Log in or Register

SEARCH

Monday, April 30, 2012

Home » News » In Sequence

Comparison of Desktop Sequencers Clarifies Differences, but No Single Platform Comes out on Top

April 24, 2012

Like 0 0 Share 2

By Julia Karow

Type size:

Email

Printer-friendly version

RSS Feed

This article has been updated with comments from a researcher at the US Food and Drug Administration.

Desktop sequencers differ in the quality of data and genome assemblies they produce, as well as in their cost, run time, throughput, and other metrics, but no single platform excels in all categories.

This is the conclusion of the first independent published comparison of the 454 GS Junior, Life Technologies Ion Torrent PGM, and Illumina MiSeq sequencers, which appeared online in *Nature Biotechnology* this week. The study, led by scientists at the University of Birmingham, tested the three platforms for the *de novo* assembly of a bacterial genome.

Notably, the MiSeq had the lowest error rate, while both the GS Junior and in particular the PGM produced considerable homopolymer-associated indel errors. The GS Junior, on the other hand, generated the most contiguous assemblies.

"All platforms are fit for purpose, and they all generate very useful information, certainly much more information than people were used to before whole-genome sequencing came along," said Nick Loman, a bioinformatician at the Centre for Systems Biology at the University of Birmingham and the first author of the study. "But having said that, there are significant differences in the assemblies produced and the kind of confidence you have in the biological inference." While NGS data can always be confirmed by Sanger sequencing, "it's much nicer if you can put the sample in and walk away with a strong answer without having to do too much extra validation."

The sequence data for the study was collected last summer and stems from an isolate of the *E. coli* O104:H4 strain that caused a deadly outbreak of food poisoning in Germany a year ago. At the time, the Ion Torrent PGM had been commercially available for about six months, and the MiSeq was not on the market yet. A number of research groups and instrument vendors used the outbreak as a test case for the performance of their desktop sequencers. Various blogs published analyses of the data while Illumina and Life Tech made claims and counterclaims about their platforms (*IS 6/7/2011*, *IS 7/5/2011* and *IS 9/13/2011*).

By conducting a thorough side-by-side comparison, the UK researchers are hoping to offer users some guidance. "The community was looking for a bit of clarity," said Loman. "We felt that because we had access to data sets from each of the benchtop instruments, we would be in a good position to write up an independent comparison that would give people who are thinking about buying a platform, or bought a platform, some independent verification of the kind of performance one might expect to get from these instruments.

"People are going to want to know not necessarily what's possible in the future, but the kind of results they can expect if they are an average user," he said.

To generate a high-quality draft reference genome of the *E. coli* strain for the comparison, the researchers assembled long-read data from the 454 GS FLX+ as well as paired-end GS FLX Titanium data from an 8-kilobase library, both contributed by 454. That draft assembly consisted of three scaffolds with a total of 153 gaps. In addition,

In this issue of In Sequence

- With Illumina Acquisition on Ice, What is Roche's Next Move in the Sequencing Space?
- **Comparison of Desktop Sequencers Clarifies Differences, but No Single Platform Comes out on Top**
- Record Orders for MiSeq, Consumables Drive Illumina's Q1 Sequencing Revenues
- Survey Finds MiSeq, Ion PGM Tied for Most Likely Sequencing Purchase over Next 12 Months

Transcriptomics Technical Guide

Tech Guide Archives

At which point during ChIP-chip and/or ChIP-seq experiments do you most often run into problems?

During the ChIP/library prep stage

During the data normalization step

While working to estimate false discovery rates

Elsewhere during the analysis

While deciding which data to store, and how

At some other point

Vote

View Results

Video Spotlight

Young Investigator Profile

Blog

Papers of Note

Factors Influencing ChIP-seq Fidelity
Chen, Negre et al., Nature Methods
Dana-Farber Cancer Institute and Harvard School of Public Health's Shirley Liu and her colleagues present a "systematic evaluation of factors influencing ChIP-seq fidelity," which is based on their work using Drosophila melanogaster S2 cells to generate data for the site-specific transcription factor Suppressor of Hairy-wing and a histone modification, H3K36me3. "Paired-end sequencing revealed that single-end data underestimated ChIP-library complexity at high coverage," Liu *et al*. write, adding that "removal of reads originating at the same base reduced false-positives but had little effect on detection sensitivity."

