Published online before print March 10, 2008
Genome Research, DOI: 10.1101/gr.072033.107
ACCEPTED PREPRINT

Methods and Resources

De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer

David Hernandez[1,3], Patrice Francois[1], Laurent Farinelli[2], Magne Osteras[2], and Jacques Schrenzel[1]

[1] *Geneva University Hospitals;* [2] *Fasteris SA*

Novel high-throughput DNA sequencing technologies allow researchers to characterize a bacterial genome during a single experiment and at a moderate cost. However, the increase in sequencing throughput that is allowed by using such platforms is obtained at the expense of individual sequence read length, which must be assembled into longer contigs to be exploitable. This study focuses on the Illumina sequencing platform that produces millions of very short sequences that are 35 bases in length. We propose a de novo assembler software that is dedicated to process such data. Based on a classical overlap graph representation and on the detection of potentially spurious reads, our software generates a set of accurate contigs of several kilobases that cover most of the bacterial genome. The assembly results were validated by comparing datasets that were obtained experimentally for *Staphylococcus aureus* strain MW2 and *Helicobacter acinonychis* strain Sheeba with that of their published genomes acquired by conventional sequencing of 1.5 - 3.0 kb fragments. We also provide indications that the broad coverage achieved by high throughput sequencing might allow for the detection of clonal polymorphisms in the set of DNA molecules being sequenced.

Correspondence: [3] E-mail: david.hernandez@genomic.ch