**NORGAL TOOL**

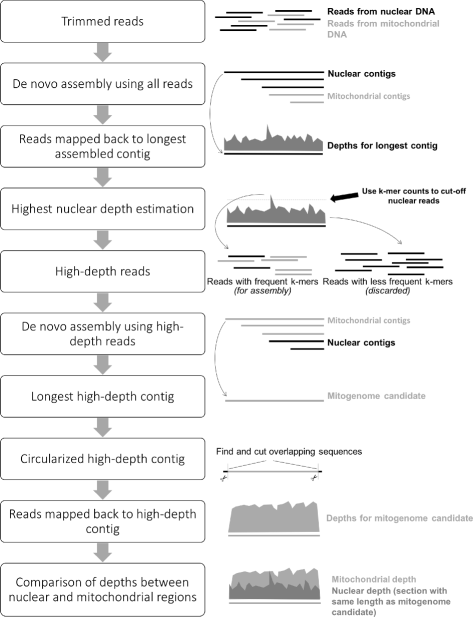
Norgal (de Novo ORGAneLle extractor) avoids this requirement by identifying a high frequency subset of k-mers that are predominantly of mitochondrial origin and performing a de novo assembly on a subset of reads that contains these k-mers.

Current methods to extract mtDNA from WGS data require a short seed sequence to initiate assembly.

The mitogenome will have a higher read depth compared to the nuclear genome and this difference in the read depth levels can be used to separate the reads into two groups.

The separation of the two types of reads is done by counting occurrences of subsequences of length k in the reads - k-mers - and classifying reads that have k-mers that are found more times than the nuclear read depth as being of non-nuclear origin.

Norgal uses raw short NGS reads from WGS data as input and outputs either a full or partial mitogenome.



1. Trim and remove adapters from NGS reads using *AdapterRemoval* and perform a de novo assembly using *MEGAHIT* with default settings and the k-mer range: 21, 49, 77 and 105. Norgal assumes that the longest assembled sequence (contig) is nuclear.
2. Map the reads back to the longest assembled sequence using *bwa mem* and calculate the read depths for each position, if the longest assembled contig is longer than 100,000 base pairs, only the first 100,000 base pairs are used as it should be enough to determine the depth. The read depths of the mapped reads to this contig are used to determine the nuclear depth threshold (ND threshold) which is defined as the mean of all non-zero read depths from the 25*th* to the 75*th* percentile range multiplied with five.
3. Count k-mers of size 31 in all reads and only keep a subset of reads that contains at least one 31-kmer with a frequency that is greater than the ND threshold. This is done using the program *BBTools*.
4. Perform a de novo assembly using *idba\_ud* with the reads containing the frequent k-mers and extract either the longest contig or optionally the longest contig with a predicted cytochrome c oxidase subunit 1 (COI) gene. The binned reads with high-frequency k-mers are used for an assembly with *idba\_ud* with default settings which does multiple assemblies with different k-mer sizes in the range: 20, 40, 60, 80 and 100. This second assembly only contains contigs that have a high read-depth of at least the ND threshold.
5. Examine circularity of the longest contig, determine read depth, identify potential mitochondrial and chloroplast contigs, and output plots comparing depths between this contig and the longest contig from the assembly in step 1. Another option is to select the longest contig that has the best hits to full RefSeq mitochondrial or pastid genomes. The extracted contig is tested for circularity by comparing the ends of the contig and finding overlaps. Any overlapping base pairs are cut and the final sequence is reported as a potential mtDNA candidate. Norgal outputs a graph with the read depths as well as the read depths of a section of the nuclear DNA. This graph with the two sets of read depths may be used for validation of the mtDNA candidate, so if the depths over the mtDNA candidate is around 10-100 higher than the depths over the nuclear region, it increases the evidence that the candidate is from the mitogenome.