ion torrent
by life technologies™

The Ion PGM™ System, with 400-base read length chemistry, enables routine high-quality *de novo* assembly of small genomes

Key findings:

- Sequencing read length advances on the Ion PGM™ System, enabled by 400-base chemistry, yield improved *de novo* assembly of *E. coli* strain K-12/DH10B, with a 10% to 80% increase in contig N50 values when compared to 200 v2 and 200 chemistries, respectively (Figure 1)
- Single read 400-base chemistry on the Ion PGM™ System delivers better contig N50 values for the *de novo* assembly of *E. coli* strain K-12/DH10B when compared to assemblies produced using 2 x 250 bp paired-end data from the MiSeq™ Personal Sequencer
- The number of total contigs in 400-base assemblies decreased when compared to 200-base assemblies, indicative of a less fragmented and more complete assembly
- Improvements in the 200-base chemistry, in combination with new Torrent Suite Software v3.4, resulted in better accuracy rates and a 63% increase in contig N50 values for *de novo* assembly of *E. coli* strain K-12/DH10B

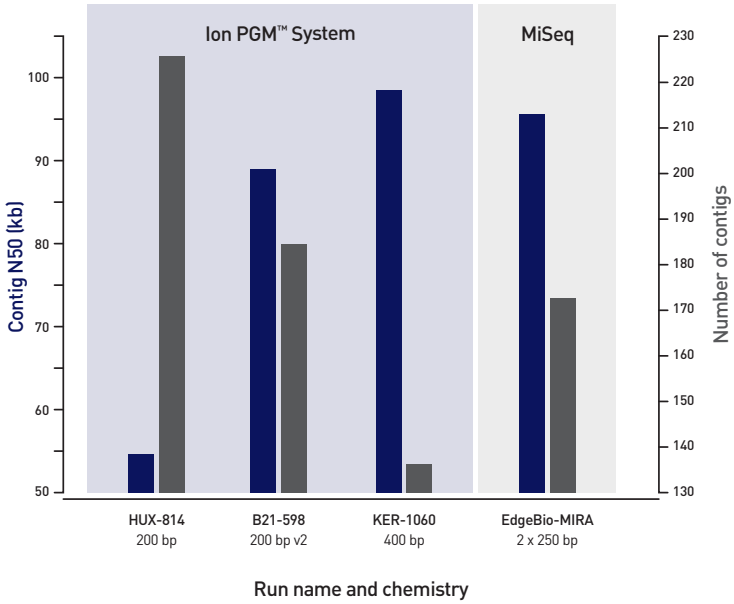

Figure 1. Improvements in 200-base chemistry and the introduction of 400-base chemistry demonstrate the progression of *de novo* assembly on the Ion PGM™ system. With the continued development of 200-base chemistry and the release of long-read 400-base chemistry, the N50 contig lengths increase (in blue, left-hand axis) with concomitant decrease in number of contigs (in grey, right-hand axis) for assembly of *E. coli* (strain K-12/DH10B) sequencing reads. The single-read 400-base data delivers corresponding contig N50 values and an improved, less fragmented assembly (fewer contigs) when compared to an assembly (using the MIRA assembler) of 2 x 250 bp paired-end data from the Illumina MiSeq™ Personal Sequencer.

Better-quality sequence assembly with increasing read lengths

Fast, accurate whole-genome "shotgun" sequencing of microbes is critical for disease surveillance, outbreak investigation, and disease etiology determination. However, the final assembly of sequencing data inevitably contains a number of contigs separated by gaps of unknown size that can create difficulties in determining contig order and orientation without the use of reference strain sequences or additional sequencing information such as mate-pair data. Consequently, improved *de novo* assemblies are beneficial in identifying novel structural and functional genomic arrangements in newly sequenced strains.

life technologies™

One way to improve *de novo* assembly is the use of longer sequencing reads. With short read lengths, determining the relative position of near exact copies of the same repeat or mobile elements in differing regions of the genome may not be possible without enough sequence information outside of repeat or mobile elements. Longer reads enable sequencing through repeated sequences and regions of low complexity sequence that typically hinder accurate assembly. The evolution of Ion semiconductor sequencing enables researchers to take advantage of increased throughput, higher accuracy, and longer reads.

