BMC Bioinformatics

BioMed Central

Meeting abstract

Open Access

Flipping NextGen: using biological systems to characterize NextGen sequencing technologies

Jarret Glasscock*, Ryan Richt and Matt Hickenbotham

Address: Cofactor Genomics, St. Louis, MO 63103, USA

Email: Jarret Glasscock* - Jarret_Glasscock@cofactorgenomics.com

* Corresponding author

from UT-ORNL-KBRIN Bioinformatics Summit 2009
Pikeville, TN, USA. 20–22 March 2009

Published: 25 June 2009

BMC Bioinformatics 2009, **10**(Suppl 7):A18 doi:10.1186/1471-2105-10-S7-A18

This abstract is available from: http://www.biomedcentral.com/1471-2105/10/S7/A18

Background

At a current 12 gigabases per sequencing run (and growing), there have been significant advancements in DNA sequencing technologies resulting in next generation (NextGen) sequencing platforms that produce 5 orders of magnitude more data than platforms used for the human genome project.

Results

A broad range of genomes was surveyed in order to assess characteristics necessary to sufficiently analyze these biological systems. In the context of genome re-sequencing projects we found 15 bp was needed to uniquely map 98% of loci in many bacteria, while 20 bp was needed before hitting lower asymptotes to uniquely characterize a fraction of more complex genomes (Figure 1).

Transcriptomes on the other hand were much less variable and required fewer bases (x) to uniquely map a much larger percentage (y) of their sequence space. For example, more than 98% of the complex human transcriptome could be uniquely characterized with as few as 20 bp.

Finally, de-novo sequencing (i.e. without a reference) would require a minimum of 1/2 of the sequence length to be unique in order to allow sufficient contig extension in the assembly process. For example, 40–50 bp reads are necessary for de-novo characterization of these systems uniquely defined by 20–25 bp reads. As of 2009, short read NextGen sequencing technologies have moved to 50 bp and beyond, ushering in what is expected to be the start of a revolution in genomics.

Conclusion

These results establish a lower bound on sequence length (x) required to sufficiently conduct re-sequencing, transcriptome, and de-novo sequencing projects. The asymptotic nature of the results also provides a guide for what percentage of the total space (y) we might expect to define in genomes/transcriptomes of similar size and complexity.

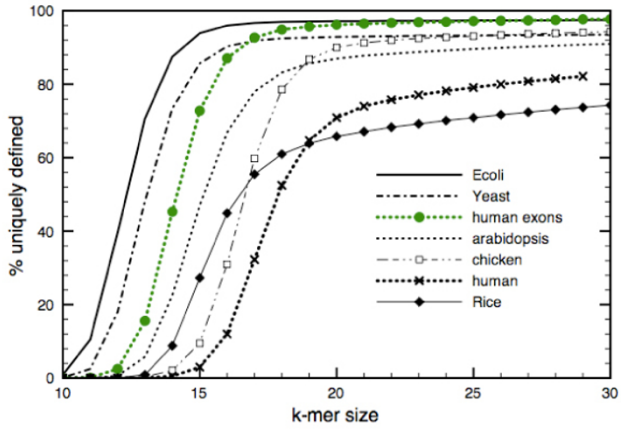

Figure 1
Percent of genome or transcriptome (Y-axis) uniquely defined by a read length (X-axis).