CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Infoscience - École polytechnique fédérale de Lausanne

© 2010 Nature America, Inc. All rights reserved.

Primer-initiated sequence synthesis to detect and assemble structural variants

To the Editor: Structural variants constitute the largest portion of nucleotide variation in genomes, yet their comprehensive characterization based on high-throughput sequencing technologies is still challenging[1]. Here we present primer-initiated sequence synthesis for genomes (PrInSeS-G), a software tool that detects and assembles sequence variants (1 base pair (bp) to ~10 kilobases (kb)) from single- or paired-end short reads. PrInSeS-G first aligns the reads to a reference genome using Maq[2], after which it targets regions that have a fluctuation in coverage, which may be indicative of sequence variation, and performs local sequence assembly across the affected regions. Assembly is seeded by a short fragment of the reference sequence preceding the low-coverage region, the 'primer' (**Fig. 1a**). The primer is extended by one base at a time using overlapping reads until it reaches a predefined fragment of the reference sequence termed 'terminator'. The use of these short reference fragments allows the direct mapping of the assembled 'contig'. Thus, unlike current structural-variant mappers, PrInSeS-G simultaneously assembles and maps sequence variation.

To evaluate the performance of PrInSeS-G, we simulated paired-end reads from a 1 megabase (Mb) region of human chromosome 21 in which we introduced 500 insertions and deletions (indels) of 5 bp to 10 kb. PrInSeS-G's performance was robust, but the false positive and false negative rates were, as we expected, affected by indel size and by the quality of the alignment profile, which itself is dependent on the read format and the overall coverage (**Supplementary Fig. 1** and **Supplementary Table 1**). Moreover, PrInSeS-G's performance compared favorably with that of the structural variation mapper Breakdancer[3] (**Supplementary Data** and **Supplementary Methods**) with the important benefit that PrInSeS-G yielded sequence information on the variants.

To evaluate the performance of PrInSeS-G on real data, we used single-ended reads from *Salmonella paratyphi* A AKU12601. First, we used the AKU12601 genome as reference to detect false positives. PrInSeS-G detected 74 non–single-nucleotide polymorphism (non-SNP) variants. For 54 of these, we did not obtain improved read depth after realigning the reads to the new consensus sequence, indicating that this validation approach is efficient at removing potential false positives. To estimate the true positive rate, we used the genome of a related *S. paratyphi* strain, ATCC9150, as reference template and found that PrInSeS-G assembled 68% of detectable variants

