

Асемблиране на геноми с използването на технологията за фрагментно секвениране Illumina. Анонсиране на прокариотни геноми

Гл. ас. Владимир Толчков, Национален
Център по Заразни и Паразитни Болести,
отдел Микробиология, сектор Микробиом
tolchkov@gmail.com

В дадената презентация е представена работа с асемблер *galaxy*, анотацията на геномите – чрез база данни Prokka (usegalaxy.org), асемблираният и аотиран геном е на *Mycobacterium tuberculosis* секвениран целогеномно по технолофията Illumina.

- Резултат от секвенирането – нуклеотидни фрагменти, визуализирани във формат FASTQ, съдържащ информация за апарата, координатите където се е намирал фрагмента в плаката
- @HSR:363:h2tcnbcxy:1:1101:16331:1998
1:N:0:TAAGGCGA...

Асемблер на геноми

- Програма, наслагваща секвенираните фрагменти, подредени хаотично в изходящия от секвенатора файл върху базиданни, от които сглобява цял геном.
- Сглобеният геном се визуализира във формат FASTA, разделен на фрагменти – континги, колкото са по-големи и по-малко на брой, толкова по-качествено е асемблирането.

Асемблиран фрагмент във формат FASTA

>1 length=227706

depth=1.01xTGACACTGAGACGCAAAGGGCCCCCATTTCGTGCCGAAATGGGGTGCTTTTGCG
TCTGCTCGGCCCAGGTCCAAGAAGTTGCCAAGAATTTCGCTGAGGGGTGTGCCGGAGTCTGTG
GTCATGTCGTATTCGTATTTTCGTTGAGCTTCCCAGGCTCGAGGACATCGAGCCTGGCGCGCACA
CCGACGTTCTGATTGCGAACTCACGGGTGGACCAGGGGGCGTATCCGCGCGGGCGGTGGAGGCG
GTCTTCGACGCCCATCCGGCCCTTGGCACCGTATTCGAGCCGCGCGTTGACACCTTGACTTCTC
GCCCCGGGCGGGCGGGGTGGGGCTGGGGAGTGGAACCCCCGGGAGCCGCCGTCGCGGAGG
TGATCGCACGGCACAGCGCGAGCTTCGATATGTACACCGGCAGGTTGTTTCGCGGTTTTCTCTGCT
CCCCGGAAGTCCCGACCGGCTAGTACTTACCGCCAGCCGCCTCTGCGTGGACGATGCCTCGTG
GCAGACCGTGGTCGAAGACCTGGTGAGGCAATACGACGAGAGTGTGCTGGTGCCAGCACGGT
AGGCATCCTGGTGTCCGCGGCGGGTCGGCGAGCCGGCCTGCCTACCGAAGGGTAGGCGCGG
GTGGGCGCAGAACCGGGGATGGGCAAACCCGCCGGCTAACGGCCCCGATTTCGTGCGGTATGAC
CTCGAACGTGAGTTCATGGTCGCCGATGCGAATGTGGTCGCCGTCGTTGAGGGTGGCTGTGGT
CGCGATGCGCCGCCACGCACGTAGACGCCGTTGACCGATCGCAGGTCGGTGATCATGAAGCT
TTCACCGGTGTTGACGATAACGGCATGGTAGGGACTGACTTTGCCGTCCGGCAGCACCATGTC
GTTACTTTTGCTACGCCCCGATACGAAGGGGGCAACCGGCCAAGTGGAGACCGGTGCCCCGCGG
CGTCGCGTATTGCGGCGCGCGGGCGAACGGTCGGTCATGCCGGGAGAGTGTTTCGAG...

Асемблиран фрагмент в GRAPH файл

S1TGACACTGAGACGCAAAGGGCCCCCATTTTCGTGCCGAAATGGGGTGCTTTTGCGTCTGCTCGGCCCAGG
TCCAAGAAGTTGCCAAGAATTCGCTGAGGGGTGTGCCGGAGTCTGTGGTCATGTCGTATTCGTATTTCTGTTG
AGCTTCCCAGGCTCGAGGACATCGAGCCTGGCGCGCACACCGACGTTCTGATTGCGAACTCACGGGTG...

...CTCGTGGATGCCCCGCGC

LN:i:227706

dp:f:1.011076220438104S

Анонсиране на геноми

- Определяне на структурни елементи (гени, регулаторни участъци, промотори и т.н.)
- 2 подхода – търсене на всички налични в съответните бази данни и задаване ръчно на представляващи за нас интерес секвенции.

Последователност на анонсиране

1. Наслагване на секвенции на рРНК, иРНК и риРНК от съответните бази данни
2. Изтриване от аотирания геном на вече локализираните гени за рРНК
3. Локализиране на други гени, не кодиращи белтък
4. Изтриване на всичко открито до тук
5. Локализиране на кодиращи гени.