In this application note we show how improvements in read length chemistry can be used to produce a much higher-quality assembled sequence (larger and fewer contigs with improved contig N50 values). Genome assembly metrics from datasets generated using 400-base chemistry on the Ion Torrent PGM™ System are compared with recent *E. coli* (strain K-12/DH10B) data generated using 2 x 250 bp paired-end sequencing on the Illumina MiSeq® Personal Sequencer (Johnson, 2012), demonstrating improved performance of single sequencing reads when compared to paired-end reads for *de novo* genome assembly. In addition, the single reads generated by 400-base sequencing outperform paired-end sequencing in terms of time to result, with only 7.3 hours of sequencing time for Ion 318™ Chip compared to 39 hours for 2 x 250 bp paired-end sequencing reads on the MiSeq™ Personal Sequencer. Time to result can be critical. For example, the speed of Ion semiconductor sequencing was key in the rapid characterization of the shiga toxin-producing *Esherichia coli* outbreak in Northern Germany, allowing researchers to quickly to respond to a serious public health risk. (Mellmann et al., 2011; Rohde et al., 2011).

Results

400-base reads provide greater throughput and high raw accuracy

A comparison of 200-base and 400-base Ion Torrent sequencing runs demonstrate the higher-quality assemblies possible for 400-base sequencing, while illustrating the improvements in the Ion PGM™ Sequencing Kit 200 v2 and the Torrent Suite Software.

Results show that the Ion PGM™ Sequencing 400 Kit yields long read lengths and higher run throughputs than the 200-base runs (Table 1). For the two 400-base runs, the most common read length was 403 and 416 bases with run throughputs that were >2 Gb on a single Ion 318™ Chip. Alternatively, the 200-base runs demonstrated throughputs that were >1 Gb, with the most common read length of 249 and 261 bases using the Ion PGM™ 200 Sequencing Kit and the Ion PGM™ Sequencing Kit 200 v2, respectively.

Using Torrent Suite Software v3.4, the 200-base v2 and 400-base sequencing chemistries exhibited a high mean raw read accuracy of 99.4% and ≥99.1%, respectively (Table 1). Raw read accuracy is measured at each base across the length of the read and is based on 1x sequencing coverage, not based on consensus accuracy across multiple reads for the same base position.

Improved error frequencies for the new 200-base v2 chemistry

Base-called files were aligned to the reference genome for calculation of mapping metrics and error frequencies for the new long-read sequencing enzyme in the Ion PGM™ Sequencing 400 Kit and the high-accuracy chemistry of the Ion PGM™ Sequencing Kit 200 v2. All chemistries demonstrated >98% of reads mapping to the reference, equivalent to the 98% read alignment achieved with MiSeq™ Personal Sequencer 2 x 250 bp paired-end reads data assembled using the Novalign assembler (Johnson, 2012). As a result of chemistry and Torrent Suite Software improvements, both the substitution (sub) and insertion/deletion (indel) error frequencies per 100 bp were improved for the Ion PGM™ Sequencing Kit 200 v2 over the Ion PGM™ Sequencing Kit 200, whereas the sub and indel error frequencies per 100 bp were minimally impacted by the use of the new long-read chemistry incorporated in the Ion PGM™ Sequencing 400 Kit (Table 1).

Better contig N50 values

Accurate *de novo* assemblies are essential to identify novel structural and functional genomic arrangements in newly sequenced microbial strains. The output of *de novo* assemblies is a set of contigs. Ideally a single contig would result from an assembly but repeats, mobile elements, and other regions of low-complexity sequence hinder the assembly process. Important metrics used to assess the quality of an assembly are contig N50 and the total number of contigs that result from an assembly. Contig N50 is the size of the smallest contig, in an ordered set of the largest contigs, that cover 50% of the total size of all the contigs comprising a *de novo* assembled genome. Consequently, larger N50 values correlate to more complete and less fragmented assembly.

Longer read lengths improved *de novo* assemblies, with the 400-base runs generating N50 values of 98.5 and 98.6 kb

Table 1. Run metrics and error frequencies.

