**RepeatExplorer**

RepeatExplorer is a computational pipeline for discovery and characterization of repetitive sequences in eukaryotic genomes. The pipeline uses high-throughput genome sequencing data as an input and performs graph-based clustering analysis of sequence read similarities to identify repetitive elements within analysed samples. It should be noted that although the repeat identification algorithm generally works for any genome, some parts of the pipeline (e.g. protein domain-based classification of mobile elements) were primarily developed for application to plant genomics. However, there is a possibility to supply a custom repeat database to improve sensitivity in classification of non-plant repeats.

**Steps**

* **Getting your data to server**
* Datasets can be downloaded directly from the EBI Short Read Archive using **Get Data 🡪 EBI SRA** tool. Enter the ENA accession number in the search window, locate the corresponding dataset and select download link in the "Galaxy" column
* **Pre-processing of sequence reads**
* The clustering analysis requires a single file containing read sequences in FASTA format as an input.
* If such a file can be uploaded by the user, no pre-processing is required
* However, data obtained from sequencing facilities or downloaded from public archives are usually in FASTQ format combining nucleotide sequence information with sequencing quality scores.
* There is a number of programs for analysing and pre-processing raw sequence reads in **Tools 🡪 NGS: QC and manipulation 🡪 FastQC:Read QC**
* **Grooming the data**
* **The FASTQ illumine format must be converted to FASTQ sanger format using**  **Tools 🡪 NGS: QC and manipulation 🡪 FastQC:Read QC 🡪 FASTQ Groomer**
* **Trim the data**
* **If there is any discrepancy in the sequence then the sequence should be trimmed using Tools 🡪 NGS: QC and manipulation 🡪 FastQC:Read QC 🡪 FASTQ Trimmer**
* **FASTQ to FASTA**
* **The FASTQ format is converted into FASTA using Tools 🡪 NGS: QC and manipulation 🡪 FastQC:Read QC 🡪 FASTQ to FASTA converter**
* **TAREAN**
* Tandem repeat analyser (TAREAN), a computational pipeline which was built on the principles of graph-based repeat clustering, enhanced and supplemented with additional tools facilitating unsupervised identification and characterization of satellite repeats from unassembled sequence reads
* The pre-processed FASTA file is passed into **Tools 🡪 TAREAN 🡪 Tandem Repeat Analyzer**