Megan Curry & Reece Theakston

Modern Genetics 425

Dr.Beach

10-18-2018

**Project 1: Eukaryotic Genomes - Pilobolus**

Whole genome sequencing, commonly known as WGS is a method used to read entire genomic sequences. When using the WGS tutorial, certain steps for analysis have to be taken such as: FastQc, Trimmomatic, KmerGenie, SOAPdenovo, and MaSuRCA. A Maker step will be discussed, but not used for this tutorial. FastQc will be used to obtain a data quality summary of all of the individual sequencing reads in the dataset provided. The Trimmomatic program cleans the data and KmerGenie finds the optimal k-mer sizes for the desired dataset while producing the largest genome assembly, this is obtained through DNA sequencing. A Kmer is a list all of the possible subsequences from a genomic read. SOAPdenovo will be used for error correcting functionality. Though MaSuRCA does the same, it also assembles the whole genome from short-read Illumina and creates a smaller number of “super-reads” from the paired-end reads. A “super-read” contains the sequence information from the original reads. Once all of these steps have been finished the data can be used for multiple research projects and findings. This tutorial will demonstrate how WGS can be used in future research such as CRISPR and its effects on cancer patients through their individual genetics.

The first method of WGS is FastQc, and it will be used to check if the sequence ran produced high quality data. The data that was used was already in the system prior to us using it, but in order to perform a new set the user will need to enter their own data into a FastQc program. After entering the data , log into the cluster and make a directory for trimmomatic. Copy all of the FastQ and qsub files from a shared directory to the trimmomatic directory. Once this is done, view your files in the trimmomatic folder. There will be three files, the FastQ program only needs to be ran on one of the data files to create a quality score boxplot. After FastQc is finished, use the program Cyberduck to transfer the html file produced to your laptop. Open the file to view the following data and diagnostics from FastQc. The program will give them two files; file one is specifically for all forward sequence reads, and the second file is for all reverse sequence reads. The files should include: basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, duplicate sequences, and overrepresented sequences and K-mers. Once the analysis is done running Trimmomatic is next.

Next, while using Trimmomatic a quality filter will be ran on the ends of one of the raw reads by using options such as: ILLUMINACLIP (removes artifacts from sequencing), TRAILING and LEADING (requires a Phred score that is the minimum quality score of the base pairs), SLIDINGWINDOW (insures that the window of bases don’t drop below the average Phred score), and MINLEN (gives the minimum length necessary to be retained in the dataset). The Trimmomatic program will be ran on the Juniata HHMI cluster as a qsub script. A qsub script is a program that tells the computer to run a job on “worker” or “compute” nodes. Before running the program edit the directories and make sure the address (username) of the data files corresponds to the job. To do this use the nano program to view the qsub, and directly edit what is needed. After this is done, save and run the trimmomatic program using qsub. The qsub program will submit your jobs to the cluster queue to save the work. Rerun Fastq and open the file in Cyberduck to compare the original data and the data that has been trimmed and filtered.

The third program used is KmerGenie. KmerGenie will produce a K-mer graph after the program is ran, and be used used to understand the expected assembly size. An assembly size is the total length of DNA in contigs (set of overlapping DNA segments that represent a region of DNA) and scaffolds(comprised of contigs) which result from running the assembly program. An assembly size can be less than the genome size due to repetitive elements collapsing into a single contig. An assembly size can also be larger due to read errors and heterozygosity (causes the sequence to look different); resulting in the assembly of different contigs. With respect to producing the largest genome assembly, KmerGenie is used to find the optimal k-mer size for a dataset. KmerGenie works by running five different commands sequentially. *Module load Python*, *module load RMod*, and *module load KmerGenie* load the programs necessary to make KmerGenie. The line “cat/share/apps/sharedData/GCAT/EukGenWorkshop6/2\_ kmergenie/\*.fq > input.fastq” merges raw Fastq files for Ensifer from the shared directory into a single file in your current working directory. The line “kmergenie input.fq” invokes the program in order to make k-mer histograms and determine an optimal k-mer size. The following line “rm input\*” removes unnecessary intermediate datasets.To run the program create a directory for KmerGenie and move into it as your working directory. Examine the contents of a shared folder; it contains four copies of the FASTQ files that was made previously. It also contains the qsub file named “kmergenie.qsub,” copy this file and open the qsub file to edit it. Edit the username and email address as before to insure that everything is running in the right directory. After editing, send the job to the worker nodes, and use the program Cyberduck to transfer the report.html files generate. Once this is done you can open your files and review your results.

