De novo assembly and annotation of the Candida cylindracea genome: a pipeline for rare organisms

Carvalho MJ[1], Pinheiro M[1], Santos MAS[1], Moura G[1]

universidade de aveiro · ibimed instituto de biomedicina

[1] Institute of Biomedicine - iBiMED, Department of Medical Sciences, University of Aveiro, 3810-193 Aveiro, Portugal

Introduction

The fungal CTG clade species are a group of diverse pathogenic and non-pathogenic fungi that translate the universal leucine CUG codon as serine. In previous studies, we have demonstrated that Ser-CUG translation erased the ancestral Leu-CUG codons from their genomes and that extant Ser-CUG codons evolved recently from serine codons or codons assigned to amino acids with chemical properties similar to serine. We have also demonstrated that these new Ser-CUG codons are used at low level and are present in non-conserved genes. However, the unconventional fungus *Candida cylindracea,* which is known to produce low-specificity lipases, uses the Ser-CUG codons at high level in almost all proteins, raising the puzzling questions of how such codons re-emerged in this genome in a short time span? And how did conserved proteins tolerate serines at CUG sites?
To answer these questions, we have sequenced the genome of *C. cylindracea* using both Illumina short read and Oxford Nanopore long read sequencing. The genome is diploid and posed significant challenges to the assembly and annotation pipelines that are publicly available. To overcome these difficulties, we are attempting several approaches. We present below the first data of the assembly and annotation of the *C. cylindracea* genome.

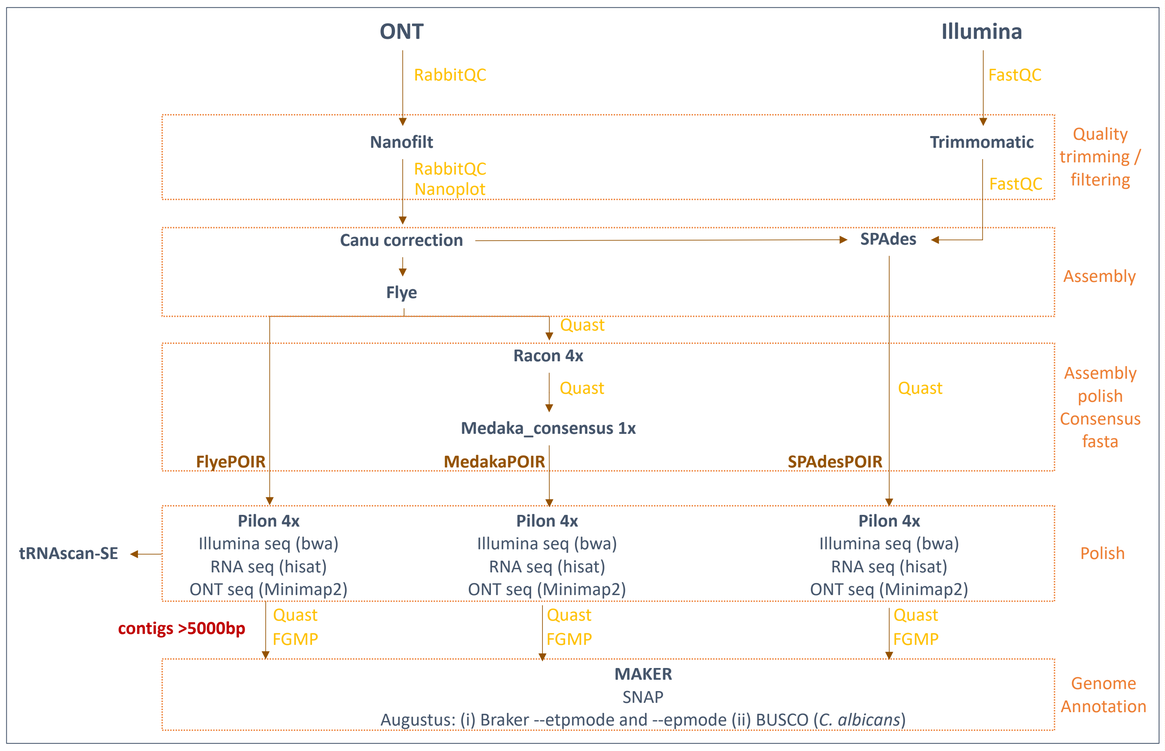
Methods

Genomic DNA from *Candida cylindracea* (ATCC 14830) was purified using the QIAGEN Genomic DNA Purification kit with Genomic-Tips 100/G.