Chemistry/ Run name	Number of bases (Gb)	Number of reads (M)	Average read length (bp)	Modal read length (bp)	Mean raw read percent accuracy	Percent reads aligned	Indel error rate/100 bp	Sub error rate/100 bp
200 bp – HUX-814	1.7	7.553	232	249	99.1	99.35	0.79	0.11
200 bp v2 – B21-595	1.2	5.489	233	261	99.4	98.88	0.497	0.08
400 bp – C19-543	2.2	6.907	332	403	99.3	99.75	0.6	0.085
400 bp – KER-1060	2.1	6.310	345	416	99.1	99.68	0.76	0.13

Table 2. *De novo* assembly metrics.

Chemistry/Run name	Average read depth	Number of assembled bases (Mb)	Reads assembled	Number of contigs	Contig N50 (kb)	Contig N90 (kb)	Mean contig length (kb)	Min contig length (bp)	Max contig length (kb)
200 bp – HUX-814	54.16	4.56	1,143,055	226	54.6	13.2	20.2	336	192.5
200 bp v2 – B21-595	48.38	4.54	972,855	185	89.1	25.4	24.6	524	327.6
400 bp – C19-543	54.48	4.57	771,959	158	98.5	27.9	28.9	715	323.0
400 bp – KER-1060	55.92	4.56	759,339	137	98.6	28.9	33.3	706	344.8
MiSeq™ 2 x 250 (MIRA assembly)	57.59	4.61	1,149,448	173	95.6	29.6	21.4	73	292.6

(Table 2). This represents a 10% and 80% improvement over the 200-base v2 run and 200-base run, respectively. *De novo* assembly metrics from the 400-base runs and the MiSeq™ Personal Sequencer 2 x 250 bp paired-end reads are very similar with N50 of 95.6 kb attained using the MIRA assembler and to an N50 of 97.1 kb achieved with untrimmed data (note that the data were subsampled to 30x) assembled using Novoalign assembler (Johnson, 2012). The N50 length improved 63% for the 200-base v2 run over the assembly achieved with the previous 200-base chemistry.

Less fragmented and more complete assembly with high reference coverage

The total number of contigs indicates the level of fragmentation for an assembly. The two 400-base single-read assemblies produced total contig numbers of 158 and 137, less fragmented than the 173 contigs resulting from MiSeq™ Personal Sequencer 2 x 250 bp paired-end reads assembled using the MIRA assembler. There was a progressive decrease in contig numbers through 200-base chemistry kit development and advances in 400-base chemistry that mirrors the progression in contig N50 length improvements (Figure 1).

Contig N50 values and other assembly metrics indicate the completeness of an assembly, but not the accuracy nor coverage of that genome. To assess coverage, the contigs from each assembly were aligned to the reference sequence with *nucmer* from MUMmer package v3.23 (http://mummer.sourceforge.net/). The assembly of 400-base runs resulted in >92% coverage of the reference with the 200-base runs covering >93% of the reference (Table 3).

400-base *de novo* assemblies reflect the DH10B genomic architecture

Typical assembly metrics like N50 do not account for the structure of a particular genome. For instance, due to relatedness of paralogous regions, mobile elements and other low complexity regions within a genome, an assembly of high-quality moderately sized contigs that span inter-repeat/mobile element regions is preferable to an assembly of fewer low-quality large contigs that are misassembled. While fewer and longer contigs are desirable for an assembly, it is equally important that the contigs are correct with fewer assembly errors. Due to insertion sequence (IS) transposition, the genomic architecture of *E. coli* strain K-12/DH10B has been substantially remodeled compared to the related *E. coli* strain K-12/MG1655 (Durfee et al., 2008). Problem regions for assembly such as mobile/phage elements comprise 309,968 bp (6.6%) of the DH10B genome including a 226,519 kb tandemly duplicated region (Figure 2). Due to the structure of this duplicated region, contigs of significant length could not be assembled across this region for any of the runs.

For the DH10B genome, the length (700-2,000 bp) and high sequence identity of the IS sequences largely determines the position of contigs in both 200-base and 400-base assemblies (Figure 2). Of the 65 IS elements (>500 bp in length) present in the DH10B genome, 85% (n = 55) and 83% (n = 54) had contig assemblies break at an IS element for the 400-base (KER-1060) and 200-base (B21-595) assemblies, respectively. Taken further, excluding the duplicated region, the size of fragments between IS elements would result in a theoretical contig N50 of 131.7 kb (min contig = 765 bp; max contig = 326.3 kb; mean contig = 71.9 kb). As a consequence of the DH10B genomic architecture, the 400-base *de novo* assemblies approach the best assembly possible without additional sequence information such as long mate-pair data.