The fourth program ran will be SOAPdenovo. SOAPdenovo is a genome assembly program that uses de Bruijin graphs (a graph that assigns K-mers from ever read to directed edges connecting nodes, Eulerian paths which go through every edge once build an assembly of the reads) comprised of four distinct functions including: pregraph (construct k-mer-graph), contig (eliminate errors and output contigs), map (map reads to contigs), and scaff (construct scaffolds). SoapDenovo is also used for error correction by calculating k-mer frequencies, filtering the k-mers below a certain frequency, and correcting frayed rope patterns. The program will be used to create de Bruijn graphs and scaffolds by mapping all paired reads to contig consensus sequences. The data that is used and that SOAPdenovo modules builds an assembly on will already have been through error correction screens. With SOAPdenovo we will assemble a genome from two already error-corrected bacterial genome files. The raw data used will include a paired-end fragment library. In an innie (forward and reverse reads facing each other) orientation, an insert length of 550bp will be included with 2x250bp paired end MiSeq reads. To run the program go to the directory with your last name and make a directory entitled “soap.” Move into the directory and copy all the data from the GCAT shared directory. Use nano to open the “config.txt” file. SOAPdenovo program uses the text file to tell it which files to use. It also tells you where to enter important characteristics of data. The file shows comments explaining parameters and what is ignored by the program (Figure 1). To view the rest of the config files for further comparison use “ls” (Figure 2). After comparing, edit the working directory as before with nano, and run SOAPdenovo using qsub. To see if the job is running use qstat to view the files. Next you will rerun your jobs at different k-mer sizes (21, 51, 97, 199) to see how assembly is affected (Figure 3). While running, check the KmerGenie results to figure out which k-mer size is going produce the best assembly for each dataset. Once this is done, you can create a table for a better comparison of k-mer sizes.

The final step used for WGS was MaSuRCA. MaSuRCA uses a different kind of a approach to the genome that combines the de Bruijn graph and Overlap Layout Consensus (OLC) assembly methods. MaSuRCA assembles whole genomes by using short-read Illumina data, or a compilation of short and long reads. The program creates a small number of “super-reads” from paired-end reads, and these super-reads contain the sequence information from the original reads. For the OLC part, the program uses a modified version of the CABOG assembler. This creates super-reads using only reads that are not substrings of larger reads. The super reads are composed by extending each of the original reads base by base in both the forward and reverse directions. The mate-pair and super-read information is incorporated into the assembly by the OLC assembler. MaSuRCA does have a built-in error corrector, but it also corrects, trims, and removes reads. The program eliminates the k-mer cutoff step to insure that it does not throw out low coverage regions. To run the program create a new masurca directory and move into it. Copy a template qsub file and a configuration file that contains the location of the compiled assembler, the data, and some parameters. Prepare the assembler to create a shell script ("assemble.sh") from the configuration file and drive the assembly process. Run the script “assemble.sh” in the qsub file to start the assembly, then rerun the qsub file. Monitor the assemblies process by using the “less” command. Lastly, the final step of the assembly produces a “CA/10-gapclose” directory which contains the genome listed in scaffolds. Run QUAST from within the directory to obtain the assembly statistics. Move into the QUAST output directory and view the assembly statistics. Compared to SOAPdenovo which throws out all contigs less than 200 bps, MaSuRCA throws out all contigs less than 500 bps. Rerun the assembly on the 100x trimmed data, but create a new directory within the masurca one insure that it does not overwrite your previous assembly. After analysis, delete the result files in order to have more space.