Визуализация

GFF (General Feature Format) съдържа в табличен вид:

1. Seqname – име на секвенцията
2. Source – име на програмата или базата данни, локализиращи дадената генетична структура
3. Feature – структурен елемент (ген, вариация, сходство)
4. Start – начало (номер на първия нуклеотид) на елемента в секвенцията
5. End – край на елемента в секвенцията
6. Score – математически оценява качеството на асемблирането за дадения feature
7. Strand – сенс (+) или антисенс (-)
8. Frame – рамка на четене 0, 1 или 2. При 0 първият нуклеотид дава начало на кодона, при 1 вторият, при 2 – третият
9. Attribute – описание на елемента

GFF

The screenshot shows the Galaxy web interface. The top navigation bar includes tabs for 'Galaxy', 'Call for Manuscript - tolchk', 'Google Преводач', 'конгрес на бам парк хотел', and 'GFF/GTF File Format'. The main content area displays a table of genomic data with columns for ID, tool, version, and various numerical values. The right sidebar contains a 'History' panel with a search bar and a list of datasets. A specific dataset, '181: Prokka on data 74: gff', is highlighted, showing its size (75,686 lines, 134 comments) and format (gff, database: 223). The bottom of the screen shows a Windows taskbar with various application icons and a system clock indicating 1:31 PM on 10/14/2019.

ID	Tool	Version	Size	Score	Other	Command
1	Prodigal:2.6	CDS	17427	18383	.	- 0 ID=DMONIKPF_00018;eC_number=
1	Prodigal:2.6	CDS	18530	19390	.	+ 0 ID=DMONIKPF_00019;eC_number=
1	Prodigal:2.6	CDS	19521	19880	.	+ 0 ID=DMONIKPF_00020;eC_number=
1	Prodigal:2.6	CDS	20701	21990	.	+ 0 ID=DMONIKPF_00021;db_xref=COG
1	Prodigal:2.6	CDS	22337	22771	.	- 0 ID=DMONIKPF_00022;eC_number=
1	Prodigal:2.6	CDS	22774	23022	.	- 0 ID=DMONIKPF_00023;db_xref=COG
1	Prodigal:2.6	CDS	23463	23801	.	+ 0 ID=DMONIKPF_00024;Name=higB2
1	Prodigal:2.6	CDS	23804	24127	.	+ 0 ID=DMONIKPF_00025;Name=higA2
1	Prodigal:2.6	CDS	24665	25012	.	+ 0 ID=DMONIKPF_00026;inference=ab
1	Prodigal:2.6	CDS	25009	25629	.	+ 0 ID=DMONIKPF_00027;inference=ab
1	Prodigal:2.6	CDS	25789	27054	.	- 0 ID=DMONIKPF_00028;inference=ab
1	Prodigal:2.6	CDS	27222	27431	.	+ 0 ID=DMONIKPF_00029;inference=ab
1	Prodigal:2.6	CDS	27466	27651	.	+ 0 ID=DMONIKPF_00030;inference=ab
1	Prodigal:2.6	CDS	27678	28862	.	- 0 ID=DMONIKPF_00031;inference=ab
1	Aragorn:1.2	tRNA	29734	29810	.	- ID=DMONIKPF_00032;inference=CC
1	Prodigal:2.6	CDS	29927	30391	.	+ 0 ID=DMONIKPF_00033;inference=ab
1	Prodigal:2.6	CDS	30422	30577	.	+ 0 ID=DMONIKPF_00034;inference=ab
1	Prodigal:2.6	CDS	30561	33539	.	- 0 ID=DMONIKPF_00035;inference=ab
1	Prodigal:2.6	CDS	33631	34629	.	- 0 ID=DMONIKPF_00036;Name=yjblC
1	Prodigal:2.6	CDS	34749	36149	.	+ 0 ID=DMONIKPF_00037;inference=ab
1	Prodigal:2.6	CDS	36230	37054	.	+ 0 ID=DMONIKPF_00038;inference=ab
1	Prodigal:2.6	CDS	37063	37236	.	- 0 ID=DMONIKPF_00039;inference=ab
1	Prodigal:2.6	CDS	37391	38734	.	+ 0 ID=DMONIKPF_00040;eC_number=
1	Prodigal:2.6	CDS	38768	39058	.	- 0 ID=DMONIKPF_00041;Name=whiB

History
search datasets

Unnamed history
160 shown, 32 deleted
5.44 GB

181: Prokka on data 74: gff
75,686 lines, 134 comments
format: gff, database: 223

Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/galaxy-
repl/main/jobdir/024/758/24758502/_
-Xmx30g -Xms256m
Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/galaxy-
repl/main/jobdir/024/758/24758502/_
-Xmx30g -Xms256m

Визуализация

- GBK формат на GeneBank, визуализира анансите както биха изглеждали ако са качени на [https://www.ncbi.nlm.nih.gov › genbank](https://www.ncbi.nlm.nih.gov/genbank)

GBK

Galaxy

Bx. поща - tolchkov@gmail

Google Преводач

Call for innovation to advan

GBK File Extension - What i

← → ↺ 🏠

https://usegalaxy.org/

🔖 ☆ ⚙️ 📝

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

Using 2%

Tools

search tools

Expression Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

LOCUS 1 184525 bp DNA linear 30-AUG-2019

DEFINITION Mycobacterium tuberculosis strain strain.