Conclusions

Ion Torrent delivers two optimized sequencing solutions:

1. The Ion PGM™ 200 v2 Sequencing Kit, for targeted gene sequencing applications where accurate variant detection is critical and FFPE samples are often used; and the Ion PGM™ Sequencing 400 Kit, for *de novo* sequencing and some targeted sequencing applications including HLA and 16S rRNA typing that benefit from longer reads, since the key target regions are long contiguous stretches that span large exons or contain multiple adjacent hypervariable regions.

2. For *de novo* assembly, the improved capability to sequence through highly repetitive and homologous regions due to sequencing read length advances on Ion PGM™ System improved contig N50 values 10 to 80% for *E. coli* strain K-12/DH10B. Further, the total number of contigs assembled decreased dramatically with 400-base chemistry. In summary, the Ion PGM™ Sequencing 400 Kit can increase alignment, improve assemblies, and produce greater sequencing throughput.

References

Durfee T, Nelson R, Baldwin S, et al. (2008). The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *J Bacteriol* 190, 2597-2606.

Johnson J. Miseq 2x250 – Does Length Really Matter? by Edge BioSystems: (http://www.edgebio.com/miseq-2x250-%E2%80%93-does-length-really-matter)

Mellmann A, Harmsen D, Cummings CA, et al. (2011). Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6, e22751.

Rohde H, Qin J, Cui Y, et al. (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl Med* 365, 718-724.

Ion PGM™ Data (200-base runs: B21-595; HUX-814 and 400-base runs: C19-543; KER-1060) used in this application note are downloadable from the datasets section of the Ion Community (ioncommunity.lifetechnologies.com/community/datasets)

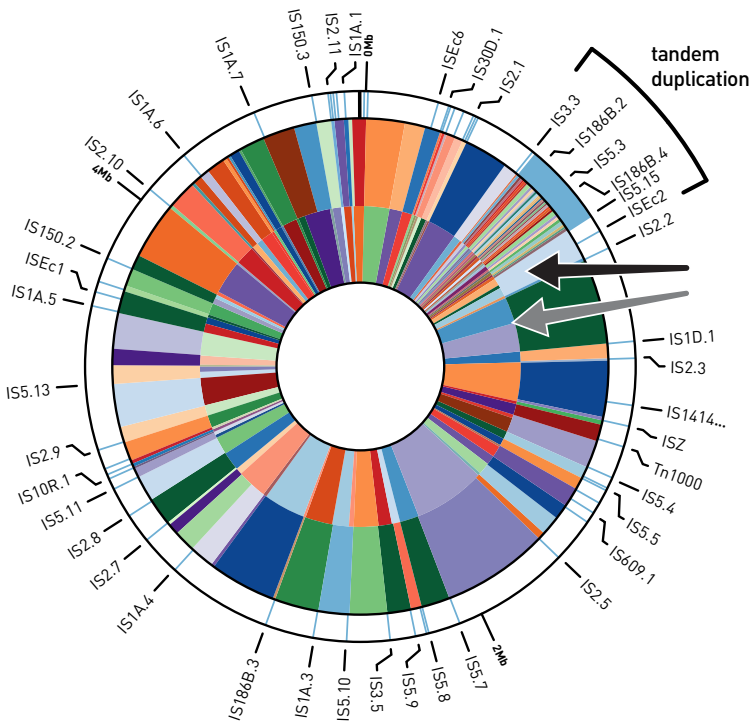

Figure 2. Circle diagram of the *E. coli* (strain K-12/DH10B) genome with the outer black ring and underlying blue boxes indicating the positions of duplicated regions within the genome including insertion sequence (IS) elements and transposons (Tn). The first outer colored ring illustrates the contig alignments to the reference for an assembly of 400-base sequencing reads from run KER-1060. The second inner colored ring illustrates the contig alignments to the reference for an assembly of 200-base v2 sequencing reads from run B21-595. Due the length and sequence homology of the IS sequences, the delineation of the contigs in both assemblies is largely determined by the positions of IS elements in the DH10B genome. Arrows indicate single contigs assembled with 400-base data that were assembled in two or more contigs with 200-base data. Note that the contig indicated by the black arrow, assembled from 400-base reads, is able to span the ISEc2 element that split contigs assembled from 200-base reads.