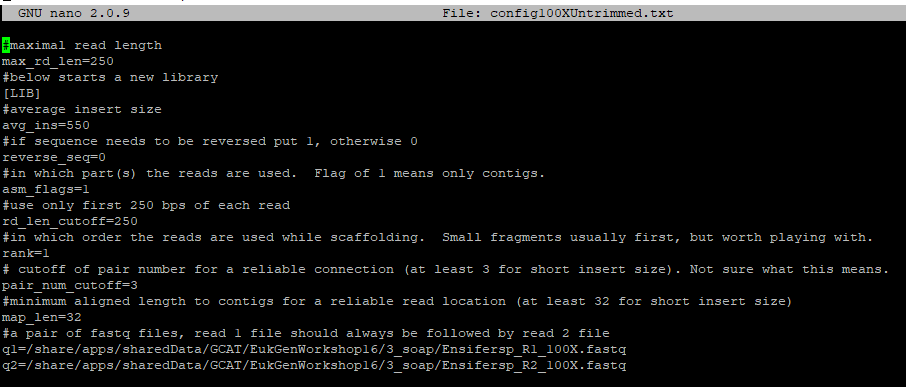
Lastly, the maker program is used as a genome annotation pipeline. The program combines two pieces of information to form structural gene annotations from the raw DNA. the first piece of information is extrinsic evidence supplied to the program based on the similarity of genomic regions to other organisms’ mRNAs or protein sequences. The second piece of information uses gene predictions from signals in the organism’s DNA. These signals are found by ab-initio (from scratch) gene predictors. The Maker program updates features including: 5’ and 3’ UTRs based on the evidence, determines alternate splice forms where the data permits, and chooses the what best matches the evidence from the gene model possibilities. Next, pieces of the eukaryotic genomes consist of SINEs (short interspersed nuclear elements) and LINEs (long interspersed nuclear elements). These two elements could contain coding genes, and these genes need to be defined as exogenous so that the gene finders do not confuse them as endogenous genes. The maker then uses a program called RepeatScout. This program is used to find DNA elements that are repeated more than 10x in the genome and it creates a draft repeat library. RepeatScout counts 12 base pair sequences and extends them to form consensus sequences. RepeatMasker, BLAST2GO, and TEclass are then used for the proteins and repeats. Maker was not used in this WGS tutorial due to technical issues. The program would not run or would crash in the middle of running. Further analysis of the issues could have been taken, but due to time conflicts, maker was not actually used.

Relevancy

All of the steps in the WGS tutorial that are used for analysis including FastQc, Trimmomatic, KmerGenie, SOAPdenovo, and MaSuRCA have been used in multiple CRISPR research projects. FastQc is being used in CRISPR experiments in order to help determine if mutations have formed replicated DNA sequences. Scientists created a custom script in the R language that is then used with the “shortread” option from biconductor.org. Once the script was applied a FastQc file was created for each clone. This process was used to de-multiplex the samples and to show if there were any mutations present. (Bell,2014). Trimmomatic is being used in some CRISPR experiments to trim illumina adapters. This is another necessary step to read and analyze replicas of DNA fragments. When involving CRISPR trimmomatic is mainly used in identifying mutations in replicated DNA.(Canver, 2017). Scientists used KmerGenie in CRISPR research to put the contigs in order. They also adjusted the kmer sizes while using the programr to compare the desired genome with different strains. KmerGenie is also being used by scientists to automatically determine kmer size. (Hébert,20140. Soapdenovo was used in CRISPR research to assemble trimmed reads from a unique contig. This unique contig was previously mapped with a gap-crossed read pair. This gap-crossed read pair was retrieved from short paired-end information.(Li,2015). MaSuRCA was used in CRISPR research to put together hybrid assemblies. MaSuRCA also has the ability to assemble hybrid assemblies of both plants and animal genomes.(Ruan,2017)

In conclusion, WGS can be ran with multiple programs and for multiple purposes. The main purpose of us using WGS was to determine the quality of replicated DNA sequences. In this tutorial we expressed every program that was beneficial and cooperative to us and the potential CRISPR research. With CRISPR research, the main use of the WGS steps was to determine if the replicated DNA sequences contained unwanted mutations. It was reported that WGS and its programs was successfully used and beneficial to the CRISPR research. We expect and hope to use this method in future career opportunities.