ACCESSION

VERSION

KEYWORDS .

SOURCE Mycobacterium tuberculosis

ORGANISM Mycobacterium tuberculosis

Unclassified.

COMMENT Annotated using prokka 1.13.3 from https://github.com/tseemann/prokka.

FEATURES Location/Qualifiers

source 1..184525

/organism="Mycobacterium tuberculosis"

/mol_type="genomic DNA"

/strain="strain"

CDS complement(345..2390)

/locus_tag="DMONIKPF_00001"

/inference="ab initio prediction:Prodigal:2.6"

/inference="similar to AA sequence:UniProtKB:P9MHZ7"

/codon_start=1

/transl_table=11

/product="putative PPE family protein PPE40"

/translation="MLLTGTRIFTKSPLFVAPFSYSLFYEYRDVHRCLSHWPGRPVGE

GAGLMNYSVLPPEINSLRMFTGAGSAPMLAASVAWDGLAAELAVAASSFGSVTSGLAG

QSWQGAAMAAAAAPYAGWLAARAAAGASQAQAKAVASAFEARAATVHPLVAA

NRNFAVQLVLSNLFQONAPAIAAAEAMYEQMWAAADVAAMVGYHGGASAAAALPSWQQ

ALRGLPLGQVASAIISGGAAMFAAPAAATAAVTPPALNTGLGNIGSWNLGGGNVGLL

NLGSNGFSGSLNLGGGNTGNANLGGGNMGFANLGSNGINTNFNGNGNGLNFGSGNLL

GNGNFGFGNAFGDGNLGSNGVSTNLGSNGFGSFNVGSGNMGMSNIGFNLGNLGNLGF

GNNGNINIGFGLTGDNLVGIGALNSGIGNMGFGNSGNNINIGFNSGNGNMGVFFNSGDG

NTFGNAGDVNTGFHNGGPFNTFGNGGNTNFNGNAGFQNMGHGAGGVNMGVSGNAG

LANTGDFNSGGVSGIGNTGSFNSGNLNTFGNAGDLNTGLFNSGDVNTGIGSTVDQ

PGSVSGFNGTGTSVSGFNNSGNLTSFGNINNSNVFDSTSGFQNIIGDANVGFNSGNSN

EGFNTGMEFNNGTYNSGVASTGTANSNGNASSGVANSNDNSGAGNAGNAGFEGOP"

History

search datasets

Unnamed history

160 shown, 32 deleted

5.44 GB

GAAATCACGGCACCAATATCGGCGATTGTGCAAAACACTTGTA

182: Prokka on data 74: gbk

181: Prokka on data 74: gff

75,686 lines, 134 comments

format: gff, database: 223

Picked up _JAVA_OPTIONS: -

Djava.io.tmpdir=/galaxy-

repl/main/jobdir/024/758/24758502/_j

-Xmx30g -Xms256m

Picked up _JAVA_OPTIONS: -

Djava.io.tmpdir=/galaxy-

repl/main/jobdir/024/758/24758502/_j

-Xmx30g -Xms256m

2:23 PM

FNA – Fasta DNA формат съдържа в чист вид дадените за анотация секвенции получени при асемблирането на отделните фрагменти (contings), подредени и номерирани по размер в низходящ ред (в дадения пример най големият континг около 230 000 нд, най-малкият – малко над 100 нд, при размер на генома около 4.3 млн нд).

FNA

Galaxy Bx. поща - tolchkov@gmail Google Преводач Call for innovation to advan fna format - Google Търсене