Table 3. *De novo* assembly quality.

Chemistry/ Run name	Percent reference coverage	Number of gaps	Gap size (bp)
200 bp – HUX-814	93.87	66	281,178
200 bp v2 – B21-595	93.89	68	285,776
400 bp – C19-543	92.58	59	259,435
400 bp – KER-1060	92.73	55	253,053

life technologies™

Supplementary information

Methods

Bacterial genome sequencing

To demonstrate the benefits of increased read lengths, the genome of *Escherichia coli* (4.686 Mb; strain K-12/DH10B) was analyzed using the Ion PGM™ System. Genomic DNA was used to prepare a fragment library using the Ion Plus Fragment Library Kit (Figure 3). Depending on the read length, libraries generated from these samples were clonally amplified with the Ion PGM™ Template OT2 200 Kit or Ion PGM™ Template OT2 400 Kit and the Ion OneTouch™ 2 System for templating and enrichment prior to chip loading (except for run HUX-814, this sample was processed using the Ion OneTouch™ 200 Template Kit v2 DL and the Ion OneTouch™ System). The average library/insert sizes were 247 bp (HUX-814) and 220 bp (B21-595) for the 200-base runs, and 415 bp (C19-543) and 429 bp (KER-1060) for the 400-base runs. Sequencing was performed using Ion 318™ Chip with total sequencing times of 4.4 hours, and 7.3 hours for 200-base and 400-base runs respectively.

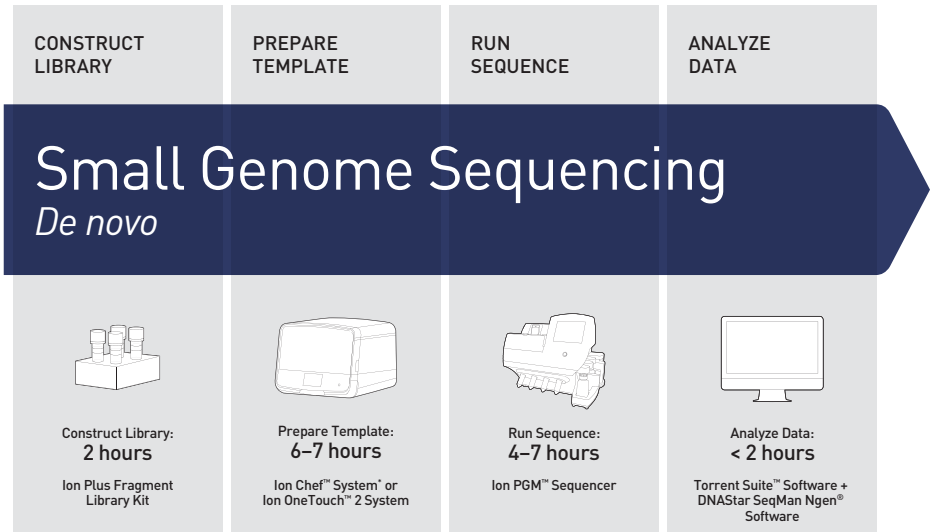

Figure 3. Life Technologies supplies an easy-to-implement, cost-effective, scalable small genome sequencing workflow for the Ion PGM™ System with a rapid <15-hour workflow from library to *de novo* assembly results. Following library construction, using the Ion Plus Fragment Library Kit, and template preparation, 400 bp sequencing runs are completed in just 3.7 hours for the Ion 314™ Chip and 7.3 hours for the Ion 318™ Chip. Primary data analysis is performed using Torrent Suite software and *de novo* assembly accomplished using the DNAStar® SeqMan Ngen software package. SeqMan Ngen® provides an easy-to-use interface for rapid reference-guided and *de novo* genome assembly that minimizes the data analysis time (<2 hours with 10.5 GB of RAM).

