Performance of High-Throughput Sequencing for the Discovery of Genetic Variation Across the Complete Size Spectrum

Andy Wing Chun Pang,*,† Jeffrey R. MacDonald,† Ryan K. C. Yuen,† Vanessa M. Hayes,‡ and Stephen W. Scherer*,†,1

*Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada, †The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada, and ‡J. Craig Venter Institute, San Diego, California 92121

ABSTRACT We observed that current high-throughput sequencing approaches only detected a fraction of the full size-spectrum of insertions, deletions, and copy number variants compared with a previously published, Sanger-sequenced human genome. The sensitivity for detection was the lowest in the 100- to 10,000-bp size range, and at DNA repeats, with copy number gains harder to delineate than losses. We discuss strategies for discovering the full spectrum of genetic variation necessary for disease association studies.

KEYWORDS
copy number variation
insertion/deletion
high-throughput sequencing
genome variation
annotation

Insertion/deletion (indel, unbalanced change <100 bp) and copy number variation (CNV, unbalanced alteration 100 bp upwards) are increasingly observed to be important in development and disease (Lee and Scherer 2010; Weischenfeldt *et al.* 2013). However, in our experience, it has been difficult to detect indels and CNVs, even when the latest high-throughput sequencing (HTS) technologies are used (Pang *et al.* 2010). Although the detection of single-nucleotide variation by HTS seems sufficient (Lam *et al.* 2012b), the short reads of HTS limit the detection of larger and more complex genetic variants, and that limitation can hamper disease studies.

MATERIALS AND METHODS

To investigate the robustness of indel/CNV calling using HTS, we assessed data from commercial genome sequencing vendors and found that Complete Genomics (CG) (Drmanac *et al.* 2010) detected the greatest number of variants and yielded a more consistent and even variant size distribution (Supporting Information, Figure S1 and Table S1). To evaluate the quality of the CG variation (unbalanced genetic variants) profile, we chose to compare the structural variation data from a comprehensively characterized personal genome, namely the HuRef Standard (Levy *et al.* 2007; Pang *et al.* 2010), to 80 CG-sequenced genomes. One of the 80 genomes was HuRef, herein called HuRef CG (Table S2). The HuRef Standard assembly is of greater quality than HTS-generated genomes, since it was produced from high-accuracy Sanger-based sequencing of long mate-pair clone-end sequences. Using a combination of sequence- and microarray-based strategies, we detected 791,873 gains (insertions: size <100 bp or retrotransposons; duplications: size ≥100 bp) and losses (deletions) in HuRef relative to the National Center for Biotechnology Information reference assembly (Levy *et al.* 2007; Pang *et al.* 2010) (Table S3). Experimental validation confirmed 88% (184/210) of the variants (Levy *et al.* 2007; Pang *et al.* 2010). Details can be found in File S1.

RESULTS AND DISCUSSION

First, by comparing the HuRef CG and HuRef Standard variation profiles, we noticed that short-read sequencing detected fewer calls and had substantial drops in discovery along the variation size spectrum (Figure 1, A and B). There were 241,033 gains and 230,737 losses in the HuRef CG data, which was a fraction of HuRef Standard's 408,403 gains and 383,470 losses (Table S3). For losses, HuRef CG detected 60% of the total number of HuRef Standard

doi: 10.1534/g3.113.008797
Manuscript received October 3, 2013; accepted for publication October 26, 2013; published Early Online November 5, 2013.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.008797/-/DC1.
[1] Corresponding author: The Centre for Applied Genomics The Hospital for Sick Children 13th Floor, Peter Gilgan Centre for Research and Learning, 686 Bay Street, Toronto, Ontario M5G 0A4, Canada. E-mail: stephen.scherer@sickkids.ca