Find on page 2 No results Options

Galaxy Analyze Data Workflow Visualize Shared Data Help User Using 2%

Tools search tools

Expression Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

>1
GATCCCAAGCCACCGATGCCGAAGCATCGGCGAGACCCCGACCCGGTAAACATCCGCA
ACGAATTAATCTCCGACGGCAACACCGAATAATTATCAGCCAGCCCCCTCCCGAGCGC
GCGACGCCGATGACACAGGCGTTGCGGCACGCTACTCCACCCGTAAGAACAACTGTAGG
GAAATCACGGCACCAATATCGCGGATTTGTCAAAACACTTGTACATTGCCAAAAATTCGG
GCCACCGATCGCCACCCCTGGTCACGCGACTCTGCCACATTGCCGCCGCGGTACCTCA
TCGTGCCGGCTAATCGCCCCAGCAACGTCGGGCTGGTAGGCGTTACGGCTGGCCGAAG
AACCCGGCTGGTTATCGCCCTGGTTGAAGGCACCCGAGGAGTTGTACCTGAGTTGGCG
ACACCGGAGTGCGATTACCGGAGTTTGGCAGTGCCGGTGCTGGCCAGCCCGAGTTGTAG
TTGACGTTGTTAAACATGCCGCTGTTGAAGAAGCCCTCTTGGAGTTGCCGAGTTAAAG
AAGCCGACATTGCGCTGCCAATATTCTGGAAGCCTGAGGTGGAGTGAAGACGTTTCGAG
TTCATGTTCCCAAGCCGAGGTAAGGTTACCCGAATTATTGAAGCCCGAGACGCTGGTG
CCGGTGTACCGAAGCCGAGACGAAACAGGCTGATCGACCGTGCTGCCGATGCCGGTG
TTGACGTCACCCGAGTTGAATAGGCTGTGTTGAGATCGCCGGCATTTCGGAAGCCGGTG
TTCAGATTGCCGAGTTGAACGAGCCAGTGTTCACCGATCCGCTGACACACCGCCC
GAGTTGAAATCGCCGGTTGGCCAAGCCCGCTTCCGGAGCCACGTTGACACACCT
CGGTTTCCATGGCCCATGTTCTGGAAGCCCGCTTACCAAAACCGAAGTTTGTGTACCA
CCGTTCCCAAAACCGGTTGGAAGGCTCCCGCTTCCAGAAGCCGGTTGACATCGCCC
CGGTTGCCGAAGCCGGTGTGCCGTCGCCGAGTTAAAGAAGCCACGTTGCCATTGCCG
GAGTTGAAGAAGCCAATGTTGTTGTTCCGAGGTTCCGGAACCATGTTCCGATTCCC
GAGTTCAACGCACCAATGCCACCATGTTGTCACCGGTGAGCCAAACCGATGTTGTTA
TTGCCGTTGTTCCCAAGCCAGGTTGTTGTTACCGAGGTTGCCGAACCGATATTACTC
ATCCCATGTTGCCGTACCCACGTTGAACGAACCAAGTTCCCGTACCGAGGTTTGTA
CTACCGAGGTTGCCGTACCCAGGTTTCCGTCACCGAAGGCGTTTCGGAACCGAAGTT
CCGTTGCCGAGGAGTTCCCGCTGCCAAAGTTGAGATTGCTTGGTTGCCGTTGCCGAAG
TTGGTGTACCAATGTTGCCGCTCCCAAGTTGGCAAGCCCAAGTTCTCCGCGCCAGG
TTGGCATTGCCGGTGTGCCGCCACCCAGGTTAGGCTGCCAAAGTTCCCGCTGCCCAAG
TTCACAAACCGAGTTTCCGCCACCCAGGTTCCAGCTGCCAATATCCCGAGACCCGTG
TTCAGCGCCGGCGGGTGACAGCGCTGTGCGAGCGCTGGGGCGGCAACATGCTAGCC
GCACCGCGGAAATCGCGCTGGCACCTGACCAACCCGGCAGGCCCGCAATGCTGC
TGCCACGATGGCAACGCCGCCGCGCGCGGATGCCCGCGTGATAGCCACCATCGCG

History search datasets

Unnamed history

160 shown, 32 deleted

5.44 GB

AGWAAAAAAGASAAKAVASAF EAARAATVHPMLVAANRN
IAAAEAMYEQMAADVAAMVGHHGASAAAAALPSWQALRGL

183: Prokka on data 74: fna

132 sequences

format: fasta, database: 223

Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/galaxy-
repl/main/jobdir/024/758/24758502/_
-Xmx30g -Xms256m
Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/galaxy-
repl/main/jobdir/024/758/24758502/_
-Xmx30g -Xms256m

Windows taskbar: 2:46 PM 10/14/2019

FAA(Fasta Amino Acid) – визуализира секвенции на белтъци, транслирани от анотирани