For Oxford Nanopore Technologies (ONT) based sequencing, two genomic DNA libraries were prepared using the Ligation Sequencing Kit (SQK-LSK109), the NEBNext Nanopore companion module (E7180S), the NEB Blunt/TA Ligase Master Mix (M0367) and the NEBNext Quick Ligation Reaction Buffer (NEB B6058). Two different sequencing runs on the MinION using FLO-MIN106 R9.4.1 flow cells were performed. One using the MinKNOW 19.06.8, MinKNOW Core 3.4.8, Bream 4.1.9, Guppy 3.0.7 and the other with the MinKNOW 20.06.4, MinKNOW Core 4.0.4, Bream 6.0.7, Guppy 4.0.9. For both runs, the standard 48-hour run script with live fast basecalling model was used.

For Illumina based sequencing, the Illumina DNA prep (20018704) with Nextera DNA CD indexes (20018707) were used for genomic DNA library preparation. Paired-end sequencing was performed on an Illumina MiniSeq using the High Output chemistry to generate fragments of up to 151bp (300 cycles).

Bioinformatics analysis is shown in the workflow.

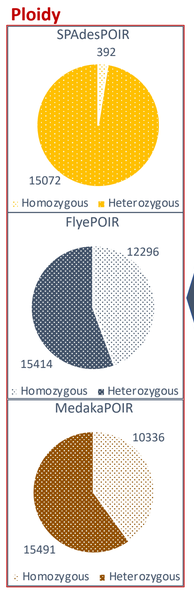


Results

Assembly QC

QUAST	SPAdesPOIR	FlyePOIR	MedakaPOIR
contigs (n)	80	21	16
≥0 bp	1208	21	16
≥5000 bp	44	20	16
≥10000 bp	38	16	15
Largest contig	1134783	2023231	2024649
Total length	10019537	10185511	10161698
≥5000 bp	9972521	10184522	10161698
≥10000 bp	9928166	10149910	10152188
≥50000 bp	9574546	10108262	10125511
N50	677169	1132984	1131684
N90	150578	326595	326644
L50	6	4	4
L90	19	9	9
GC (%)	63.05	63.03	63.06
N's	54037	0	0
FGMP			
UCEs (n)	27 out of 31	27 out of 31	27 out of 31
UCEs completion estimation	(87.1%)	(87.1%)	(87.1%)
Estimate completeness	92.1 (%)	87.9 (%)	87.9 (%)