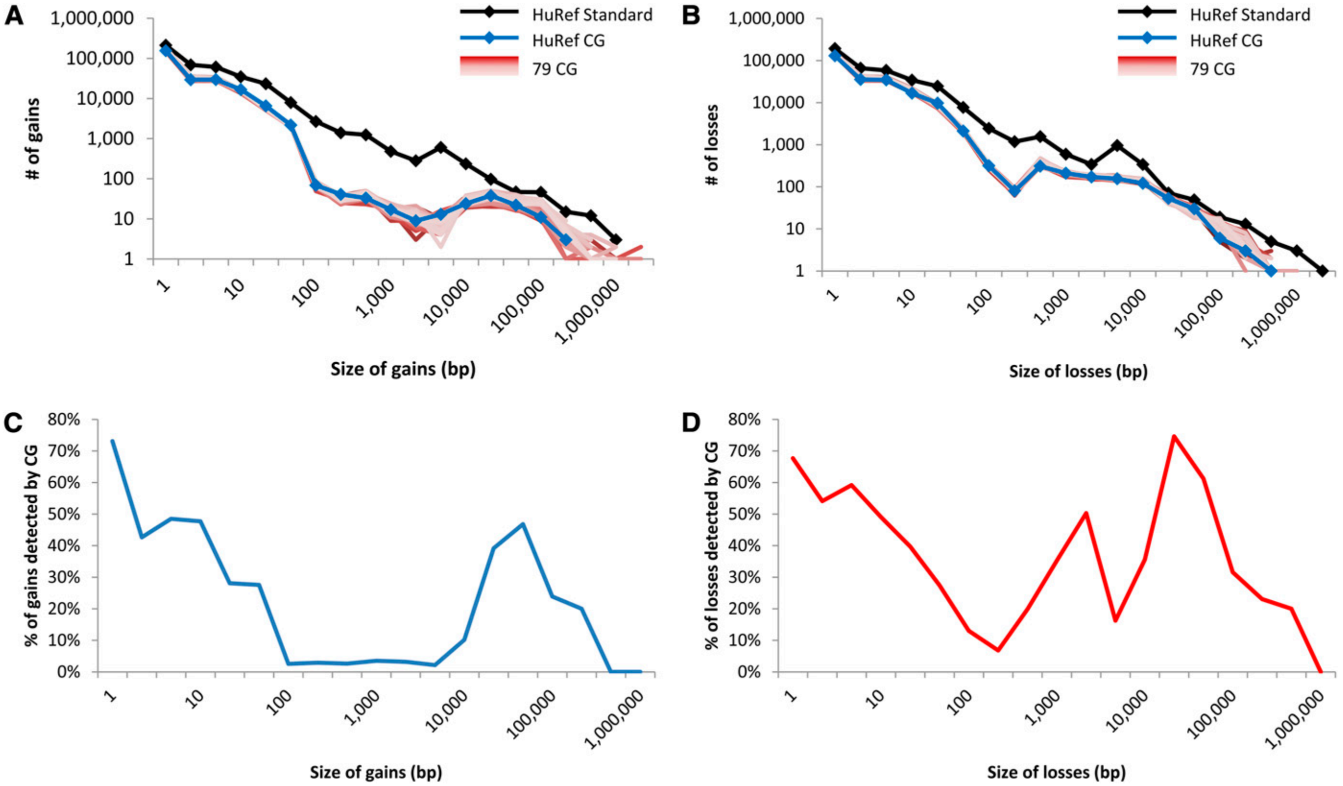

Figure 1 Variation distribution of genomes sequenced. The size distribution of nonredundant (A) gains and (B) losses detected in the HuRef and 79 other samples. The proportion of nonredundant (C) gains and (D) losses detected in HuRef by CG in comparison with HuRef Standard.

losses whose size ranged from 1 to 100 bp, 30% of that from 100 to 10 kb, and 43% of that from >10 kb; for gains, HuRef CG detected 59% of that in HuRef Standard gains with the size ranged from 1 to 100 bp but only 7% of that from 100 to 10 kb and 21% of that from >10kb (Figure 1, C and D). CG used three primary approaches to detect gains and losses: split-read, paired-end and read depth (File S2 and Table S3). Unlike the uniform negative slope of the size distribution of variants annotated in the Sanger-based HuRef Standard (Figure 1, A and B), there were notable declines in sensitivity in the CG version, particularly for gains in the paired-end detection range, which spanned from 100 bp to 10 kb (Figure 2). As acknowledged by CG (Support & Community webpage), the paired-end detection approach had difficulty in calling variants at high-identity repeats, and calling novel insertion sequences relative to the National Center for Biotechnology Information reference.