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo and links to Analyze Data, Workflow, Visualize, Shared Data, Help, and User. The left sidebar lists various tool categories: Expression Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, and VCF/BCF. The central workspace shows a FASTA file with the following content:

```
>DMONIKPF_00001 putative PPE family protein PPE40
MLLTGTRIFTKSPLFVAPFSYSLFYEYRDVHRLCSHPGRPVGEGAGLMNYSVLPPEINS
LRMFTGAGSAPMLAASVAMDGLAAELAVAASFFGSVTSGLAGQSWQGAAMAAAAAPY
AGWLAASAAAAAGASAQAKAVASAFEAAAAATVHPMLVAANRNAFVQLVLSNLFQGNAP
IAAAEAMYEQMAADVAAMVGYHGGASAAAAALPSWQQALRGLPGLGQVASAISGGAAS
FAAPAAATAAVTPPALNTGLNIGSINLGGGNVGLNLGSGNFGSLNLGGGNTGNANLGG
GNWGFANLGSIGNIGNTFNGNGNQNLNFGSGNLLGNGNFGFGNAFGDGNLGSNGVGS
GSGNFGSFNVGSGNMGMSNIGFNGNLGNNGFNNGNINIGFGLTGDNLVGIGALNSG
NMFGFNGSNNINIGFNSGNGNVGFNSGNGNTGFGNAGDVNTGFWNGGPFNTGFGNG
NFGFGNAGFQNMHGNAGGVNVGSGNAGLANTGDFNSGGVVSIGGNTGFSFNGNLNTG
GNAGDLNTGLFNSGDVNTGIGSTVDQGSVSFGNTGTSVSGFNNSGNLTSFGFNMS
FDSTSGFQNGIDANVGFNSGNSNEGFNTGFMFNIGYNSGVASTGIANSNGNASSGV
GDNSSGAFNQGDNQAGFFGQ
>DMONIKPF_00002 hypothetical protein
MPRQAGRWSPALRILGAAELIALRGYSSTSDIAAAGVQEPAIYKHFSAKRDI
LVRlavePLeLFghITAMPVPAVVKLHRWLtESDHLhasPYVLSILITPDlhQESFV
AERELVAEMERALVGLIETGQGEDVRAMHPLSAARLVQALFDALALPEFAVSPDEIVE
AMTALLSDPDLAEIRAAADALEIQTAPPDRGL
>DMONIKPF_00003 Carnitine monooxygenase oxygenase subunit
MEGMLSTONRAELGDIIDIGDYLDDNPALSLPPAAYTSSELWQLERERIFNRSWMLVA
HVDQVAKTGdyVTVSvAGEPMVVRDVGQLHALSPICRHRMLMVEPGAGRIDTLTCQY
HLWRyGLDGRlRGAPhMAANLDFNRRECLPQFAVATWNGLVWILNDADAEPtAAHLDLT
DDEFAGYRLGEMVQVEShSHEWRAMkVAAENGHENYHVLGLHRQTLEPVPVGGDL
DVRQYSRWALRLVPVTPVPEAKSLQLNEVQSKNLVVLWTFPNSALAIAGERVVWFGFIPQSI
DRVQVLGGVLTTPLEAADAATAQTsqFVMAMINDEDRGLEAVQVGAGSRFAERGHLLS
KEWPGMLAFYRNLAHALVGDHPGAS
>DMONIKPF_00004 hypothetical protein
MTSFaHPGTRGLSTVFGLMVGSAAVGSGLAVVGLAAVIAVGVAAVFRLAATLAVVLS
VVMIVVSGPthVLAALSGFCAAVYLVCRYGAGVAGSWPTTVAAVGFTFAGLAATSFPLQ
VPWLPLAAPLAVLATYVLAATRPFSR
>DMONIKPF_00005 hypothetical protein
MIQTCEVELRWASQLTIAIATCAGVALAAAVVAGRWQLIAFAAPLLGVLCISISWQRPVP
VIOVHGDPDSQRCFENEHVRVTWVTTESVDAVELTVSALAGMQFEALESVSRRTTTS
```

The right sidebar shows the History panel with a search bar and a list of datasets. The datasets are:

- 188: Prokka on data 74: tbl
- 187: Prokka on data 74: fsa
- 186: Prokka on data 74: sqn
- 185: Prokka on data 74: ffn
- 184: Prokka on data 74: faa
- 183: Prokka on data 74: fna
- 182: Prokka on data 74: gbk

FNN (Fasta nucleotide of gene region)

– показва същата информация за гените като FAA на ниво ДНК

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'FASTA format - Wikipedia', and a search bar. The main content area shows a FASTA format file with the header '>DMONIKPF_00001 putative PPE family protein PPE40' and a long sequence of nucleotides. The left sidebar contains a 'Tools' section with a search bar and a list of tools including 'Expression Tools', 'Get Data', 'Send Data', 'Collection Operations', 'GENERAL TEXT TOOLS', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Datamash', 'GENOMIC FILE MANIPULATION', 'FASTA/FASTQ', 'FASTQ Quality Control', 'SAM/BAM', and 'BED'. The right sidebar shows a 'History' section with a search bar and a list of datasets, including '188: Prokka on data 74: tbl', '187: Prokka on data 74: fsa', '186: Prokka on data 74: sqn', '185: Prokka on data 74: ffn', '184: Prokka on data 74: faa', and '183: Prokka on data 74: fna'. The bottom status bar shows the system clock as 8:09 PM on 10/14/2019.