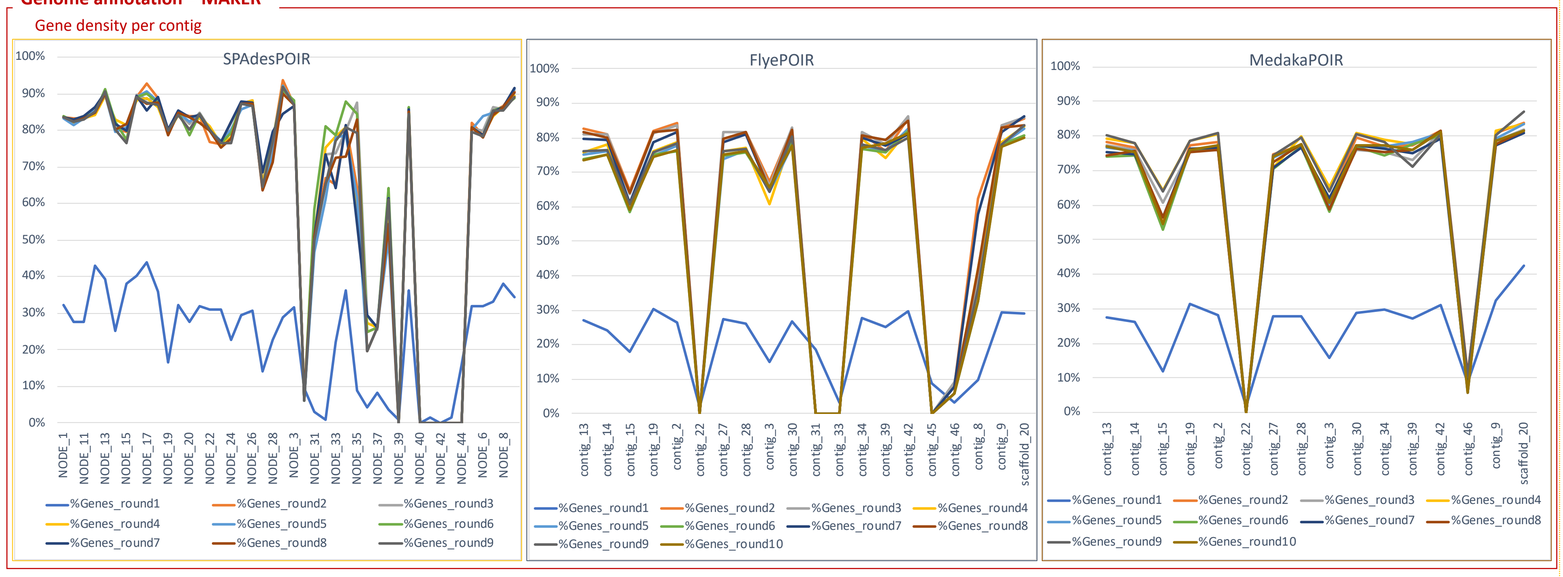
Ploidy



Ploidy was inferred by allele ratio obtained from variant calling (Illumina bam vs each assembly). Results indicate *C. cylindracea* as being diploid.

