**Supplemental material**

## Workflow for mapping reads to a reference

Illumina HiSeq2500 RNA-Seq sequences available in NCBI under accession number SRR5181508.

```
# downloading and converting raw files (reads)
fastq-dump SRR5181508

# downloading the GFF file of human genome
$ wget ftp://ftp.ensembl.org/pub/release-88/gtf/homo_sapiens/Homo_sapiens.GRCh38.88.gtf.gz
$ gzip -d Homo_sapiens.GRCh38.88.gtf.gz

# downloading the human chromosome 22
$ wget ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.
    chromosome.22.fa.gz
$ gzip -d Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz

# Filtering with sickle
# se => sigle sequences; --qual-type => type of quality; --output-file +> output file;
# -q => flag designates the minimum quality; -l => the minimum read length;
$ sickle se --fastq-file SRR5181508.fastq --qual-type sanger --output-file
    SRR5181508_FILTERED.fastq -q 30 -l 25

# Building the hisat2 index;
# -p number of processors
$ hisat2-build -p 4 Homo_sapiens.GRCh38.dna.chromosome.22.fa chromosome.22.hisat2.idx

# Mapping the filtered reads to the refrence;
# -p => number of processors; -q => input in fastq format; -S => output file
$ hisat2 -p 2 -x chromosome.22.hisat2.idx -q SRR5181508_FILTERED.fastq -S SRR5181508.sam

# Analisys
# Converting formats
$ samtools view -bS SRR5181508.sam > SRR5181508.bam

# Sorting data and converting
$ samtools sort -n SRR5181508.bam SRR5181508_SORTED.sn
$ samtools view -h -o SRR5181508_SORTED.sn.sam  SRR5181508.bam

# Counting reads in features
# -m => reads overlapping more than one feature; -s => strand-specific assay;
# -o output file; -a => skip reads with alignment quality lower than x (x = default: 10);
$ htseq-count -m intersection-nonempty -s no -a 10 SRR5181508_SORTED.sn.bam Homo_sapiens.
    GRCh38.88.gtf -o SRR5181508.count
```