SQN секвенциите са анонсирани по начин,
съвместим с изискванията на NCBI, и могат да
бъдат директно качени в GeneBank

The screenshot displays the Galaxy web interface in a browser window. The address bar shows <https://usegalaxy.org/>. The top navigation bar includes links for **Analyze Data**, **Workflow**, **Visualize**, **Shared Data**, **Help**, and **User**. On the left, a sidebar lists various tools under categories like **Expression Tools**, **GENERAL TEXT TOOLS**, and **GENOMIC FILE MANIPULATION**. The main workspace shows a workflow editor with a tool named **Seq-entry** configured with the following parameters:

```
Seq-entry ::= set {  
  class genbank ,  
  seq-set {  
    set {  
      class nuc-prot ,  
      descr {  
        source {  
          org {  
            taxname "Mycobacterium tuberculosis" ,  
            orgname {  
              mod {  
                {  
                  subtype strain ,  
                  subname "strain" } } ,  
                  gcode 11 } } } ,  
            comment "Annotated using prokka 1.13.3 from  
https://github.com/tseemann/prokka" ,  
            user {  
              type  
                str "NcbiCleanup" ,  
              data {  
                {  
                  label  
                    str "method" ,  
                  data  
                    str "SeriousSeqEntryCleanup" } ,  
                {  
                  label  
                    str "version" ,  
                  data  
                    int 8 } ,  
                {  
                  label  
                    str "month"
```

On the right, a **History** panel shows a list of datasets, including **190: Prokka on data 74: err**, **189: Prokka on data 74: tsv**, **188: Prokka on data 74: tbl**, **187: Prokka on data 74: fsa**, and **186: Prokka on data 74: sqn** (362,765 lines, format: asn1, database: 223).

- TBL показва откритите фичърси върху всеки континг от асемблирания геном

TBL показва откритите фичъри върху всеки КОНТИНГ ОТ асемблирания геном

The screenshot displays the Galaxy web interface with the Funannotate 1.0 tool results. The top navigation bar includes links for Analyze Data, Workflow, Visualize, Shared Data, Help, and User. The sidebar on the left lists various tool categories, including Expression Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS, and GENOMIC FILE MANIPULATION. The main panel shows a table of genomic features with columns for coordinates, CDS status, and detailed annotations. The right sidebar displays a history panel with a list of datasets.

Feature	Start	End	CDS	Annotations
>Feature 1	2390	345	CDS	inference ab initio prediction:Prodigal:2.6 inference similar to AA sequence:UniProtKB:P9WHZ7 locus_tag DMONIKPF_00001 product putative PPE family protein PPE40
	3062	2421	CDS	inference ab initio prediction:Prodigal:2.6 locus_tag DMONIKPF_00002 product hypothetical protein
	4230	3073	CDS	EC_number 1.14.13.- db_xref COG:COG4638 gene yeaw inference ab initio prediction:Prodigal:2.6 inference similar to AA sequence:UniProtKB:F0KFI5 locus_tag DMONIKPF_00003 product Carnitine monoxygenase oxygenase subunit
	4728	4291	CDS	inference ab initio prediction:Prodigal:2.6 locus_tag DMONIKPF_00004 product hypothetical protein
	5996	4725	CDS	inference ab initio prediction:Prodigal:2.6 locus_tag DMONIKPF_00005 product hypothetical protein
	6988	6026	CDS	EC_number 3.6.3.- gene ravA inference ab initio prediction:Prodigal:2.6 inference protein motif:HAMAP:MF_01625 locus_tag DMONIKPF_00006 product ATPase RavA
	7478	6996	CDS	inference ab initio prediction:Prodigal:2.6

The right sidebar shows a history panel with a list of datasets:

- 192: Prokka on data 74: log
- 191: Prokka on data 74: txt
- 190: Prokka on data 74: err
- 189: Prokka on data 74: tsv
- 188: Prokka on data 74: tbl
- 187: Prokka on data 74: fsa

TSV (tab separated values)

- Съдържа в табличен вид служебно зададени от асемблера - locus_tag, вид на структурния елемент - ftype неговата дължина – length_bp, името на гена, ако е ензима – неговия ЕС номер, идентификационния номер в базата данни описание COG, opisane na produkta na negowata ekspresiq

TSV

Galaxy

https://usegalaxy.org/

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

Using 2%

Tools

search tools

Expression Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

1	2	3	4	5	6	7
locus_tag	ftype	length_bp	gene	EC_number	COG	product
DMONIKPF_00001	CDS	2046				putative PPE family protein PPE40
DMONIKPF_00002	CDS	642				hypothetical protein
DMONIKPF_00003	CDS	1158	yeaW	1.14.13.-	COG4638	Carnitine monooxygenase oxygenase sub
DMONIKPF_00004	CDS	438				hypothetical protein
DMONIKPF_00005	CDS	1272				hypothetical protein
DMONIKPF_00006	CDS	963	ravA	3.6.3.-		ATPase RavA
DMONIKPF_00007	CDS	483				hypothetical protein
DMONIKPF_00008	CDS	960				hypothetical protein
DMONIKPF_00009	CDS	627				hypothetical protein
DMONIKPF_00010	CDS	1137		2.7.1.-	COG3173	Putative aminoglycoside phosphotransfer
DMONIKPF_00011	CDS	1125				hypothetical protein
DMONIKPF_00012	CDS	1365	aofH	1.4.3.-	COG1231	Putative flavin-containing monoamine oxi
DMONIKPF_00013	CDS	972	rutD_1	3.5.1.-		Putative aminoacylate hydrolase RutD
DMONIKPF_00014	CDS	483				hypothetical protein
DMONIKPF_00015	CDS	603				hypothetical protein
DMONIKPF_00016	CDS	708		1.-.-.-		putative oxidoreductase
DMONIKPF_00017	CDS	1488	aam	3.5.1.13	COG0154	Acylamidase
DMONIKPF_00018	CDS	957	dhmA	3.8.1.5	COG0596	Haloalkane dehalogenase
DMONIKPF_00019	CDS	861	bpoC_1	1.11.1.18	COG0596	Putative non-heme bromoperoxidase Bpo
DMONIKPF_00020	CDS	360	ddn_1	1.-.-.-		Deazaflavin-dependent nitroreductase
DMONIKPF_00021	CDS	1290			COG1373	putative protein
DMONIKPF_00022	CDS	435		3.1.-.-		Ribonuclease VapC49

