

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 bases (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 many 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 base 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(见图 2a)。
- 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匀性较 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 2 观察到 higher variations in coverage distribution between samples by CG compared to the other platforms (Hiseq, SOLiD 4 和 5500xl SOLiD)。由于工具

对 variant calling 的敏感度依赖于 coverage, 所以用 CG 数据跨样本分析 variant 时要考虑不同样本的 coverage distribution 会不会对最终的 variant 结果产生影响。

- 10) 文献 2 考察了没有测到或覆盖度低的区域有什么特殊之处。这个我们可以借鉴。
- 11) 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。
- 12) 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 产生的变化相对于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)。
- 13) 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 2b 作为结果展示。Motif 26 考虑的是覆盖度, 用图 2c 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。

- 14) 关于各平台测序准确度的评估,文献 1 直接计算 mismatch、deletion 和 insertion(没有用 SNP calling 工具,只用 BWA 做完 alignment 部分,然后自己统计 mapped、mismatch 和 indel,其中 Illumina 的数据还用 GATK IndelRealigner 和 TableRecalibration tools 进行优化),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 比较了 platform-specific SNVs 和 concordant SNVs。
- 15) 文献 1 的结果显示(横坐标分别为%GC 100-base windows 和同聚物长度) CG 数据的 mismatch 显著高于 Ion Torrent 和 Illumina,但是 CG 数据在 Indel 方面的表现远好于 Ion Torrent 和 Illumina。具体而言,CG 数据在 20%~80% GC(100-base windows)的 indel 只有 10^{-4} ,但这个值在两端的 GC 含量有均有升高;CG 数据的 mismatch 随 GC 百分比的变化而发生相对较小的波动。CG 数据的发生 Indel 的比例随着同聚物长度的增长而上升,在同聚物长度为 10bp 时进入平台期;mismatch 几乎不随同聚物长度的变化而发生波动,但维持在一个较高的水平($10^{-1.5}$)。Illumina 数据随同聚物长度的变化趋势跟 CG 数据一致,但是 Illumina 的发生 mismatch 的比例更低($\leq 10^{-2}$),但 Indel 出现的比例较 CG 高。与 Illumina 和 CG 数据不同,Ion Torrent 三种错误率(mismatch、insertion 和 deletion)随%GC 变化而发生的波动较小,但每种错误率都较高($\geq 10^{-2}$, mismatch 大于 Illumina,小于 CG)。Ion Torrent 数据错误率随同聚物长度而波动的范围也不大,但均维持在较高的水平(见图 1c)。综上所述,我认为用 CG 数据来挖掘 Indel,用 Illumina 数据来挖掘 mismatch。
- 16) 文献 2 和文献 3 结果均显示 Illumina 数据对 SNV 的 sensitivity 高于 CG 数据,文献 2 认为 SOLiD 4 和 5500xl SOLiD 对 SNV 的 specificity 优于 Illumina 和 CG。
- 17) 文献 3 通过比较 Illumina-specific SNV 和 CG-specific SNV 后(Sanger 验证),认为 CG 获得的 SNV 准确度大于 Illumina;CG-specific SNV 和 noncoding RNA 的关联更大;CG-specific SNV 和 Alu 元件、着丝粒以及端粒的关联更大,而 Illumina-specific SNV 和 L1、simple repeat 以及 low-complexity repeat 的关联更大(文献 2 的结果显示 CG 在 simple repeat 的覆盖度低,作者认为是由于

CG 在这种 motif 的覆盖度而导致 CG 在此发现的 SNV 少)。文献 3 认为 concordant SNVs 更可靠, 但是 platform-specific SNVs(这些 SNVs 通常都与 repeat 相关, 可能是因为 mapping difficulty 引起)也非常重要, 不能忽视。