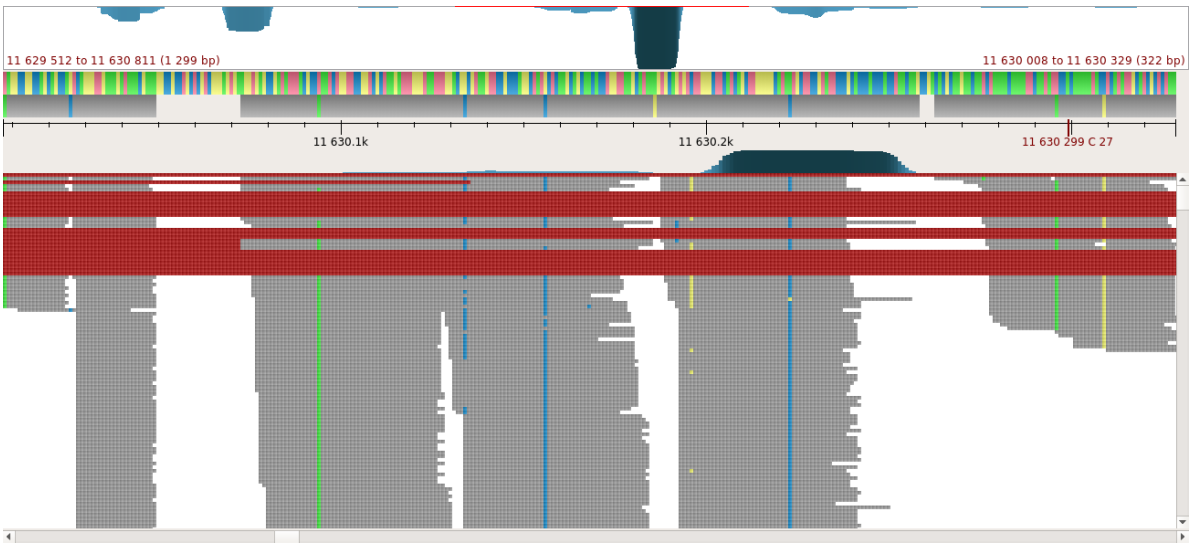



**Figure 8.** A partial view of the mapping results of Workflow 1 generated using Ugene[66].

## Workflow for *de novo* bacterial assembly

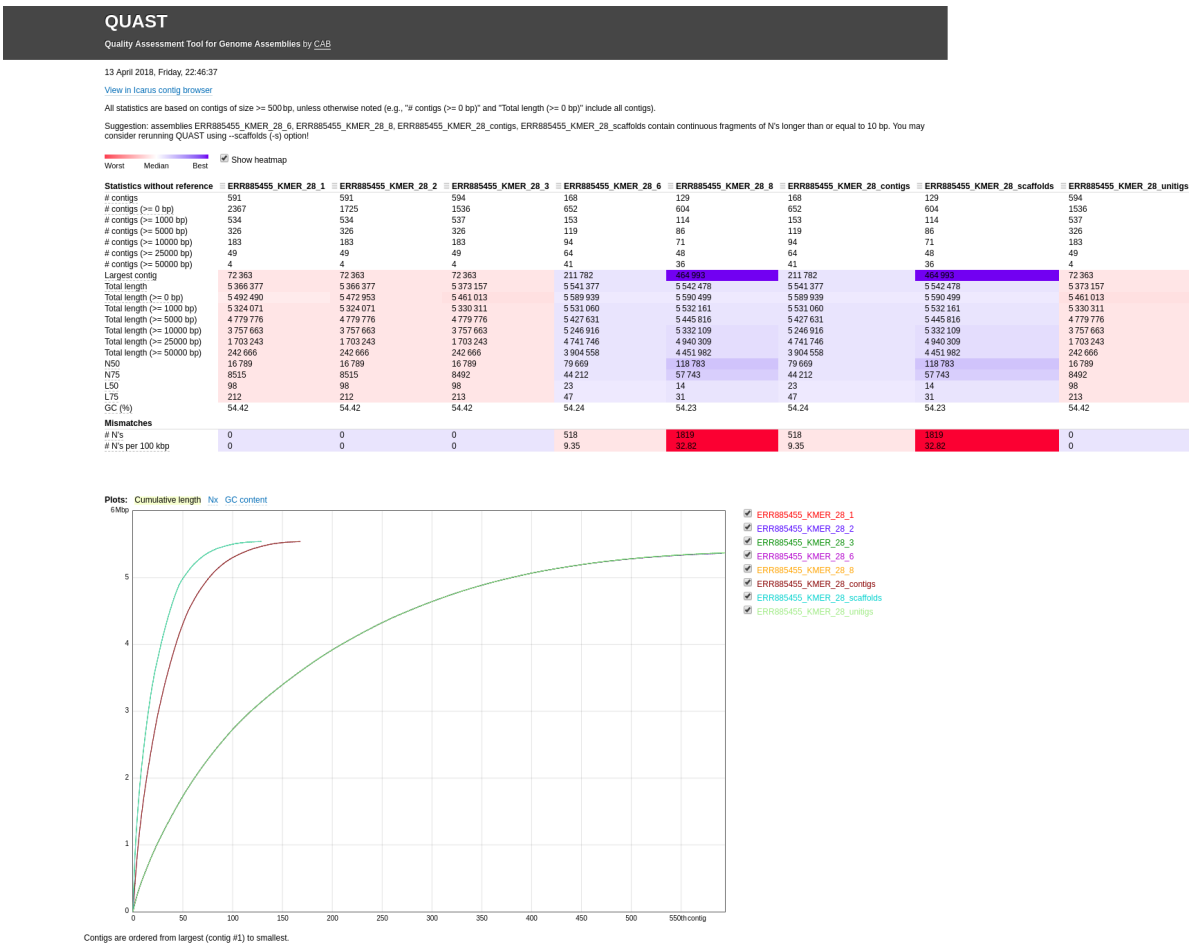
Illumina HiSeq 2000 paired end sequences available in European Nucleotide Archive (ENA) under accession number ERR885455

```
# download raw files (reads)
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR885/ERR885455/ERR885455_1.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR885/ERR885455/ERR885455_2.fastq.gz

# Filtering with Trimmomatic
$ java -jar /opt/Trimmomatic-0.36/trimmomatic-0.36.jar PE -phred33 ERR885455_1.fastq.
    gzERR885455_2.fastq.gz
ERR885455_1_FILTERED.fastq ERR885455_1_UNPAIRED.fastq ERR885455_2_FILTERED.fastq
    ERR885455_2_UNPAIRED.fastq ILLUMINACLIP:/opt/Trimmomatic-0.36/adapters/TruSeq3-PE.fa
    :2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 -validatePairs

# de novo assembly with ABySS
# abyss-pe => for paired-end reads; k => size of the kmer;
# np =>number of processors (depends on mpi modules installed);
# in => input file for forward/reverse trimmed reads; name => output file prefix
$ abyss-pe in='ERR885455_1_FILTERED.fastq ERR885455_2_FILTERED.fastq' k=28 name=
    ERR885455_KMER_28 np=4

# Statistical analisys with QUAST
$ python /opt/quast/quast.py -o ERR885455_stats ERR885455*.fa
```