History

search datasets

Unnamed history

160 shown, 32 deleted

5.44 GB

192: Prokka on data 74: log

191: Prokka on data 74: txt

190: Prokka on data 74: err

189: Prokka on data 74: tsv

188: Prokka on data 74: tbl

187: Prokka on data 74: fsa

186: Prokka on data 74: ...

ERR – доклад за грешки

Galaxy Bx. поща - tolchikov@gmail.com

https://usegalaxy.org/

Galaxy Analyze Data Workflow Visualize Shared Data Help User

Tools

search tools

Expression Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Discrepancy Report Results

Summary

FATAL: MISSING_PROTEIN_ID:4031 proteins have invalid IDs.

DISC_SOURCE_QUALS_ASNDISC:strain (all present, all same)

DISC_SOURCE_QUALS_ASNDISC:taxname (all present, all same)

DISC_FEATURE_COUNT:CDS: 4031 present

DISC_FEATURE_COUNT:tRNA: 53 present

DISC_FEATURE_COUNT:rRNA: 3 present

DISC_FEATURE_COUNT:repeat_region: 1 present

DISC_FEATURE_COUNT:tmRNA: 1 present

DISC_COUNT_NUCLEOTIDES:132 nucleotide Bioseqs are present

FEATURE_LOCATION_CONFLICT:4088 features have inconsistent gene locations.

OVERLAPPING_CDS:14 coding regions overlap another coding region with a similar or identical name.

SUSPECT_PRODUCT_NAMES:44 product_names contain suspect phrase or characters

Putative Typo

1 features May contain plural

Suspicious phrase; should this be nonfunctional?

2 features contains 'truncat'

May contain database identifier more appropriate in note; remove from product name

22 features contains three or more numbers together that may be identifiers more appropriate in note

15 features Contains underscore

Use short product name instead of descriptive phrase

3 features Is longer than 100 characters. Remove descriptive phrases or synonyms from product names. Keep valid long product names, eg long enzyme names

use protein instead of gene as appropriate

1 features contains 'genes'

FATAL: EC_NUMBER_ON_UNKNOWN_PROTEIN:2 protein features have an EC number and a protein name of 'unknown protein' or 'hypothetical protein'

FIND_BADLEN_TRNAS:4 tRNAs are too long

NO_ANNOTATION:4 bioseqs have no features

DISC_QUALITY_SCORES:Quality scores are missing on all sequences.

FATAL: DISC_SUSPECT_RRNA_PRODUCTS:1 rRNA product names contain suspect phrase

FATAL: DISC_SHORT_RRNA:1 rRNA features are too short

ONCALLER_COMMENT_PRESENT:132 comment descriptors were found (all same)

SHORT_PROT_SEQUENCES:72 protein sequences are shorter than 50 aa.

MISSING_GENOMEASSEMBLY_COMMENTS:132 bioseqs are missing GenomeAssembly structured comments

History

search datasets

Unnamed history

160 shown, 32 deleted

5.44 GB

190: Prokka on data 74: err

189: Prokka on data 74: tsv

188: Prokka on data 74: tbl

187: Prokka on data 74: fsa

186: Prokka on data 74: sqn

362,765 lines

format: asn1, database: 223

Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-

1:59 PM 10/16/20

ТХТ кратко резюме за аотириания геном

The screenshot displays the Galaxy web interface at <https://usegalaxy.org/>. The browser tabs include "BHT 1 / 16:9 LIVE", "Galaxy", and "err format bioinformatics -". The interface features a dark navigation bar with options: "Analyze Data", "Workflow", "Visualize", "Shared Data", "Help", and "User". A "Using 2%" status indicator is present in the top right.