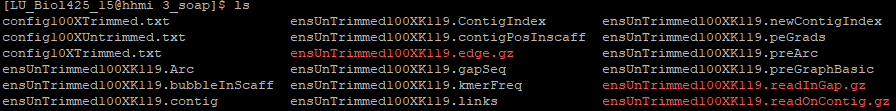
**Figures**



**Figure 1. Nano of the config 100x untrimmed file.**

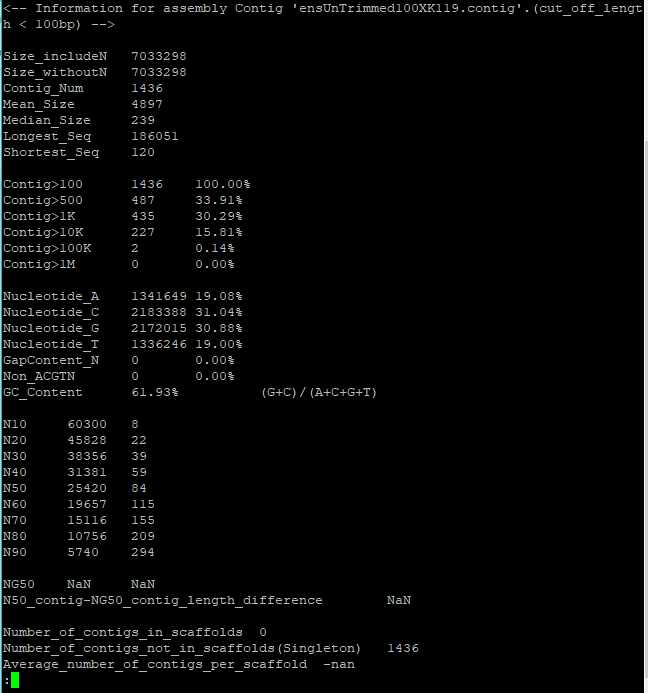
The nano shows what is ignored by the program and everything that is in the file itself.

The pound sign shows comments that explain the parameters below them. The command for this figure is nano config100xuntrimmed.qsub.



**Figure 2. List of contig files**

The command ls was used in order to show all of the contig files in SoapDenovo.



**Figure 3. Nano of Soap**

The nano file of SoapDenovo shows the desired kmer size and allows for kmer size to be changed. The command for this figure is nano soap.qsub.

Citations

Aleksey V. Zimin, Daniela Puiu1, Ming-Cheng Luo, Tingting Zhu, Sergey Koren, Guillaume Marçais, James A. Yorke, Jan Dvořák, Steven L. Salzberg. 2017 Jan. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *genome.cshlp.org*

Charles C Bell,Graham W Magor, Kevin R Gillinder† and Andrew C Perkins. 2014 Nov. A high-throughput screening strategy for detecting CRISPR-Cas9 induced mutations using next-generation sequencing. *Bmcgenomics.biomedcentral.com*

Laurent Hébert, Fabrice Touzain, Claire de Boisséson, Marie-France Breuil, Fabien Duquesne, Claire Laugier, Yannick Blanchard, Sandrine Petry. 2014 Nov. Draft Genome Sequence of Taylorella equigenitalis Strain MCE529, Isolated from a Belgian Warmblood Horse. *Mra.asm.org*

Matthew Canver, Maximilian Haeussler, Daniel E. Bauer, Stuart H. Orkin, Neville E. Sanjana, Ophir Shalem, Guo-Cheng Yuan, Feng Zhang, Jean-Paul Concordet, Luca Pinello.2017 Apr. Integrated computational guide design, execution, and analysis of arrayed and pooled CRISPR genome editing experiments. [*www.biorxiv.org*](http://www.biorxiv.org)

[Molly Gasperini](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gasperini%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28712454), [Gregory M. Findlay](https://www.semanticscholar.org/author/Gregory-M.-Findlay/40271733), [Aaron McKenna](https://www.semanticscholar.org/author/Aaron-McKenna/1679420), [Jennifer H Milbank](https://www.semanticscholar.org/author/Jennifer-H-Milbank/47429000), [Choli Lee](https://www.semanticscholar.org/author/Choli-Lee/3272591), [Melissa D. Zhang](https://www.semanticscholar.org/author/Melissa-D.-Zhang/37264638), [Darren A Cusanovich](https://www.semanticscholar.org/author/Darren-A-Cusanovich/6639465), [Jay Shendure](https://www.semanticscholar.org/author/Jay-Shendure/2431330). 2017 July. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. [*www.ncbi.nlm.nih.gov*](http://www.ncbi.nlm.nih.gov)