Role of mtDNA in TLR9-Mediated Inflammatory Response in Cardiomyocytes
Oka, Hikoso et al., Nature
Here, a team led by investigators at the Osaka University Graduate School of Medicine in Japan show that mitochondrial DNA that escapes from autophagy "leads to TLR9-mediated inflammatory responses in cardiomyocytes, and is capable of inducing myocarditis and dilated cardiomyopathy."

People on the Move

Upcomi

the scientists had several other reference assemblies from other groups available to verify their results.

They analyzed data from two runs on the GS Junior, two runs on the Ion Torrent and one run on the MiSeq. Data for the GS Junior was provided by the UK's Health Protection Agency; data for the PGM was generated at the University of Birmingham using the Ion 316 chip, available to them under early access at the time; and MiSeq data was provided by Illumina's UK group.

While the startup costs for all three platforms are under $150,000 — ranging from about $80,000 for the Ion Torrent PGM to $125,000 for the MiSeq — they differ in operating costs.

The researchers estimated a cost of $0.50 per megabase for the MiSeq; between $0.63 and $22.50 for the PGM, depending on the chip used; and $31 for the GS Junior. Costs per run were $225 to $625 for the PGM, $750 for the MiSeq, and $1,100 for the GS Junior.

The GS Junior provided the longest reads, with a mean read length of about 520 bases, followed by the Ion Torrent with about 120 bases. The MiSeq, which provided paired-end reads, had a mean read length of 140 bases.

In terms of output per run, the MiSeq led with 1.7 gigabases, the Ion Torrent generated about 300 megabases, and the GS Junior about 70 megabases.

Run times also differed significantly between the platforms: while PGM runs last about 3 hours, a run on the GS Junior takes about 9 hours and a MiSeq run about 27 hours.

For throughput, the PGM came out on top with 80-100 megabases per hour; followed by the MiSeq, with 60 megabases per hour; and the GS Junior, with 9 megabases per hour.

The study did not compare the time required for sample preparation, but the authors noted that the MiSeq workflow "has the fewest manual steps."

All three platforms provided "generally even" coverage across the chromosome of the reference genome.

After aligning the reads from each platform to the reference, the scientists generated so-called alignment quality scores for each instrument, using a previously published scoring system that considers substitutions, insertions and deletions. They used BWA for the alignment, which Loman said works well with all data types, though they also tested other alignment tools.

The MiSeq generated the highest quality reads, they found, because it had the lowest substitution error rate and almost no insertion or deletion errors. By comparison, both the GS Junior and the PGM had considerable indel errors, which were more prominent for the PGM. Most of these errors resulted from homopolymer runs.

For certain applications, where indels are of little interest, those errors can be dealt with algorithmically, Loman said, but for projects where indels are expected, "those indel errors are quite important to understand."

According to Ion Torrent, one reason for the large number of indel errors is that the researchers used a mapping pipeline that is not optimized for Ion Torrent data. "Both our indel and homopolymer performance is significantly better than stated when using our recommended, open-source algorithms," Maneesh Jain, Ion Torrent's vice president of marketing and business development, told *In Sequence*.

"Our most current algorithms cut our indel errors in half and increase the size of called indels as large as 70 bases," he said.

The UK scientists also generated several draft *de novo* assemblies for each of the platforms, and compared their quality in terms of total assembly size, N50 size, and other metrics. They mostly used the MIRA assembler, which supports all three data types, but also employed assemblers considered the industry standard for a certain platform, for example Velvet for MiSeq data and Newbler for 454 data. Loman said there was good agreement between results from different assemblers for the same data type.

All three platforms generated "useful" draft genome assemblies, the scientists noted in the paper, covering the vast majority of the reference genome and coding sequences, though none produced a truly finished genome.

Assemblies from one or both runs of the Ion Torrent, a single run of the GS Junior, or the MiSeq run were "heavily fragmented." They improved when data from the two GS Junior runs were combined or when paired-end information was used to scaffold contigs from the MiSeq data.

While the 454 GS Junior generated the most contiguous assemblies with the fewest gaps and longest contigs, that result could have been even better if the researchers had used a different sequencing strategy, combining shotgun sequencing with an 8-kilobase paired-end library, said Todd Arnold, vice president of R&D at 454. On the GS Junior, that approach "often results in single scaffold assemblies of bacterial genomes," he noted, and is "not only a cost-effective method for generating high-quality genome references but also is extremely useful for genome annotation, comparative analysis, the study of pathogenic islands, and horizontal gene transfer."