- 18) 综上所述(第 14~16 点), 若要评估各测序平台的准确度有两种途径。其一是文献 1 的做法, 直接统计错误率(mismatch 和 Indel), 以%GC 100-bases window 和同聚物长度作为横坐标。这种做法是默认个体之间的基因组没有差异, 这样结果会使错误率变大, 或者可以先把已知的变异位点筛一遍, 比较剩下的 variant。其二是以 Affymetrix SNP 发现的 SNP 作为 reference, 比较各测序数据 SNP 的 sensitivity 和 specificity。此外, 可以参照文献 3 的思路, 找出 platform-specific SNV 在哪些 region 有富集。
- 19) 文献 1 提到不同建库方法会对测序结果有影响。但是文献中只比较了 Illumina 不同的建库方法, 其他测序平台(Ion Torrent、Pacific Bioscience、CG)都是按各测序平台自己的建库方法做的, 似乎不存在可比性。总得来说, 没有 PCR 过程的建库方法或测序方法, 测序结果的偏向性较小。
- 20) 3 个文献都提到合并不同测序平台的结果。从 coverage bias 的角度而言, 文献 1 认为在绝大多数情况下, 合并不同测序平台数据并不能减少 bias; 文献 2 这种策略只能合并不同平台已有的优势, 不能弥补各自的劣势。从 error bias 或 SNV detection 的角度而言, 文献 2 和文献 3 建议同时测 Hiseq2000 和 CG(If budget permits, sequencing genomes with both Hiseq2000 and CG allows the combination of Hiseq2000' s strength in sensitivity of SNV calling even at low coverage with CG' s strength in uniformly covering the entire genome)。
- 21) 表 4 为文献 2 给出的 4 种测序平台的运行数据(文献 2 发表时间是 2013 年 6 月)。

表 1:

		Motifs	如何获得	是否选做判断标准
文献 1	1	GC \leq 10%,200-base regions in which the middle 100 bases have \leq 10%GC content	in-house scripts	N
	2	GC \geq 75%,200-base regions in which the middle 100 bases have \geq 75%GC content	in-house scripts	N
	3	GC \geq 85%,200-base regions in which the middle 100 bases have \geq 85%GC content	in-house scripts	N
	4	(AT) ¹⁵ ,130-base regions in which the middle 30 bases are repeated AT dinucleotides	in-house scripts	Y
	5	G C \geq 80%,130-base regions in which the middle 30 bases are either 80% Gs or 80% Cs	in-house scripts	Y
	6	bad promoters	public available	N
	7	%GC (100-base windows)	in-house scripts	Y

表 2:

		Motifs	如何获得	是否选做判断标准
文献 2	1	Repeats (all)	downloaded from the UCSC Genome Bioinformatics Site	Y
	2	DNA	a custom Perl script and BEDTools	Y
	3	LINE	a custom Perl script and BEDTools	Y
	4	LTR	a custom Perl script and BEDTools	Y
	5	RC	a custom Perl script and BEDTools	Y
	6	RNA	a custom Perl script and BEDTools	Y
	7	rRNA	a custom Perl script and BEDTools	Y
	8	Satellite	a custom Perl script and BEDTools	Y
	9	scRNA	a custom Perl script and BEDTools	Y
	10	SINE	a custom Perl script and BEDTools	Y
	11	snRNA	a custom Perl script and BEDTools	Y
	12	srpRNA	a custom Perl script and BEDTools	Y
	13	tRNA	a custom Perl script and BEDTools	Y
	14	Low complexity	a custom Perl script and BEDTools	Y
	15	Simple repeats	a custom Perl script and BEDTools	Y
	16	Cancer Gene Census	downloaded from the Wellcome Trust Sanger Institute	Y
	17	Cosmic	downloaded from the Wellcome Trust Sanger Institute	Y
	18	Segmental Duplications	downloaded from the UCSC Genome Bioinformatics Site	Y
	19	Self Chain	downloaded from the UCSC Genome Bioinformatics Site	Y
	20	CpG Island shores	downloaded from the UCSC Genome Bioinformatics Site	Y

续表 2:

		Motifs	如何获得	是否选做判断标准
文献 2	21	CpG islands	downloaded from the UCSC Genome Bioinformatics Site	Y
	22	Promoters	downloaded from the UCSC Genome Bioinformatics Site	Y
	23	Introns	a custom Perl script and BEDTools	Y
	24	Exons	downloaded from the UCSC Genome Bioinformatics Site	Y
	25	Manmal conservation	downloaded from the UCSC Genome Bioinformatics Site	Y
	26	%GC (1000-base windows)	a custom Perl script	Y


注： 表示是我自己推测的，其他都是文章中给出的。

图 1a:

Table 2 Data sets and their relative coverage on bias motifs

					Relative coverage					
Data set					GC extremes			Special motifs		
Sample	#	Library method	Sequencing platform	Coverage (x)	GC ≤ 10%	GC ≥ 75%	GC ≥ 85%	(AT) ¹⁵	G C ≥ 80%	Bad promoters
<i>P. falciparum</i>	1	Fisher <i>et al.</i> ^a with Kapa reagents	Illumina MiSeq	150	0.58	-	-	0.43	-	-
3D7	2	Ion Torrent standard	Ion Torrent PGM	103	0.39	-	-	0.11	-	-
	3	Pacific Biosciences standard	Pacific Biosciences RS	104	0.89	-	-	0.85	-	-
<i>E. coli</i>	4	Fisher <i>et al.</i> ^a with Kapa reagents	Illumina MiSeq	380	-	0.82	-	-	-	-
K12 MG1655	5	Ion Torrent standard	Ion Torrent PGM	311	-	0.31	-	-	-	-
	6	Pacific Biosciences standard	Pacific Biosciences RS	115	-	0.97	-	-	-	-
<i>R. sphaeroides</i>	7	Fisher <i>et al.</i> ^a with Kapa reagents	Illumina MiSeq	388	-	0.94	0.60	-	-	-
24.1	8	Ion Torrent standard	Ion Torrent PGM	302	-	0.39	0.10	-	-	-
	9	Pacific Biosciences standard	Pacific Biosciences RS	142	-	0.97	0.87	-	-	-
Human	10	Aird <i>et al.</i> with Phusion	Illumina HiSeq v2	028	0.58	0.27	0.071	0.38	0.19	0.027
NA12878	11	Aird <i>et al.</i> with Phusion+betaine	Illumina HiSeq v2	048	0.44	0.44	0.28	0.26	0.20	0.14
	12	Aird <i>et al.</i> with AccuPrime	Illumina HiSeq v2	075	0.42	0.42	0.23	0.23	0.38	0.16
	13	Fisher <i>et al.</i> ^a	Illumina HiSeq v3	070	0.29	1.1	0.56	0.23	0.44	0.39
	14	Fisher <i>et al.</i> ^a with Kapa reagents	Illumina HiSeq v3	120	0.41	0.88	0.48	0.25	0.65	0.36
	14'	Fisher <i>et al.</i> ^a with Kapa reagents	Illumina HiSeq v3	000.5	0.41 ± 0.0032	0.88 ± 0.0047	0.48 ± 0.0067	0.25 ± 0.0042	0.65 ± 0.012	0.37 ± 0.022
	15	Ion Torrent standard	Ion Torrent PGM	001.1	0.27	0.36	0.068	0.19	0.26	0.046
	16	Complete Genomics standard	Complete Genomics	079	0.24	0.53	0.18	0.28	0.61	0.092

^aLow-input variation of Fisher *et al.* [31] (see Materials and methods). Data sets from samples, library construction methods and sequencing platforms are shown, along with their total coverage of the genome, and relative coverage, for each of five bias motifs and a set of 'bad promoters' (see text). Entries are blank if the samples' genome had no instances of the given motif. Data set 14' is the summary of ten random subsamplings from data set 14, with coverage reduced to 0.5x, and we show the mean and standard deviations for the relative coverage measurements from it (see text).

图 1b:

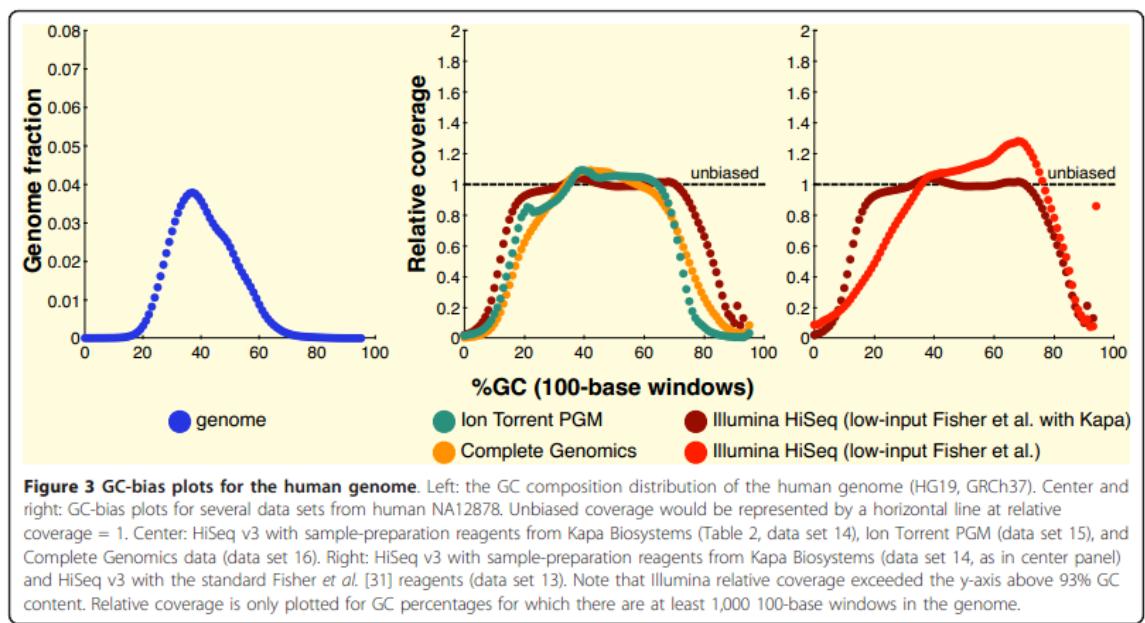


图 1c:

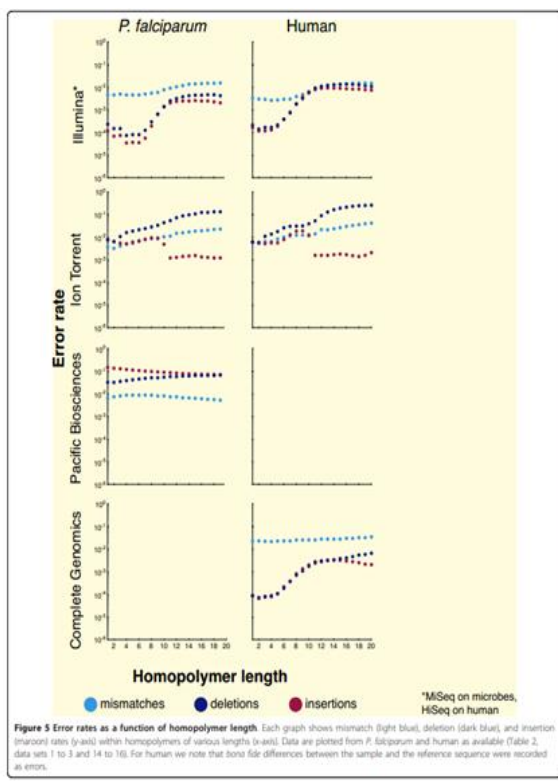
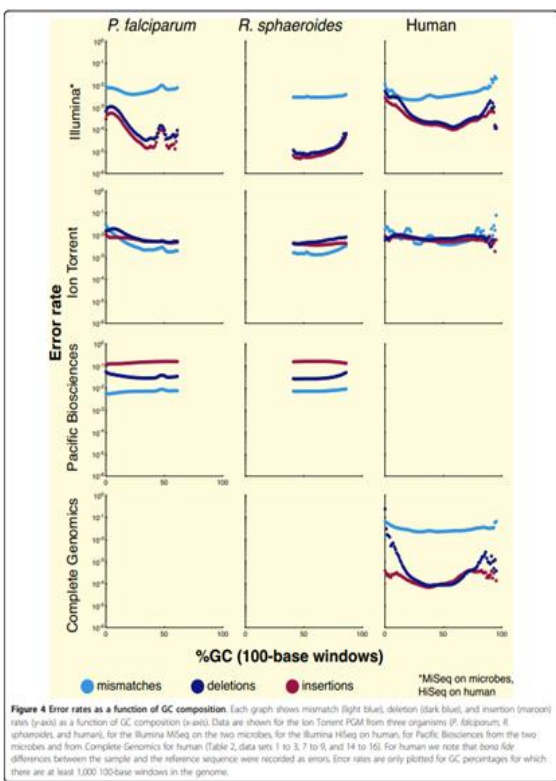


Figure 2a:

Table 3. Number of bases covered on average across all samples from Table 1, and average number of bases covered with less than 5 reads, for each platform assessed.

	Complete Genomics	Complete Genomics 30x	HiSeq2000	SOLID 4	5500xl SOLiD
Total number of bases covered	2,826,524,353	2,817,003,995	2,801,114,390	2,795,379,490	2,772,621,192
Number of bases covered with less than 5 reads	15,938,617	38,555,229	17,727,532	100,145,774	99,297,132

doi:10.1371/journal.pone.0066621.t003

Figure 2b:

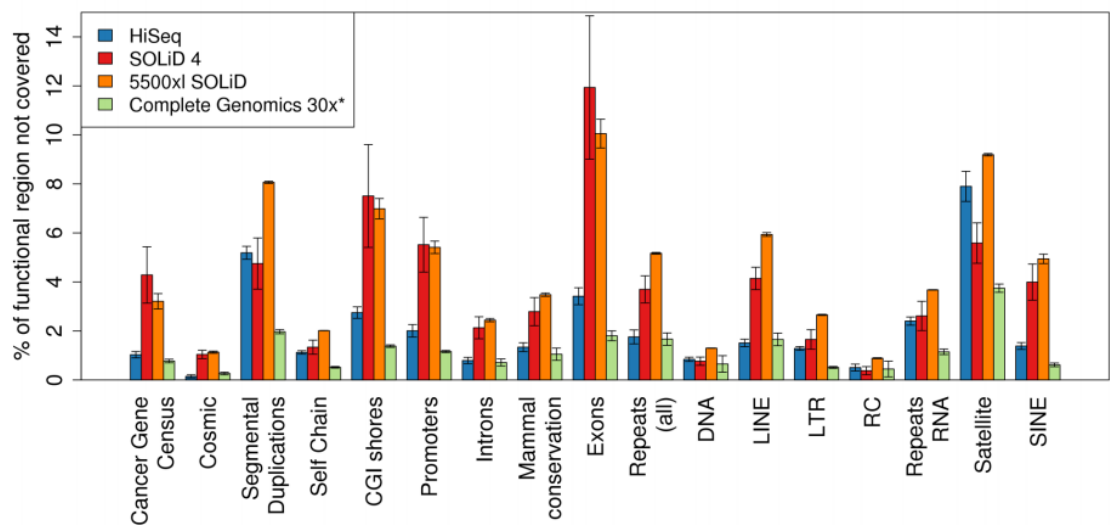


Figure 2c:

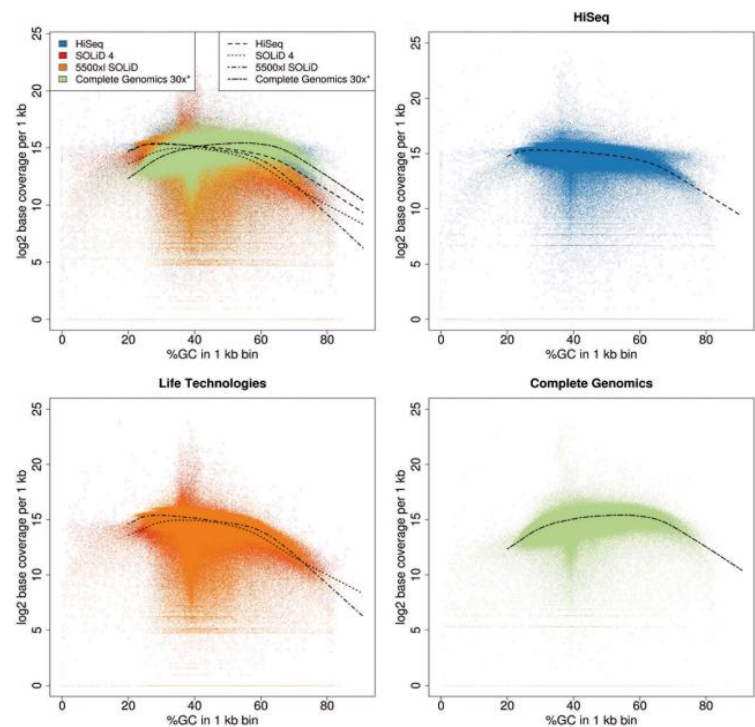


Figure 1. GC bias for each platform. Log₂ base coverage in 1 kb windows versus GC content for HiSeq2000, SOLiD 4, 5500xl SOLiD, and Complete Genomics data. The first panel shows an overlay of all four technologies. The upper right panel shows HiSeq2000 only (blue), the lower left SOLiD 4 and 5500xl SOLiD (red and orange, respectively), and the lower right Complete Genomics at downsampled 30x coverage (light green). Smoothed loess curves are fitted to each dataset to represent the local coverage trend. Exemplary data from patient sample MB24 is shown.

doi:10.1371/journal.pone.0066621.g001

表 4:

	Complete Genomics	HiSeq2000	SOLiD 4	5500xl SOLiD
DNA input amount	8-16 µg	1 µg	1 µg	1 µg
Read length (bp)	2 x 35 (5+10+10+10)	2 x 100	50+35	75+35
Fragment length (bp)	~400	400	230	230
Throughput	30-90 GB/day	55 GB/day	5-7 GB/day	20-30 GB/day