Genome annotation – MAKER

Gene density per contig

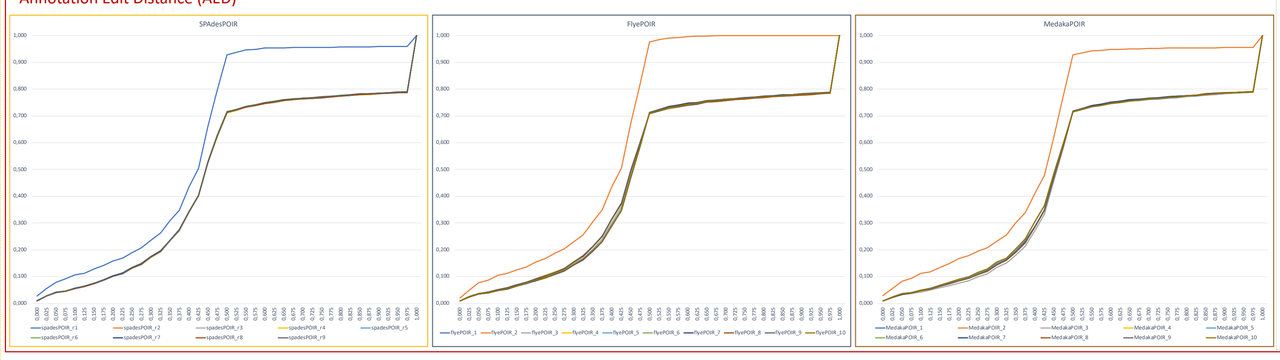


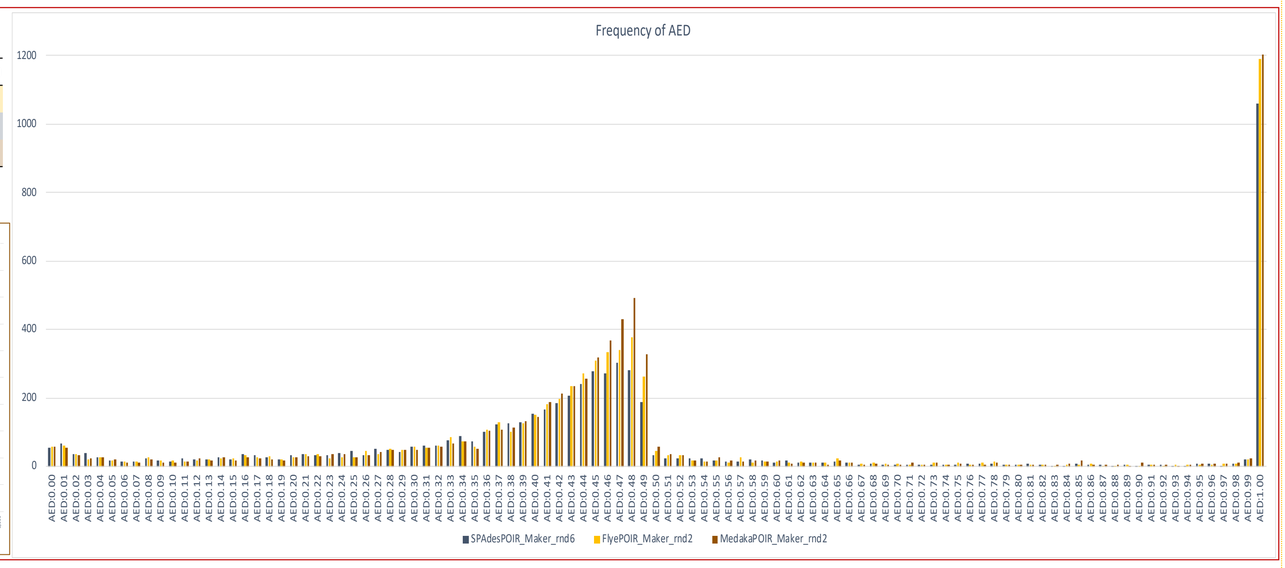
Genome annotation – MAKER

Number of genes identified per MAKER run

	MAKER_round1	MAKER_round2	MAKER_round3	MAKER_round4	MAKER_round5	MAKER_round6	MAKER_round7	MAKER_round8	MAKER_round9	MAKER_round10
SPAdesPOIR	2388	5873	5872	5842	5835	5786	5805	5829	5792	NA
FlyePOIR	2143	6284	5972	6120	6109	6039	5987	5914	6099	6071
MedakaPOIR	2332	6558	6153	6055	6072	6091	6097	6079	6001	6051

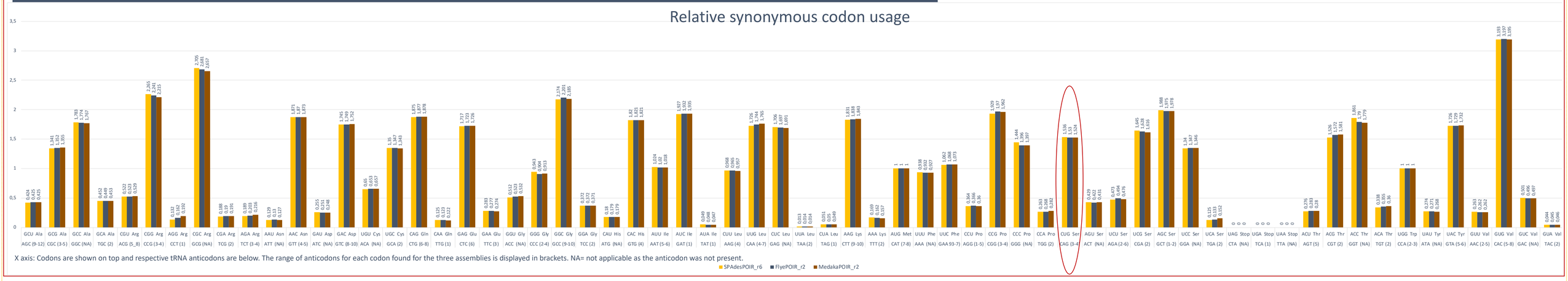
Annotation Edit Distance (AED)





tRNA characterisation – tRNAscan-SE; RSUC

The number of tRNAs found was 169 for SPAdesPOIR, 194 for FlyePOIR and 180 for MedakaPOIR. Per contig, 1-21 distinct tRNAs were found in SPAdesPOIR and 1-23 were found in both FlyePOIR and MedakaPOIR.



Concluding Remarks

- Results indicate *Candida cylindracea* is a diploid organism.
- The number of genes annotated by MAKER is in accordance to what is expected given the number of genes described for *Candida albicans* and *Saccharomyces cerevisiae* (~6000 genes).
- The quality of annotation is still low → Test other annotation tools? Use a reference species such as *Babjeviella inositovora* (closely related as assessed by BLAST) to refine Augustus gene prediction?
- The CUG codon is one of the serine codons mostly used.

CANCYL-POCI-01-0145-FEDER-031849 is supported by the Operational Program POCI, in its FEDER/FNR component, and the Foundation for Science and Technology, in its State Budget component. GenomePT POCI-01-0145-FEDER-022184; PBE POCI-O1-0145-FEDER-031238; iBiMED UIDP/04501/2020-Programático

FCT Fundação para a Ciência e a Tecnologia · COMPETE 2020 · PORTUGAL 2020 · UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional · GenomePT