Compared to the other two platforms, the PGM assemblies had "large numbers of gaps" as well as many miscalls in long homopolymeric tracts.

Assemblies from all three platforms covered roughly the same proportion of the reference genome, around 96 percent.

To assess the usefulness of the assemblies for understanding the biology and pathogenicity of the *E. coli* strain, the researchers looked whether they could identify in full length a list of 31 clinically important genes.

Assemblies from all platforms revealed the presence and sequence of the two Shiga toxin genes, maybe the most important feature of this outbreak strain. In total, the best MiSeq assembly captured 29 of the 31 genes in full length, the best GS Junior assembly 26, and the best Ion Torrent assembly 23.

The researchers also looked at whether the assemblies could generate accurate multi-locus sequence typing, or MLST, profiles — the traditionally used typing method. While all MiSeq assemblies were able to do so, some GS Junior and PGM assemblies contained indel errors in at least one gene.

None of the assemblies were able to render two of the bacterium's plasmids in a single contig.

Overall, all three platforms generated meaningful biological results, but for some, "you have to work a lot harder and use more of your biological intuition to figure out what's going on," Loman said, for example if genes are split up into fragments. "Obviously it makes your analysis much easier if you are getting full-length genes out, and ideally, you are getting them in a full-length context, so you know whether a gene is on a plasmid or it's on a chromosome. That’s very useful information in an outbreak situation."

One question the paper did not address is how data quality affects the answers obtained during an outbreak investigation, according to Marc Allard, a scientist in the Office of Regulatory Science at the US Food and Drug Administration. Scientists usually want to know whether the isolate has been seen before; how clinical, environmental, and foodborne isolates cluster; and what environmental or food source a comparative genomic analysis points to, he said.

It would also be important to know whether draft genomes from multiple platforms can be combined for accurate strain clustering. "This latter concept will be crucial to the adoption and deployment of a national and international NGS-based pathogen surveillance system," Allard told *In Sequence* by e-mail.

Allard's lab at the FDA has been testing the three desktop sequencing systems as well, which he said have all performed "adequately" with data output matching or exceeding the vendors' specifications.

Genome Data Not Equal

According to Loman, an important finding of the comparison is that "genome data is not equal." Considerable differences result not only from the type of instrument used, but also from the way it is run — for example the depth of coverage generated, read length, or paired-end reads. Desktop sequencer users thus "need to be aware that it's not just a question of getting a whole genome, you have to take those various factors into account."

Users also need to be aware of the strengths and weaknesses of each instrument type, and make their choice according to their desired applications. For example, "a lot of people might assume naturally you can just generate an MLST profile from any whole-genome data from any instrument," he said, however, his team found this was difficult to do in some cases due to low coverage or insufficient read quality.

Since the data for the comparison was generated, at least some of the platforms have been upgraded, raising doubts about the validity of the results today.

The PGM, for example, can now generate 200-base pair reads; a MiSeq upgrade this summer will increase read length to 2x250 base pairs, increase the number of clusters per run, and further reduce run times; and Roche last week announced several upgrades it is planning for the GS Junior, including longer reads and workflow automation.

Ion Torrent's Jain said the performance of the PGM would have been "significantly better" than in the paper if the researchers had used 200-base pair reads, along with its optimized algorithms.

But by and large, the runs and data presented in the paper "are probably consistent with what people are experiencing in their labs" today, Loman said. "The general conclusions are going to be valid for some time to come."

Loman hopes that other researchers will continue to systematically compare data from the desktop sequencers, a moving target. To facilitate this, he and his colleagues published their assembly files, analysis scripts, and strategies in a public online Github repository.

Comparing datasets from different platforms for the same set of reference genomes — for example, genomes of varying complexity such as a bacterium, a yeast, and a human genome — would be "very useful," he said. "Hopefully, this is at least a stake in the ground, a first attempt of trying to do that."

In the meantime, the Birmingham researchers have purchased a MiSeq, which will complement their existing PGM. A major consideration for getting the MiSeq was its easy workflow, Loman said, "a big advantage" for the group, which does not have a lot of wet lab staff.