To estimate false negatives in the CG profiles, we generated a compilation of variation from published studies (File S2, Figure S2, and Table S4). We identified a set of high confidence calls in the HuRef sample, by identifying HuRef Standard variants that were also detected in the population reference. We then examined the size distribution curves of HuRef CG variants against the curves representing the HuRef Standard variants also detected in the population reference, and we found that the HuRef CG curves were consistently below the curves of confirmed HuRef Standard. This analysis shows that there were variants missing in the HuRef CG profile; undercalling of gains greater than 100 bp was particularly severe (Figure S3). However, we emphasize that other short-read sequencing technologies also have similar problems, with large gains missing (Figure S1).

When comparing the HuRef CG data to the HuRef Standard, we determined that some of the missing gains and losses were from regions containing repeats. We found a notable reduction of calls in loci with retrotransposable repeats, tandem repeats and segmental duplications (two-tailed χ^2 test; $P < 2.2e\text{-}16$) (Figure S4, A and B and C and D). It is difficult to align HTS reads to tandem repeat loci whose length can be longer than the short reads, and consequently, variant-detection at these loci is hampered. Similarly, short inserts can prevent aligning and assembling of paired reads to regions with retrotranposons and segmental duplications. These observations highlight the importance of having long reads and inserts for alignment and variant calling. As for centromeric and telomeric repeats, both Sanger sequencing and HTS have difficulty with these locations.

We evaluated false-positive results in the HuRef CG profile by comparing this data set to both the HuRef Standard and the profiles from the other 79 CG-sequenced genomes in this study, and we conservatively estimated that 11.4% of the HuRef CG gains and 3.9% of the losses could be false (File S2, Figure S5, and Table S5). Again, detection of gains was worse than losses.

From our comparison of the HuRef CG and HuRef Standard datasets, we observed that CG also had notable strengths. First, the HuRef CG loss size distribution was fairly uniform when compared to the expected HuRef Standard (Figure 1B). Second, CG was highly precise in determining variant size, with the exception of overcalling by the read-depth approach (Figure S6). Increasing the sequence coverage plus decreasing the bin-size may reduce this overestimation. Finally, the HuRef CG variant profiles were similar to the profiles of the other 79 CG genomes, highlighting consistency across experiments (File S2 and Figure 1, A and B).

Taking advantage of the availability of a comprehensive set of variation from a fully sequenced genome, we have analyzed the performance of detecting insertion and deletion by a HTS technology. Overall, we conclude that only a fraction of kown variation was captured, with notable shortcomings in detecting insertions and

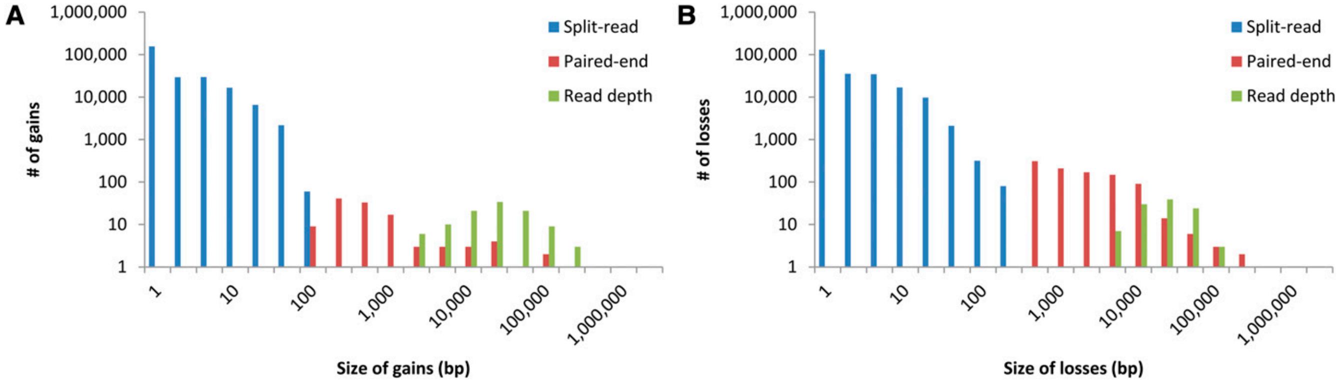

Figure 2 Size distribution of HuRef CG gains and losses detected by each discovery strategy examined: split-read, paired-end mapping and read depth. (A) Gains. (B) Losses.