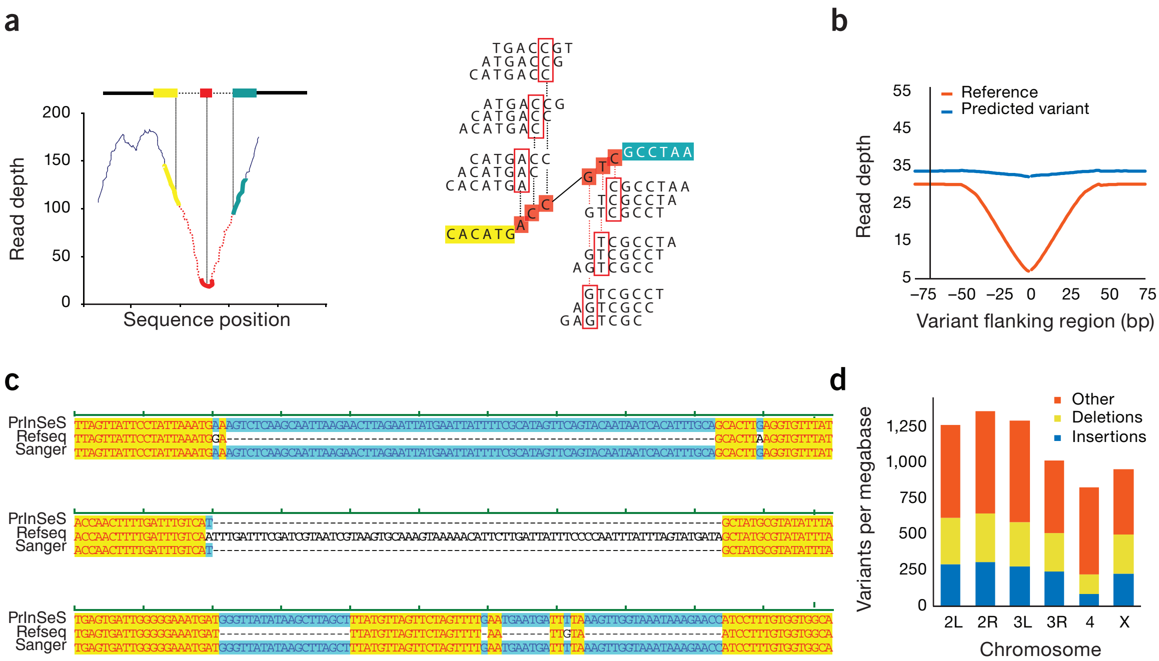

Figure 1 | PrInSeS-G principle and validation. (**a**) PrInSeS-G detects areas of low read coverage indicating putative variants (red). It attempts to *de novo* assemble this region by using a short fragment of the reference sequence as 'primer' (yellow) to align reads that partially overlap (red). The assembly continues until PrInSeS-G reaches a given 'terminator' sequence (cyan). This assembly process is independently performed in both directions. (**b**) Average read depth for the reference sequence and for the corresponding sequence that includes a predicted variant for all *D. melanogaster* RAL-304 regions that resulted in alignment improvement. (**c**) Examples of Sanger sequencing validations of variants detected in the *D. melanogaster* RAL-304 strain as compared to the reference genome sequence (top, insertion; middle, deletion; and bottom, substitution). (**d**) Summary of non-SNP variants per megabase per chromosome or chromosome arm, as indicated. 'Other' includes more complex variants such as substitutions.

(28 of 41; **Supplementary Data**, **Supplementary Methods** and **Supplementary Table 2**). To enable comparison with Breakdancer on real data, we performed a similar analysis using paired-end reads from *Mycobacterium tuberculosis* and, consistent with the simulation data, found that PrInSeS-G outperformed BreakDancer in terms of false positive and false negative rates (**Supplementary Data**, **Supplementary Table 3** and **Supplementary Methods**).

Next we used PrInSeS-G to analyze uncharacterized genomes. First, we analyzed single-end reads from *M. tuberculosis* strain 18b (**Supplementary Methods**) and detected 275 consensus non-SNP variants for which the read depth improved after realigning the reads to the new consensus sequence (**Supplementary Data**, **Supplementary Figs. 2** and **3**, and **Supplementary Table 4**). We then mapped structural variation in *Drosophila melanogaster* whole genomes using a combination of single- and paired-end read data from the *Drosophila* Genetic Reference Panel[4] with an average combined coverage of 31. Initially we identified 121,198 non-SNP variants; 93% of these resulted in improved read alignment (**Fig. 1b** and **Supplementary Fig. 4**). After selecting these and removing overlapping variants (**Supplementary Methods**), we reached a final 'consensus' list of 107,517 non-SNP variants up to ~10 kb, thereby generating, to our knowledge for the first time, a comprehensive catalog of naturally occurring *D. melanogaster* variants (**Supplementary Tables 5** and **6**). Sanger sequencing validation for selected variants confirmed 84% of these (26 out of 31) (**Fig. 1c**, **Supplementary Data** and **Supplementary Methods**). Our data were consistent with those obtained using microarray comparative genome hybridization[5,6] in that we found significantly fewer variants per megabase on the X chromomsome compared to the autosomal chromosomes (741 versus 928, respectively; *G* test (G) = 21.0, $P < 4.6 \times 10^{-6}$, degrees of freedom (d.f.) = 1; **Fig. 1d**) and in exons compared to nonexonic regions (223 versus 1,148, respectively; $G = 683$, $P < 2.2 \times 10^{-16}$, d.f. = 1; **Supplementary Fig. 5**). In addition, genes with structural variants had significantly more expression variation than those without variants (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; **Supplementary Fig. 6** and **Supplementary Table 7**), consistent with the notion that genes tend to be closely associated with variants that impact their expression[5].

We expect that imminent read length increases and future software development will ameliorate PrInSeS-G's overall variant detection sensitivity and specificity, and should eventually enable it to characterize heterozygous variants in mammalian whole genomes.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank the members of the Lausanne DNA array facility for performing the Illumina sequencing, N. Gheldof for help with the figures, J. Sacchettini for providing the *M. tuberculosis* H37Rv-JS read data, and the members of the Deplancke laboratory for critical reading of the manuscript. A substantial part of the data analysis was performed at the Vital-IT Center for high-performance computing of the Swiss Institute of Bioinformatics. This work was supported by funds from the Swiss National Science Foundation, Systems X, a Marie Curie International Reintegration grant (B.D.) and the Seventh Research Framework Programme, and by Institutional support from the Ecole Polytechnique Fédérale de Lausanne.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Andreas Massouras[1,5], Korneel Hens[1,5], Carine Gubelmann[1], Swapna Uplekar[2], Frederik Decouttere[3], Jacques Rougemont[4], Stewart T Cole[2] & Bart Deplancke[1]

[1]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [2]Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [3]Genohm, Zwijnaarde, Belgium. [4]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland. [5]These authors contributed equally to this work.
e-mail: bart.deplancke@epfl.ch

PUBLISHED ONLINE 13 JUNE 2010; DOI:10.1038/NMETH.F.308

1. Medvedev, P., Stanciu, M. & Brudno, M. *Nat. Methods* **6**, S13–S20 (2009).
2. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008).
3. Chen, K. *et al. Nat. Methods* **6**, 677–681 (2009).
4. Ayroles, J.F. *et al. Nat. Genet.* **41**, 299–307 (2009).
5. Dopman, E.B. & Hartl, D.L. *Proc. Natl. Acad. Sci. USA* **104**, 19920–19925 (2007).
6. Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O. & Long, M. *Science* **320**, 1629–1631 (2008).

© 2010 Nature America, Inc. All rights reserved.