High-quality draft assemblies of large and small genomes from massively parallel DNA sequence data

Dariusz Przybylski, Iain MacCallum, Sante Gnerre, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, Giles Hall, Terrance P. Shea, Sean Sykes, Aaron M. Berlin, Daniel Aird, Maura Costello, Riza Daza, Louise Williams, Robert Nicol, Andreas Gnirke, Chad Nusbaum, Eric S. Lander, David B. Jaffe

Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge MA 02142
dariusz@broadinstitute.org

Massively parallel sequencing (MPS) technologies are revolutionizing genomics by making it possible to generate billions of relatively short (~100 base) sequence reads at very low cost. While such data can be readily used for a wide range of biomedical applications, it has proven difficult to use them to generate high-quality *de novo* assemblies of large, repeat-rich vertebrate genomes. To date, the genome assemblies generated from such data have fallen far short of those obtained with the older (but 1000 times more expensive) capillary-based sequencing approach.

We report the development of a new genome assembly algorithm, ALLPATHS-LG, and its application to MPS data from fifteen vertebrate genomes. The resulting draft genome assemblies have good accuracy, short-range contiguity, long-range connectivity and coverage of the genome. In particular, the base accuracy is high (≥ 99.95%) and the scaffold sizes (*e.g.* N50 size = 11.5 Mb for human and 17.4 Mb for mouse) are similar to those obtained with capillary-based sequencing.

While high-quality assembly of large genomes remains a key challenge of the field, in fact the assembly of small genomes is often challenging, and presently limited by defects in amplification-based MPS data, including read length and uneven coverage. Unamplified single-molecule sequencing data (having complementary properties) can now be generated on the Pacific Biosciences platform. At current yields, this is highly practical for small genomes, for which sample prep costs dominate. Using a modified version of ALLPATHS-LG, we demonstrate hybrid (Illumina plus Pacific Biosciences) assemblies of bacterial genomes. These assemblies are much better than the Illumina-only assemblies of the same genomes. In fact they close nearly all small gaps.

The ALLPATHS-LG program is available at http://www.broadinstitute.org/software/allpaths-lg/blog.