duplications in the 100-bp to 10-kb size range, and at repetitive DNA sequences. Many of these deficiencies are associated with short reads and insert lengths (File S2, Figure 1, A and B, Figure S4 and Figure S7, Table S6 and Table S7). Generating longer reads (Loomis *et al.* 2013) or libraries of multiple insert lengths can mitigate these shortcomings. Greater depth of coverage can also partially recover some of the missing calls. Among our 80 CG-sequenced samples (File S2, Figure S8 and Figure S9), we noticed that the sequenced-depth and the number of variants reported were positively correlated (gains: R = 0.36, P = 0.00097; losses: R = 0.41, P = 0.00017; Figure S10). Computationally, one should continue to apply multiple complementary variant detection strategies: split-read, paired-end, read depth, and one-end-anchor approaches (Hajirasouliha *et al.* 2010). Moreover, whole-genome assembly comparison approach should be considered (Khaja *et al.* 2006; Levy *et al.* 2007), as our analysis has shown that this approach can yield the greatest number, type and size range of variation (Table S3). However, current *de novo* assembly of short sequences is often restricted by the presence of repeats. A possible solution is a hybrid assembly constructed with a mixture of shallow coverage (~5×) of mate-pair long-reads with deeper coverage (~25×) of paired-end short-reads (Schatz *et al.* 2010; Gnerre *et al.* 2011). Alternatively, sequencing can be performed in conjunction with microarray or single-molecule physical mapping (Lam *et al.* 2012a) to detect larger variation. Physical mapping or other complexity-reduction processes [*e.g.*, Long Fragment Read (Peters *et al.* 2012)] should improve alignment and the accuracy of variant discovery. Finally, some common variants (minor allele frequency >5%) that are missed by HTS could be imputed by nearby tag SNPs, although it may not be applicable to some rare variants as it has been shown that ~20% of biallelic CNVs cannot be readily captured (Mills *et al.* 2011). Ultimately, if HTS is to become a primary technology in clinical laboratories it will further benefit from improvement, particularly in capturing rare indels, CNVs and more complex rearrangements that are associated with diseases.

ACKNOWLEDGMENTS

We thank Ewen Kirkness for input regarding Complete Genomics variants characteristics, and Lynette Lau for computational analysis. This work was funded by grants from the University of Toronto McLaughlin Centre, Genome Canada and the Canadian Institutes for Health Research (CIHR). Stephen W. Scherer holds the GlaxoSmithKline-CIHR Chair in Genome Sciences at the University of Toronto and The Hospital for Sick Children. We would like to thank the J. Craig Venter Family Foundation, La Jolla, CA for funds for the HuRef sequencing. There is no conflict of interest to declare for this study.

LITERATURE CITED

Drmanac, R., A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns *et al.*, 2010 Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78–81.

Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108: 1513–1518.

Hajirasouliha, I., F. Hormozdiari, C. Alkan, J. M. Kidd, I. Birol *et al.*, 2010 Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26: 1277–1283.

Khaja, R., J. Zhang, J. R. MacDonald, Y. He, A. M. Joseph-George *et al.*, 2006 Genome assembly comparison identifies structural variants in the human genome. Nat. Genet. 38: 1413–1418.

Lam, E. T., A. Hastie, C. Lin, D. Ehrlich, S. K. Das *et al.*, 2012a Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. 30: 771–776.

Lam, H. Y., M. J. Clark, R. Chen, G. Natsoulis, M. O'Huallachain *et al.*, 2012b Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. 30: 78–82.

Lee, C., and S. W. Scherer, 2010 The clinical context of copy number variation in the human genome. Expert Rev. Mol. Med. 12: e8.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern *et al.*, 2007 The diploid genome sequence of an individual human. PLoS Biol. 5: e254.

Loomis, E. W., J. S. Eid, P. Peluso, J. Yin, L. Hickey *et al.*, 2013 Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 23: 121–128.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen *et al.*, 2011 Mapping copy number variation by population-scale genome sequencing. Nature 470: 59–65.

Pang, A. W., J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq *et al.*, 2010 Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11: R52.

Peters, B. A., B. G. Kermani, A. B. Sparks, O. Alferov, P. Hong *et al.*, 2012 Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487: 190–195.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg, 2010 Assembly of large genomes using second-generation sequencing. Genome Res. 20: 1165–1173.

Weischenfeldt, J., O. Symmons, F. Spitz, and J. O. Korbel, 2013 Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. 14: 125–138.

Communicating editor: F. P. Roth