Data analysis

Conversion of raw signal to base calls was performed on the Torrent Server using Torrent Suite Software v3.4 for runs using the Ion PGM™ Sequencing Kit 200 v2 and the Ion PGM™ Sequencing 400 Kit (run HUX-814 performed using the Ion PGM™ Sequencing Kit 200 was analyzed using Torrent Suite Software v2.2). Torrent Suite Software v3.4 incorporates algorithmic improvements in both raw accuracy and variant calling, even in challenging regions such as long homopolymers. Base-called files were aligned to the reference genome for calculation of run quality metrics. Output FASTQ files were used for subsequent *de novo* assembly using MIRA v3.9.4 on the Torrent Server as a Torrent Suite plug-in. The data from single runs were randomly downsampled to ~55x coverage. To facilitate *de novo* assembly comparison between 400-base runs and 2 x 250 bp paired-end data, the MiSeq™ Personal Sequencer data was down-sampled to 58x and assembled using MIRA v3.9.4. Open-source assembly tools such as MIRA are described to work proficiently with both Ion PGM™ and MiSeq™ Personal Sequencer data (Loman et al., 2012) Further, comparable assembly metrics were obtained for MiSeq™ Personal Sequencer data using non–open-source assemblers such as Novoalign (Johnson, 2012).

How subsampling impacts *de novo* assembly using DNASTAR® SeqMan NGen

SeqMan NGen® software provides rapid reference-guided and *de novo* genome assembly on a desktop computer and on the cloud. With an easy-to-use interface, small genomes can be assembled using computers with as little as 16 GB of RAM, compute power affordable and readily available to most scientists. An important consideration for improved *de novo* genome assembly is the number of reads used with the constraint that the assembly does not require excessive RAM resources. An Ion PGM™ 400-base run, KER-1060, was subsampled at 600 k, 800 k, 1 M, 1.5 M, and 2 M reads from a 6.3 M total read run to determine the level of coverage that provides the best assembly (as measured by the contig N50 metric), but does not adversely affect assembly time. The assemblies obtained using the SeqMan NGen® v11.0, using default parameters for 400-base Ion PGM™ data, indicate that 1–2 M reads for this particular data set result in the greatest N50 values with 1.5 M reads requiring only 10.5 GB of RAM and <2 hours of assembly time (Figure 4). The results from SeqMan NGen® would ideally be visualized in SeqMan Pro from DNASTAR to support further analysis of contigs, creation of scaffolds, annotation, and, ultimately, finishing of the genome, if desired.

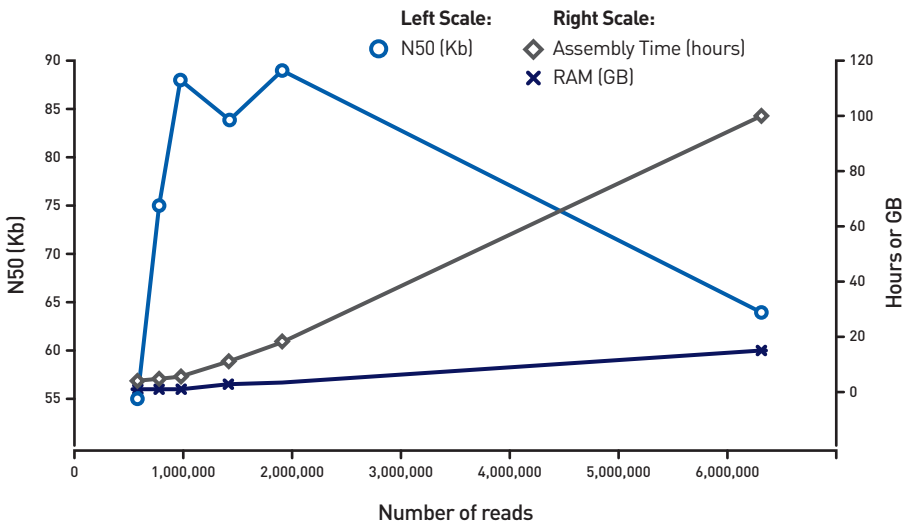

Figure 4. Subsampling reads and the impact on assembly. The contig N50 lengths increase (in light blue, left-hand axis) with an associated increase in assembly time in hours (in grey, right-hand axis) and RAM (in dark blue, right-hand axis) for assemblies of up to 2 million *E. coli* (strain K-12/DH10B) sequencing reads. The additional 4.3 million reads inhibit assembly with detrimental effects on assembly time and RAM.

References

Johnson J. (2012) Miseq 2x250 – Does Length Really Matter? by Edge BioSystems: (http://www.edgebio.com/miseq-2x250-%E2%80%93-does-length-really-matter).

Loman NJ, Misra RV, Dallman TJ, et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30, 434-439.

life technologies™