CG 结果比 Hiseq 好。Motif 7 的结果显示三种测序平台在 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，见图 1b。
11) 文献 2 采用的 motif（genomic regions）有 26 种，详见表 2。这些 motif 都可以作为衡量标准。其中 motif 1~25 考虑的是这些 motif 有多少没有被测到（定义如果一个 base 得到的 read support 小于 3 即认为它没有被测到），用图 2a 作为结果展示。Motif 26 考虑的是覆盖度，用图 2b 作为结果展示。在 motif 1~25 中，CG 表现最出色，体现在 CG shows a uniform coverage of almost all regions with a generally very low percentage(<2%) of bases not covered.但 CG 在 simple repeat 的 motif 表现没其他平台好(Hiseq，SOLiD 4 和 5500xl SOLiD)。Hiseq 的表现紧跟其后。对于 motif 26 而言，CG 在 GC-rich 区域的 GC bias 较其他 3 种平台小，但是在（深度为 30x）GC-poor 区域的 GC bias 比其他三种平台大。Hiseq 的表现比 Life technology 的两种平台好。
12) 关于各平台测序准确度的评估，文献 1 直接算 mismatch、deletion 和 insertion 的比例，文献 2 通过比较各平台发现的 SNP 和 Affymetrix SNP6 arrays 得到的 SNP 来判断，文献 3 通过比较 platform-specific SNVs 和 concordant SNVs。
13) 文献 1 计算 mismatch 的方法是 XXX，结果显示 CG 的在 indel 方面的表现极

好，但是 CG 的 mismatch 强差人意。