**QUAST**

Quality Assessment Tool for Genome Assemblies by CAB

13 April 2018, Friday, 22:46:37

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Suggestion: assemblies ERR885455_KMER_28_6, ERR885455_KMER_28_8, ERR885455_KMER_28_contigs, ERR885455_KMER_28_scaffolds contain continuous fragments of N's longer than or equal to 10 bp. You may consider rerunning QUAST using --scaffolds (-s) option!

Worst Median Best — Show heatmap

| Statistics without reference | ERR885455_KMER_28_1 | ERR885455_KMER_28_2 | ERR885455_KMER_28_3 | ERR885455_KMER_28_6 | ERR885455_KMER_28_8 | ERR885455_KMER_28_contigs | ERR885455_KMER_28_scaffolds | ERR885455_KMER_28_unitigs |
|---|---|---|---|---|---|---|---|---|
| # contigs | 591 | 591 | 594 | 168 | 129 | 168 | 129 | 594 |
| # contigs (>= 0 bp) | 2367 | 1725 | 1536 | 652 | 604 | 652 | 604 | 1536 |
| # contigs (>= 1000 bp) | 534 | 534 | 537 | 153 | 114 | 153 | 114 | 537 |
| # contigs (>= 5000 bp) | 326 | 326 | 326 | 119 | 86 | 119 | 86 | 326 |
| # contigs (>= 10000 bp) | 183 | 183 | 183 | 94 | 71 | 94 | 71 | 183 |
| # contigs (>= 25000 bp) | 49 | 49 | 49 | 64 | 48 | 64 | 48 | 49 |
| # contigs (>= 50000 bp) | 4 | 4 | 4 | 41 | 36 | 41 | 36 | 4 |
| Largest contig | 72 363 | 72 363 | 72 363 | 211 782 | 254 992 | 211 782 | 254 992 | 72 363 |
| Total length | 5 366 377 | 5 366 377 | 5 373 157 | 5 541 377 | 5 542 478 | 5 541 377 | 5 542 478 | 5 373 157 |
| Total length (>= 0 bp) | 5 492 490 | 5 472 953 | 5 461 013 | 5 589 939 | 5 590 499 | 5 589 939 | 5 590 499 | 5 461 013 |
| Total length (>= 1000 bp) | 5 324 071 | 5 324 071 | 5 330 311 | 5 531 060 | 5 532 161 | 5 531 060 | 5 532 161 | 5 330 311 |
| Total length (>= 5000 bp) | 4 779 776 | 4 779 776 | 4 779 776 | 5 427 631 | 5 445 816 | 5 427 631 | 5 445 816 | 4 779 776 |
| Total length (>= 10000 bp) | 3 757 663 | 3 757 663 | 3 757 663 | 5 246 916 | 5 332 109 | 5 246 916 | 5 332 109 | 3 757 663 |
| Total length (>= 25000 bp) | 1 703 243 | 1 703 243 | 1 703 243 | 4 741 746 | 4 940 309 | 4 741 746 | 4 940 309 | 1 703 243 |
| Total length (>= 50000 bp) | 242 666 | 242 666 | 242 666 | 3 904 558 | 4 451 982 | 3 904 558 | 4 451 982 | 242 666 |
| N50 | 16 789 | 16 789 | 16 789 | 79 669 | 118 783 | 79 669 | 118 783 | 16 789 |
| N75 | 8515 | 8515 | 8492 | 44 212 | 57 743 | 44 212 | 57 743 | 8492 |
| L50 | 98 | 98 | 98 | 23 | 14 | 23 | 14 | 98 |
| L75 | 212 | 212 | 213 | 47 | 31 | 47 | 31 | 213 |
| GC (%) | 54.42 | 54.42 | 54.42 | 54.24 | 54.23 | 54.24 | 54.23 | 54.42 |
| **Mismatches** | | | | | | | | |
| # N's | 0 | 0 | 0 | 518 | 1819 | 518 | 1819 | 0 |
| # N's per 100 kbp | 0 | 0 | 0 | 9.35 | 32.82 | 9.35 | 32.82 | 0 |

Plots: Cumulative length Nx GC content



Contigs are ordered from largest (contig #1) to smallest.

**Figure 9.** Quark stats for the assembly of *Enterobacter kobei* using a K-mer = 28.

## Workflow for identifying bacterial drug-resistant genes

```
# download raw files (reads)
$ fastq-dump --split-files ERR037801

# download drug resistence genes (MEGARes, doi:10.1093/nar/gkw1009)
$ wget https://megares.meglab.org/download/megares_v1.01/megares_database_v1.01.fasta

# filetring
$ java -jar Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 ERR037801_1.fastq ERR037801_2
    .fastq ERR037801_1_FILTERED.fastq ERR037801_1_UNPAIRED.fastq ERR037801_2_FILTERED.fastq
     ERR037801_2_UNPAIRED.fastq ILLUMINACLIP:Trimmomatic-0.39/adapters/TruSeq3-PE.fa
    :2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 -validatePairs

# Assembly
$ python3 SPAdes-3.13.0-Linux/bin/spades.py --pe1-1 ERR037801_1_FILTERED.fastq --pe1-2
    ERR037801_2_FILTERED.fastq -t 4 --careful --cov-cutoff auto -o ERR037801_assembly

# assembly stats
$ quast-5.0.2/quast.py /home/ERR037801_assembly/contigs.fasta

# resistance genes prediction

$ glimmer3.02/bin/build-icm MEGARes < megares_database_v1.01.fasta

$ glimmer3.02/bin/glimmer3 /home/ERR037801_assembly/contigs.fasta MEGARes
    drug_resistence_genes

$ glimmer3.02/bin/extract /home/ERR037801_assembly/contigs.fasta drug_resistence_genes.
    predict > putative_drug_resistence_genes.fasta

# gene annotation

$ makeblastdb -dbtype nucl -in megares_database_v1.01.fasta -out MEGARes

$ blastn -query putative_drug_resistence_genes.fasta -evalue 10e-5 -db MEGARes -out
    putative_drug_resistence_genes_annotated.fasta
```