Assembling whole genomes using Illumina mate pair reads

Ole Schulz-Trieglaff[1,2], Matthew Hims[1], Niall Gormley[1], Geoffrey Smith[1] and Dirk Evers[1]

Several recent publications demonstrated the versatility of DNA reads from next-generation sequencing platforms for the de-novo assembly of large and complex genomes. Examples include the Human pan genome [3], giant panda [4], cucumber [2], strawberry [5] and cacoa [1].

There is also the 10K Genome project which has the aim to de-novo assemble a genomic zoo, a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. Nevertheless, whole genome de-novo assembly is still a computationally demanding task and a challenging theoretical problem. There is thus a pressing need for methods to make fast and low-cost de-novo assemblies feasible.

In this poster, we focus on sequencing reads from the Illumina sequencing platforms such as HiSeq 2000 and Genome Analyzer II. They nowadays generate billions of reads from a single run with a length of 100bp and more and insert sizes of up to 10Kb.

We present recent advances in informatics and chemistry for de-novo assembly: progress in sample preparation, sequencing chemistry and informatics allow us to sequence and assemble even difficult bacterial genomes using a single sample preparation and a single sequencing run. Furthermore, long insert mate pairs and short inserts at high coverage allow us to generate high-quality draft assemblies of complex, mammalian-sized genomes.

Finally, we present assemblies of several large genomes and give recommendations for data preprocessing, quality control of the assemblies and how comparisons of whole-genome assemblies from unknown organisms can be used to obtain new biological insights.

References

[1] Xavier Argout et al. The genome of theobroma cacao. *Nature Genetics*, 43(2):101–108, February 2011.

[2] Sanwen Huang et al. The genome of the cucumber, cucumis sativus l. *Nature Genetics*, 41(12):1275–1281, December 2009.

[3] Ruiqiang Li et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, February 2010.

[4] Ruiqiang Li et al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7284):1106–1106, February 2010.

[5] Vladimir Shulaev et al. The genome of woodland strawberry (fragaria vesca). *Nature Genetics*, 43(2):109–116, February 2011.

[1]Illumina Cambridge, Chesterford Research Park, Cambridge CB10 1XL, United Kingdom.
[2]Corresponding author: oschulz-trieglaff@illumina.com