His team is also looking forward to the arrival of the 400-base pair kits for the PGM, and the 2x250 base pair reads for the MiSeq, which could both potentially replace their existing 454 GS FLX for 16S amplicon sequencing.

According to Stephan Schuster, a professor at Pennsylvania State University, both the MiSeq and PGM platforms have advantages for certain applications. His lab currently owns three MiSeqs and two PGM machines, in addition to four 454 GS FLX+ systems.

For Schuster, the main attraction of the MiSeq is its low sequencing error rate. "I would call it the first sequencer that gets close to 'final quality' data," he said. The instrument also has very short cycle times, gearing it toward high throughput and low costs.

Schuster's lab has had early access to 2x250 bp reads for the MiSeq, resulting in contiguous reads of about 480 base pairs. It now generates those reads for about half its MiSeq samples, he said.

But the emulsion PCR process used during sample prep for both the PGM and the 454 platforms has advantages for sequencing DNA from "dirty" samples, Schuster noted, such as ancient DNA, DNA from soil, or DNA from forensic samples. Those samples often contain inhibitory substances that are associated with the DNA, and "this is where the platforms that use emulsion PCR play out their strengths," he said.

He said he also looks forward to reads in excess of 400 base pairs for the PGM, which, combined with their low cost, could start competing with 454 Titanium reads.

The fact that users move from one sequencing platform to another was also echoed by Illumina. "We believe that the success we have seen in rapidly receiving several PGMs and 454s as trade-ins as part of a recent program correlates with the performance [of the MiSeq] indicated in the publication," said Rob Tarbox, a product marketing manager at Illumina.

According to Schuster, maintaining a choice between different sequencing platforms — including desktop sequencers — will be important going forward, "because otherwise, we end up in a situation as with Sanger sequencing, where for 30 years, there was not a second technology you could use to validate sequences."

Julia Karow tracks trends in next-generation sequencing for research and clinical applications for GenomeWeb's *In Sequence* and *Clinical Sequencing News*. E-mail her here or follow her GenomeWeb Twitter accounts at @InSequence and @ClinSeqNews.

Related Stories

Monsanto Team Sequences Succulent Plant on PGM; First Step in Platform 'Bake-off' for Plant Genomes

January 31, 2012 / In Sequence

With Illumina Acquisition on Ice, What is Roche's Next Move in the Sequencing Space?

🔒
April 24, 2012 / In Sequence

Survey Finds MiSeq, Ion PGM Tied for Most Likely Sequencing Purchase over Next 12 Months 🔒
April 24, 2012 / In Sequence

Macrogen Expands Sequencing Fleet in Preparation for Clinical Sequencing Service Launch 🔒
March 28, 2012 / Clinical Sequencing News

Single-Cell Sequencing Can Isolate Individual Bacterial Species from Metagenomic Samples 🔒
March 13, 2012 / In Sequence

Science

Investigators from the US, France, and Cameroon brought together information on genome-wide ancestry patterns, signals of selection, and more in their search for genetic factors behind diminutive stature of Western African Pygmy populations. A broad search for variants with ties to height in the Pygmy populations pointed to SNPs in and around genes in growth hormone and insulin-related pathways.

Business

Waters surprised the market by reporting a 2 percent decline for its first-quarter revenues, missing analysts' consensus estimate on the top and bottom line. The firm said the shortfall was due to weaker sales in certain developing markets as well as reluctance by several larger pharmaceutical firms to release their capital budgets. However, it said that it expects the pharma market to improve through the year.

Funding

The US Senate Committee on Appropriations has approved a $240 million increase in funding for the National Science Foundation to $7.3 billion, or five percent more than it received this fiscal year, as well as a 10 percent boost for NIST to $826 million for fiscal-year 2013. The overwhelming vote of 28 to 1 supporting these increases may provide cause for relief for those concerned about the potential for steep cuts.

Genome Technology Magazine

Basic research allows for a better understanding of cancer and, eventually, improved patient outcomes. Zhu Chen, China's minister of health, and Shanghai Jiao Tong University's Zhen-Yi Wang recently received the seventh annual Szent-Györgyi prize from the National Foundation for Cancer Research for their work on a treatment for acute promyelocytic leukemia. *Genome Technology* spoke with Chen, Wang, and past prize winners about the state of cancer research.

About | FAQ | Subscriptions | Contact us