Cynomolgus and Chinese rhesus macaque genome assembly and analysis

Guangmei Yan (ygm@mail.sysu.edu.cn)
The South China Center for Innovative Pharmaceuticals, Guangzhou 510663, China
Jun Wang (wangj@genomics.org.cn)
Beijing Genomics Institute, Shenzhen
Xiaoning Wang (xnwang@scut.edu.cn)
The South China Center for Innovative Pharmaceuticals, Guangzhou 510663, China
Jian Wang (wangjian@genomics.org.cn)
Beijing Genomics Institute, Shenzhen
Zhiyong Huang
Beijing Genomics Institute, Shenzhen
Guojie Zhang
Beijing Genomics Institute, Shenzhen
Xiaodong Fang
Cai Li
Fei Ling

Method Article

Keywords: cynomolgus,Chinese rhesus macaque,macaca,assembly,analyse

Posted Date: November 4th, 2011

DOI: https://doi.org/10.1038/protex.2011.264

Abstract

cynomolgus and Chinese rhesus macaque sequencing, assembly and analyse

Procedure

1. SOAP denovo assembly SOAPdenovo employs the de Bruijn graph algorithm in order both to simplify the task of assembly and to reduce computational complexity. Low quality reads were filtered and potential sequencing errors were removed by k-mer frequency-based error correction. We filtered the following type of reads: 1. Reads having an ‘N’ over 10% of its length. 2. Reads from short insert-size libraries having more than 65% of bases with quality ≤7, and reads from large insert-size libraries that contained more than 80% of bases with a quality ≤ 7. 3. Reads with more than 10 bp from the adapter sequence \(allowing ≤2 bp mismatches). 4. Small insert size paired-end reads that overlapped ≥10 bp between two ends. 5. Read1 and read2 of two paired-end reads that were completely identical \(and were hence considered to be the products of PCR duplication). 6. Reads having a k-mer frequency <4 \(to minimize the influence of sequencing errors). SOAPdenovo first constructs the _de Bruijn_ graph by splitting the reads from short insert size libraries \(200-500bp) into 31-mers and then merging the 31-mers \(30bp overlaps with 1 bp overhangs); contigs were then collected which exhibited unambiguous connections in _de Bruijn_ graphs. Reads from mate-paired libraries \(insert size >2k) were aligned onto the contigs for scaffold building using the paired-end information. This paired-end information was subsequently used to link contigs into scaffolds, step by step, from short insert sizes to long insert sizes. **2. RNA-seq sequencing** 1| Homogenise frozen tissues in Trizol reagent in a bead mill with 5mm stainless steel beads. 2| Follow the Trizol procedure, including two alcohol precipitations and suspension of the final RNA pellet in RNAse-free water. 3| Construct RNA sequencing libraries using an Illumina standard mRNA-Seq Prep Kit. Briefly: Use oligo\(dT) magnetic beads to purify the poly-A containing mRNA molecules. Further fragment the mRNA into short lengths by controlled temperature, and then randomly primed during first strand synthesis by reverse transcription. Follow this with second-strand synthesis with DNA polymerase I to create double-stranded cDNA fragments. Subject double stranded cDNA to end repair by Klenow and T4 DNA polymerases and A-tailed by Klenow lacking exonuclease activity. 4| Ligation to Illumina Paired-End Sequencing adapters, size selection by gel electrophoresis and then PCR amplification complete the library preparation. Sequence the paired-end libraries sequenced on a Illumina Genome Analyzer for 100 bp at each end. **3. Gene prediction** use BLAT to map genes of IR \(MMUL_0_1) and human \(Ensembl release-56) onto two macaca genome, Orthologous regions were then determined by best-BLAT hit and synteny-based analysis, followed by the application of "Exonerate":http://www.ebi.ac.uk/~guy/exonerate/ and "GENEWISE":http://www.ebi.ac.uk/Tools/Wise2/index.html to refine gene model at each locus. **4. Assembly quality validation in neutral mode** Neutral InDel model1 can be used to validate the quality of our genome assemblies.When aligning two closely related genome sequences, the frequencies of lengths of successive alignment blocks \(which were split by gaps during the alignment), termed Inter-gap Segments \(IGS), may be expected to follow a geometric frequency distribution under a standard neutral

model.Within the neutral evolving regions, incorrect InDels introduced during the assembly process would result in the observed IGS length distribution departing from the geometric distribution. The introduced InDels would generate an excess of short IGS over the number predicted by the neutral InDel model. By quantifying this excess, several parameters viz. the proportion \(), average density \(D), and number \(Ng) of the clustered erroneous gaps in the genome alignments can be estimated.

References

1 Meader, S., Hillier, L. W., Locke, D., Ponting, C. P. & Lunter, G. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res 20, 675-684.