Title:
KILAPE: A novel pipeline for the assembly of complex genomes from next-generation sequencing data.

Authors:
Philipp Koch, Bryan R. Downie, Andreas Petzold, Niels Jahn, Marco Groth, Stefan Taudien, Kathrin Reichwald and Matthias Platzer

With the advent of next generation sequencing technologies (NGS), whole genome sequencing (WGS) can be conducted for much reduced prices and timescales. However, accurate assembly and scaffolding of complex genomes from NGS WGS data is a tremendous challenge.

In this light, we have developed the novel pipeline KILAPE (K-masking and Iterative Local Assembly of Paired Ends) as a universal, automated method to improve genome assemblies obtained by tools like Velvet, SOAPdenovo, CLC Assembly Cell or ALLPATHS-LG. It uses a less-repetitive fraction of reads to scaffold contigs and performs local assemblies to fill gaps in an iterative way.

Having applied our pipeline to human sequence data downloaded from the Short Read Archive, we improved a non-scaffolded ALLPATHS-LG assembly from 243,429 contigs, 2.62 Gb total and 24 kb N50 length to 13,215 scaffolds, 2.66 Gb and 609 kb, respectively.

Using a mixture of Illumina and Roche sequencing data, we have utilized CLC Assembly Cell and KILAPE to *de novo* assemble the ~1.6-1.9 Gb genome of the turquoise killifish *Nothobranchius furzeri*, an exceptionally short-lived species which over recent years emerged as a new vertebrate model for age research. The assembly contains 59,554 scaffolds with a total length of 1.5 Gb and a N50 scaffold size of 65 kb.

The most recent KILAPE benchmarks and quality control data on these assemblies and additional training and test data sets will be presented at the conference.