Steven R. Head, [H. Kiyomi Komori](https://www.ncbi.nlm.nih.gov/pubmed/?term=Komori%20HK%5BAuthor%5D&cauthor=true&cauthor_uid=24502796), [Sarah A. LaMere](https://www.ncbi.nlm.nih.gov/pubmed/?term=LaMere%20SA%5BAuthor%5D&cauthor=true&cauthor_uid=24502796), [Thomas Whisenant](https://www.ncbi.nlm.nih.gov/pubmed/?term=Whisenant%20T%5BAuthor%5D&cauthor=true&cauthor_uid=24502796),[Filip Van Nieuwerburgh](https://www.ncbi.nlm.nih.gov/pubmed/?term=Van%20Nieuwerburgh%20F%5BAuthor%5D&cauthor=true&cauthor_uid=24502796),[Daniel R. Salomon](https://www.ncbi.nlm.nih.gov/pubmed/?term=Salomon%20DR%5BAuthor%5D&cauthor=true&cauthor_uid=24502796), and [Phillip Ordoukhanian](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ordoukhanian%20P%5BAuthor%5D&cauthor=true&cauthor_uid=24502796). 2014 Feb. Library construction for next-generation sequencing: Overviews and challenges. *www.ncbi.nlm.nih.gov*

Sung-Eun Shin, Jong-Min Lim, Hyun Gi Koh, Eun Kyung Kim, Nam Kyu Kang, Seungjib Jeon, Sohee Kwon, Won-Sub Shin, Bongsoo Lee, Kwon Hwangbo, Jungeun Kim, Sung Hyeok Ye, Jae-Young Yun, Hogyun Seo, Hee-Mock Oh, Kyung-Jin Kim, Jin-Soo Kim, Won- Joong Jeong, Yong Keun Chang, Byeong-ryool Jeongc. 2016 Jun.CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii.

Wei Li, [Johannes Köster](https://www.ncbi.nlm.nih.gov/pubmed/?term=K%26%23x000f6%3Bster%20J%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Han Xu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Xu%20H%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Chen-Hao Chen](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chen%20CH%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Tengfei Xiao](https://www.ncbi.nlm.nih.gov/pubmed/?term=Xiao%20T%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Jun S. Liu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Liu%20JS%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Myles Brown](https://www.ncbi.nlm.nih.gov/pubmed/?term=Brown%20M%5BAuthor%5D&cauthor=true&cauthor_uid=26673418), [Shirley Liu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Liu%20XS%5BAuthor%5D&cauthor=true&cauthor_uid=26673418). 2015 Dec.Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR.*www.ncbi.nlm.nih.gov*

[Xueyan Li](https://www.ncbi.nlm.nih.gov/pubmed/?term=Li%20X%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Wei Zhang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhang%20W%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Guichun Liu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Liu%20G%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Lu Zhang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhang%20L%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Li Zhao](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhao%20L%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Xiaodong Fang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Fang%20X%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Lei Chen](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chen%20L%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Yang Dong](https://www.ncbi.nlm.nih.gov/pubmed/?term=Dong%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Yuan Chen](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chen%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Yun Ding](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ding%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Ruoping Zhao](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhao%20R%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Mingji Feng](https://www.ncbi.nlm.nih.gov/pubmed/?term=Feng%20M%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Yabing Zhu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhu%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Yue Feng](https://www.ncbi.nlm.nih.gov/pubmed/?term=Feng%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Xuanting Jiang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Jiang%20X%5BAuthor%5D&cauthor=true&cauthor_uid=26354079),[Deying Zhu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhu%20D%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Hui Xiang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Xiang%20H%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Xikan Feng](https://www.ncbi.nlm.nih.gov/pubmed/?term=Feng%20X%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Shuaicheng Li](https://www.ncbi.nlm.nih.gov/pubmed/?term=Li%20S%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Jun Wang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20J%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Guojie Zhang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhang%20G%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Marcus R. Kronforst](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kronforst%20MR%5BAuthor%5D&cauthor=true&cauthor_uid=26354079), [Wen Wang](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20W%5BAuthor%5D&cauthor=true&cauthor_uid=26354079). 2015 Sep. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *www.ncbi.nlm.nih.gov*

[Zhi Ruan](http://europepmc.org/search/?scope=fulltext&page=1&query=AUTH:%22Ruan%20Z%22),[Yan Chen](http://europepmc.org/search/?scope=fulltext&page=1&query=AUTH:%22Chen%20Y%22), [Jianfeng Wang](http://europepmc.org/search/?scope=fulltext&page=1&query=AUTH:%22Wang%20J%22). 2017 Nov.Glimpse into the genome sequence of a multidrug-resistant Acinetobacter pittii ST950 clinical isolate carrying the bla OXA-72 and bla OXA-533 genes in China.*europepmc.org*