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

Rayan Chikhi, Dominique Lavenier

BMC Bioinformatics

BioMed Central

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

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 RepAnalyse 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

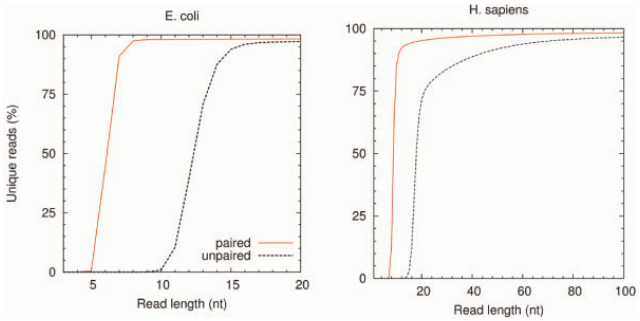

Figure 1
Percentage of unique paired and unpaired reads as a function of read length for the *E. coli* and *H. sapiens* genomes. Paired uniqueness is computed with a mate-pair distance of 300 nt.

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 ± 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.