Comparing Performance Data – Taking a Different Perspective

By Cliff Reid | Mar 1, 2012 | 09:00 AM

A recent paper in Nature Biotechnology (Lam et. al, 18 Dec 2011) reports on the results of sequencing the same human genome at an average coverage depth of 76x using two different sequencing technologies. The authors, all from the Snyder lab at Stanford University, compared the accuracy and sensitivity of results obtained using an Illumina HiSeq 2000 instrument and Complete Genomics' whole genome sequencing service.

To our knowledge, this is the only published direct comparison of the two platforms, and we applaud the authors for having undertaken this analysis. We eagerly dived into their comparison data to see what we could learn.

1/50th As Many Errors as Illumina …

We agree with their conclusion that Complete Genomics' data is more accurate than Illumina's. Using the Snyder team's Sanger validation data, one can calculate an estimate of how much more accurate:

	Complete	Illumina
Unique SNPs (Errors or Real)	99,578	345,100
Sanger Validation Rate	0.944	0.133
Estimated Total Errors	5,576	299,202

By this estimate, Complete Genomics' platform produced about 1/50th as many errors as Illumina's. The Sanger validation data consisted of a limited number of data points, so the error bars are large (see the Details discussion below for precise numbers), but this is the best estimate that can be made from the Sanger validation data (statistically speaking, this is the Maximum Likelihood Estimator).

... And Higher Sensitivity than Illumina

But when we did the same calculation to estimate the sensitivity of the two platforms, we came to a different conclusion than the authors. While they did not present their calculation of sensitivity, they stated in the Discussion section of the paper:

> "Based on the transition/transversion ratio and Sanger sequencing, CG appears to be more accurate, but also slightly less sensitive."

Using the Sanger validation data, the estimated sensitivity (total number of validated SNPs) of the two platforms is:

	Complete	Illumina
Est. Validated Common SNPs	3,295,023	3,295,023
Est. Validated Unique SNPs	94,002	45,898
Est. Validated Total SNPs	3,389,025	3,340,921

By this estimate Complete Genomics is 1.4% more sensitive than Illumina, not slightly less sensitive. To understand this discrepancy, we contacted Dr. Hugo Lam, the first author of the paper.

Dr. Lam told us that to compare sensitivity of the two platforms they did not use the Sanger validation rate, due to the lack of statistical power. Instead, they used a different validation methodology. They compared the sensitivity of the Complete Genomics and Illumina platforms by selecting targeted portions of the sample using Agilent SureSelect target enrichment, then sequencing those enriched targets at very high coverage (greater than 1,000x) on the Illumina platform.

The paper states (and Dr. Lam reaffirmed) that this validation approach has the drawback of bias toward Illumina. For example, when Illumina gets a SNP wrong in the test sample and then gets that SNP wrong again in the target enriched validation sample (i.e. it makes a systematic error), and Complete Genomics gets that SNP right, the SNP is reported as a Complete Genomics error and an Illumina success. By using the Illumina platform for validation, the more systematic errors Illumina makes, the higher Illumina scores on accuracy, and conversely, the lower Complete Genomics scores on accuracy – despite getting the right answer. Because of such systematic errors, a reliable validation process requires the use of another validation technology, such as large-scale Sanger sequencing, to measure the performance of the technologies being evaluated.

About This Blog

Dr. Clifford Rei
Chairman, Pres and CEO

Dr. Reid is a successful, seri entrepreneur. He enjoys commercializing disruptive computing and life sciences technologies.

Dr. Rade Drma
Chief Scientific Officer

Dr. Drmanac is a genome sequencing pioneer; his inver include massively parallel DN sequencing by hybridization a combinatorial probe ligation. A group leader at the Argonne National Labs, he was part of Human Genome Project. In 1 he cofounded Hyseq, one of first genomic companies.

Legal Notice

Subscribe

From Twitter

@kibosch @slate We think B Banyai of @CompleteGenom superhero too! http://t.co/QuEsCKkt

@Markoff @NYTimes: Bill Ba of @CompleteGenomic used digital expertise in factory des lower cost of mapping the hu genome

@Markoff @NYTimes: Cost Gene Sequencing Falls, Rais Hopes for Medical Advances Complete is making it happen http://t.co/QlEyHrvr

Content Archive

March 2012
October 2011
September 2011

July 2011
June 2011
April 2011

The paper also points to the magnitude of the problem caused by validating the Illumina platform with the Illumina platform. The Sanger validation data can be used to estimate the confidence in the results of the target enrichment validation data. If the Illumina unique SNPs really were 64.3% true SNPs as reported, then the likelihood of getting the Sanger validation results (2 of 15 validated SNPs) is less than 1 in 10,000. While the exact Illumina SNP validation rate is unknown, the Sanger data tells us that we can be more than 99.99% confident that it is less than the 64.3% calculated by this biased validation approach. For these reasons, we believe 64.3% is not the correct number to use in calculating the sensitivity of the Illumina platform in this study.

We are pleased to see platform comparison data published, we enjoy sharing our views on data analysis best practices, and we hope to see more such publications in the future. We commend the members of the Synder Lab for their work, and thank Dr. Lam for his follow-up conversations with us. In summary, we agree with the authors that Complete Genomics technology is more accurate than the Illumina HiSeq 2000 technology, and we suggest their Sanger validation data shows that the Complete Genomics technology is also more sensitive than the Illumina technology.

Details discussion

We created two figures to illustrate our analysis. Figure 1 shows the three initial sets of SNPs: the 3,295K SNPs found by both technologies (shown in green), the 99,578 SNPs found only by Complete Genomics (yellow), and the 345,100 SNPs found only by Illumina (orange). The key question is: of the SNPs found by one technology but not the other, how many are real and how many are errors?