Tools Panel (Left):

- search tools
- Expression Tools
- Get Data
- Send Data
- Collection Operations
- GENERAL TEXT TOOLS**
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Datamash
- GENOMIC FILE MANIPULATION**
- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF

Dataset Information (Center):

organism: *Mycobacterium tuberculosis* strain
contigs: 132
bases: 4283712
tRNA: 53
repeat_region: 1
CDS: 4031
rRNA: 3
tmRNA: 1

History Panel (Right):

search datasets

Unnamed history
160 shown, 32 deleted
5.44 GB

ID	Name	Format	Actions
192: Prokka on data 74:	log	log	View, Edit, Delete
191: Prokka on data 74:	txt	txt	View, Edit, Delete
190: Prokka on data 74:	err	err	View, Edit, Delete
189: Prokka on data 74:	tsv	tsv	View, Edit, Delete
188: Prokka on data 74:	tbl	tbl	View, Edit, Delete
187: Prokka on data 74:	fsa	fsa	View, Edit, Delete
186: Prokka on data 74:	View, Edit, Delete

The Windows taskbar at the bottom shows the system clock as 11:11 PM on 10/14/2019, along with various background application icons.

LOG – показва процеса на асемблиране по секунди

The screenshot displays the Galaxy web interface in a browser window. The address bar shows <https://usegalaxy.org/>. The top navigation bar includes links for Analyze Data, Workflow, Visualize, Shared Data, Help, and User, along with a 'Using 2%' indicator.

The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: Expression Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, and VCF/BCF.

The main content area displays a log for the Prokka tool, showing the following output:

```
[01:26:23] This is prokka 1.13.3
[01:26:23] Written by Torsten Seemann <torsten.seemann@gmail.com>
[01:26:23] Homepage is https://github.com/tseemann/prokka
[01:26:23] Local time is Fri Aug 30 01:26:23 2019
[01:26:23] You are not telling me who you are!
[01:26:23] Operating system is linux
[01:26:23] You have BioPerl 1.007002
[01:26:23] System has 32 cores.
[01:26:23] Will use maximum of 6 cores.
[01:26:23] Annotating as >>> Bacteria <<<
[01:26:23] Generating locus_tag from '/galaxy-repl/main/files/033/517/dataset_33517817.dat' contents.
[01:26:23] Setting --locustag DMONIKPF from MD5 d687249f822adc6fae9249e54f355aba
[01:26:23] Creating new output folder: outdir
[01:26:23] Running: mkdir -p outdir
[01:26:23] Using filename prefix: prokka.XXX
[01:26:23] Setting HMMER_NCPU=1
[01:26:23] Writing log to: outdir/prokka.log
[01:26:23] Command: /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/prokka --cpus 6 --quiet --outdir outdir --prefix prokka --increment 1 --gffver 3 --mincontig 200 --genus Mycobacterium --species tuberculosis --kingdom Bacteria --gcode 11 --evaluate 1e-06 /galaxy-repl/main/files/033/517/dataset_33517817.dat
[01:26:23] Appending to PATH: /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin
[01:26:23] Looking for 'aragorn' - found
[01:26:23] /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/aragorn
[01:26:23] Determined aragorn version is 1.2
[01:26:23] Looking for 'barrnap' - found
[01:26:23] /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/barrnap
[01:26:23] Determined barrnap version is 0.9
[01:26:23] Looking for 'blastp' - found
[01:26:23] /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/blastp
[01:26:24] Determined blastp version is 2.7
[01:26:24] Looking for 'cmppress' - found
[01:26:24] /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/cmppress
[01:26:24] Determined cmppress version is 1.1
[01:26:24] Looking for 'cmscan' - found
[01:26:24] /cvmfs/main.galaxyproject.org/deps/_conda/envs/__prokka@1.13/bin/cmscan
[01:26:24] Determined cmscan version is 1.1
[01:26:24] Looking for 'egrep' - found /bin/egrep
```

The right sidebar shows the 'History' section with a search bar and a list of datasets. The 'Unnamed history' section displays 160 shown, 32 deleted datasets, totaling 5.44 GB. The list includes:

- 192: Prokka on data 74: log
- 191: Prokka on data 74: txt
- 190: Prokka on data 74: err
- 189: Prokka on data 74: tsv
- 188: Prokka on data 74: tbl
- 187: Prokka on data 74: fsa
- 186: Prokka on data 74: ...

The bottom status bar shows the system clock as 11:16 PM on 10/14/2019.

Результати

РЕЗИСТЕНТНОСТ	БРОЙ ЩАМОВЕ
RMP INH EMB	17
RMP INH PZA EMB	5
RMP INH	3
FQ RMP INH EMB	5
RMP INH PZA	1
FQ RPM AMK KAN CPR PZA EMB	1
FQ RPM AMK KAN CPR INH EMB	1
FQ RPM AMK KAN CPR EMB	1
ОБЩО	34

Мутации, предизвикващи резистентност

- АБ ГЕН МУТАЦИИ
- RMP rpoB S450L H445Y H445D N432L
- INH fabG1 INTERGENIC S315T
- EMB embB M306V

АНТИБИОТИК	ГЕН	МУТАЦИИ
RPM	proB	S450L H445Y H445D N432L
IHN	fabG1	INTERGENIC S315T
EMB	EMBb	M306V
PZA	pncA	L4S H82R G97D INTERGENIC
FQ	gyrA	A90V D94H
AMK KAN CPR	rrs	RIBOSOMAL

Благодаря Ви за вниманието!