

Paired-end read length lower bounds for genome re-sequencing

Rayan Chikhi, Dominique Lavenier

► **To cite this version:**

Rayan Chikhi, Dominique Lavenier. Paired-end read length lower bounds for genome re-sequencing. BMC Bioinformatics, BioMed Central, 2009, 10.1186/1471-2105-10-S13-O2 . inria-00426856

HAL Id: inria-00426856

<https://hal.inria.fr/inria-00426856>

Submitted on 28 Oct 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Oral presentation

Open Access

Paired-end read length lower bounds for genome re-sequencing

Rayan Chikhi and Dominique Lavenier

Address: ENS Cachan/IRISA, Campus de Beaulieu, 35042 Rennes, France

from Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

BMC Bioinformatics 2009, 10(Suppl 13):O2 doi: 10.1186/1471-2105-10-S13-O2

This article is available from: <http://www.biomedcentral.com/1471-2105/10/S13/O2>

© 2009 Chikhi and Lavenier; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

Next-generation sequencing technology is enabling massive production of high-quality paired-end reads. Many platforms (Illumina Genome Analyzer, Applied Biosystems SOLID, Helicos HeliScope) are currently able to produce "ultra-short" paired reads of lengths starting at 25 nt. An analysis by Whiteford et al. [1] on sequencing using unpaired reads shows that ultra-short reads theoretically allow whole genome re-sequencing and *de novo* assembly of only small eukaryotic genomes. Chaisson, Brinza and Pevzner [2] recently determined that the paired read length threshold for *de novo* assembly of the *E. coli* genome is ≈ 35 nt, and ≈ 60 nt for the *S. cerevisiae* genome. The latter read length is unfeasible for some next-generation technologies. By conducting an analysis extending Whiteford et al. results, we investigate to what extent genome re-sequencing is feasible with ultra-short paired reads. We obtain theoretical read length lower bounds for re-sequencing that are also applicable to paired-end *de novo* assembly.

Methods

A novel algorithm that utilizes a suffix array has been specifically designed to compute the uniqueness of paired reads with fixed or variable mate-pair distance. The algorithm is a non-trivial extension of the RepAnalyze algorithm [3] to paired reads. Bacterial and eukaryotic genomes are analyzed to determine the uniqueness of paired reads given a fixed mate-pair distance of 300 nt.

Longer mate-pair distances with high variability are also considered for the *E. coli* genome.

Discussion

Simulation results indicate that 97.4% of the *E. coli* genome is covered with unique paired reads of length 8 nt, and 90% of the *H. sapiens* genome is covered with unique paired reads of length 11 nt (see Figure 1). These results suggest that for large genomes, re-sequencing requires significantly shorter (for *H. sapiens*, at least 67% shorter) paired reads to achieve coverage comparable to unpaired reads. Moreover, a trade-off exists between read length and mate-pair distance: given

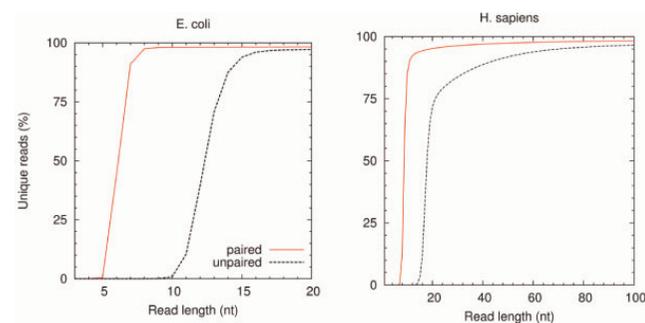


Figure 1
Percentage of unique paired and unpaired reads as a function of read length for the *E. coli* and *H. sapiens* genomes. Paired uniqueness is computed with a mate-pair distance of 300 nt.

a fixed mate-pair distance of 5,000 nt (resp. 2,000 nt), the whole *E. coli* genome can be unambiguously probed by paired reads of length above 18 nt (resp. 700 nt). When the uncertainty in mate-pair distance is $\pm 10\%$, only a small part of the genome cannot be uniquely probed (resp. 0.3% and 0.1% in the previous cases).

References

1. Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, Bradley M and Neylon C: **An analysis of the feasibility of short read sequencing.** *Nucleic Acids Research* 2005, **33(19)**: e171.
2. Chaisson MJ, Brinza D and Pevzner PA: **de novo fragment assembly with short mate-paired reads: Does the read length matter?** *Genome Research* 2009, **19(2)**:336–346.
3. Whiteford N: **String Matching in DNA Sequences: Implications for Short Read Sequencing and Repeat Visualisation.** PhD thesis, University of Southampton; 2007.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