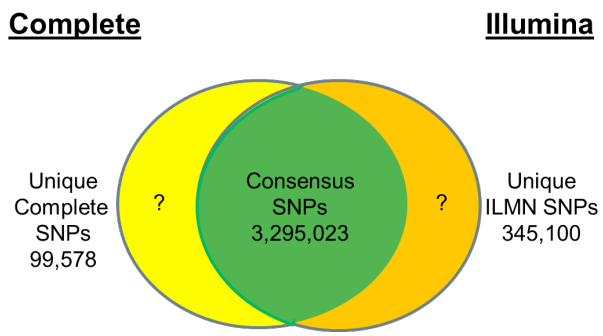

Figure 1. SNPs called by each platform.

To answer this question, the authors selected a few SNPs from each unique set (the yellow and orange SNPs in Figure 1), then went back to the original DNA sample and resequenced those SNPs using an alternative technology (Sanger sequencing, accurate but too expensive for whole genome sequencing, but good for small validation projects such as this). Unfortunately, they tested only 18 unique Complete Genomics SNPs and 15 unique Illumina SNPs, so there is significant statistical uncertainty in the test, but the direction is clear: the Complete Genomics validation rate was 94% (17 of 18) and the Illumina validation rate was 13% (2 of 15).

From these numbers, we can calculate approximately the accuracy and sensitivity of the two technologies (Figure 2). Complete Genomics found an estimated 3,389K true SNPs with 5,576 errors, and Illumina found an estimated 3,341K true SNPs with 299,202 errors. By these numbers, Complete Genomics made about 1/50th as many errors (5,576 vs. 299,202) and found 1.4% more true SNPs (3,389K vs. 3,341K).

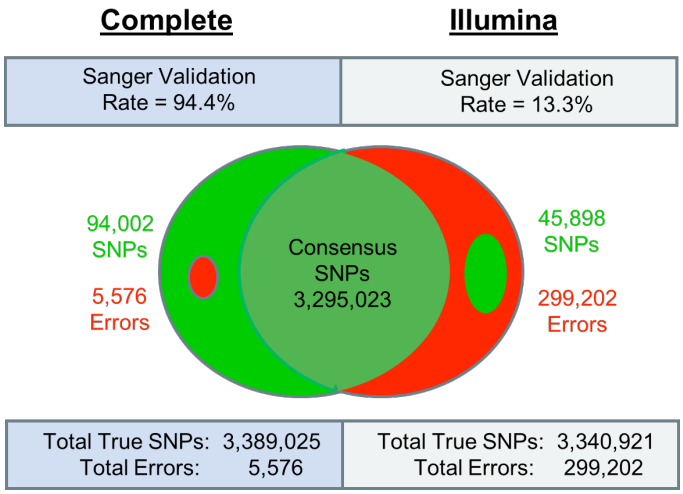

Figure 2: Sanger Validation of SNPs.

The error bars on these estimates are large due to the small number of SNPs selected for Sanger validation. The uncertainly in the mean of the binominal distribution is about ±10% at the 75% confidence level and about ±20% at the 95% confidence level. More precisely, the upper estimate of the mean of the Illumina validation rate is 0.25 and 0.36 at

the 75% and 95% confidence levels respectively, and the lower estimate of the mean of the Complete Genomics validation rate is 0.86 and 0.76 at the 75% and 95% confidence levels respectively. The maximum likelihood estimator, which is what we used in the above calculations, is 0.133 for Illumina and 0.944 for Complete Genomics.

Comments:

Is it true that your product is useful for human genomes only? If so, a pity and short-sighted as so much knowledge about human genes comes from the study of other organisms.

Posted By: jody on 03/08/2012 at 04:55 AM PST

Thanks for your comment. When we started Complete Genomics in 2005, we decided to focus on the one organism that we felt was worth sequencing a million times: humans. By exclusively focusing on human DNA we have been able to optimize our technology and operations for high-throughput whole human genome sequencing, allowing for high quality and accuracy of our data. While we agree that sequencing other organisms is also important in understanding human genes, this has been and will continue to be our focus. Complete Genomics

Posted By: Complete Genomics on 03/08/2012 at 10:28 AM PST

Hi Cliff, Very nice post, thank you. I've written up a substantial personal response (I'm one of the authors of the paper in question, but I'm only representing myself, not the other authors, here) at http://mendeliandisorder.blogspot.com/2012/03/cliff-reid-on-cg-vs-illumina.html Briefly, I agree that CG is highly accurate, but I don't think it's reasonable to use the Sanger data from this study the way you have in this post. In fact, the rather poor performance of the initial Sanger experiment was a major reason we proceeded with using the SureSelect target enrichment assay as our validation strategy because it became apparent that it would require so many Sanger experiments to gain meaning, and we hypothesized that the 2/15 validation rate by Sanger was not realistic. And that shines through in the SureSelect experiment, as well as in subsequent experiments using these data (comparing with other platforms). That said, the difference across the board between CG and Illumina when it comes to variant detection is noticeable. For CG's part, the detection of "bad" variants that pass quality filters is starkly lower than those detected by Illumina. It's what leads us to conclude that the CG approach is highly accurate. Michael J. Clark

Posted By: Michael James Clark on 03/12/2012 at 07:33 PM PDT

We are glad that this blog post is stimulating discussion. Thanks for sharing your analysis and perspective.

Posted By: Complete Genomics on 03/14/2012 at 01:56 PM PDT

Post Your Own Comment:

Username:

Email:

Comment:

Submit

Copyright © 2012 Complete Genomics Incorporated. All rights reserved. Use of this website signifies your agreement to the Terms of Use and Online Privacy Policy. Contact Webmaster