



CNR-IBIOM: Infrastructure, data and analysis resources

Fosso B., Balech B., Tangaro M.A. and Pesole G.

LifeWatch ERIC thematic meeting on Genomics 26-28 February 2020 - Porto



Outlook

The research infrastructure

Bio-molecular data resources

- ✤ ITSoneDB
- COXI-DB

Analysis resources

- MetaShot
- BioMaS



The research infrastructure



Bioinformatics software/services



С

A data-supported history of bioinformatics tools

Phenomic

arXiv:1807.06808

Over 14 thousands entries in Bio.tools







Building an infrastructure for life scientists

Services for science are moving to the Cloud.

The "European Open Science Cloud"

aims to create a trusted environment for hosting and processing research data to support EU science in its global leading role.

Develop Cloud services which can be exploited by life science community.





Galaxy

Galaxy (galaxyproject.org) is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data.

Through a coherent work environment and an **user-friendly web interface** it organizes data, tools and workflows providing reproducibility, transparency and data sharing functionalities to users.

Galaxy

← → C' @ ③ 90.147.17	0.126/galaxy?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fdevteam%2Ffastqc%2Ffastqc%2F0.728 😇 🏠 🔍 Cerca	▲ Ⅲ 🗢 🗉 🔛				
🔁 Galaxy / IRIDA-I	ELIXIR-ITALY test Analyze Data Workflow Visualize - Shared Data - Admin Help - User - 🎫	Using 2.3 G				
Fools 🛓	FastQC Read Quality reports (Galaxy Version 0.72) Options	History C 🌣				
search tools	Short read data from your current history	search datasets				
Get Data	26: Map with Bowtie for Illumina on data 25: mapped reads	Unnamed history				
Send Data	Contaminant list	26 shown				
Collection Operations	C 40 C Nothing selected	319.49 MB 🕑 🐿				
lft-Over	tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAACCAGAAGACGCCATACGA	26. Man with Revula for				
ext Manipulation	tab delimited mie with 2 columns; name and sequence, For example: Illumina small KNA KI Primer CAACAGAGACGCATACGA Adapter list					
Convert Formats Filter and Sort	C & D Nothing selected	mapped reads				
oin, Subtract and Group	list of adapters adapter sequences which will be explicitly searched against the library, tab delimited file with 2 columns: name and sequence.	~1,000,000 lines, 19 comments				
etch Alianments/Sequences	(adapters)	format: sam, database: sacCer3				
perate on Genomic Intervals	Submodule and Limit specifing file	Sequence file aligned.				
itatistics	C 2 D Nothing selected •	B02M2 \$				
araph/Display Data	a file that specifies which submodules are to be executed (default-all) and also specifies the thresholds for the each submodules warning parameter					
henotype Association	Disable grouping of bases for reads > 50bp	1.QNAME 2.FLAG 3.RNAME 4.POS 0HD VN:1.0 S0:unsorted				
ED Tools	Yes No	PS0 SN:chrI LN:238218				
ore-alignment	Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been	eSQ SN:chrII LN:813184				
DNA-alignment	warnedi (nogroup)	05Q SN:chrIII LN:316620				
picard-tools	Lower limit on the length of the sequence to be shown in the report	@SQ SN:chrIV LN:1531933				
cf-tools		25: 👁 🥒				
COVACS	As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths, (min length)	http://159.149.160.56				
covacs_Select_filtration SelectVariants VariantFiltration	length of Kmer to look for	/indigo_demo/Sc_IP.fastq				
wrapper for covacs, use in case of		140.0 MB format: fastgsanger, database:				
not enough snp or indels error in	7	sacCer3				
covacs_VariantRecalibrator	note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (kmers)	and and destroyed on the				
covacs_ApplyRecalibrator gatk ApplyRecalibrator wrapper	✓ Execute	uploaded fastqsanger file				
covacs VariantRecalibrator GATK		E 0 ?				
VariantRecalibrator wrapper	1 Purpose	<pre>eHWI-ST1210:99:C15JEACXX:8:1313:1798 AAGAAAATGCTGGGGTGGATTCTCAAGGAAGGCTTT</pre>				

- Need IT infrastructure (proportional to workload).
- Need IT expertise
- Need Galaxy admin expertise

Laniakea: Galaxy "on-demand" platform





nce European Infrastructure for Biodiversity and Ecosystem Research



https://laniakea-elixir-it.github.io

LANIAKEA is a cloud Galaxy instance provider, based on INDIGO-DataCloud software catalogue. Its architecture automates the creation of Galaxy-based virtualized environments exploiting the software catalogue provided by the INDIGO-DataCloud project.

- No need of local IT infrastructure / expertise
- Data privacy and security (encrypted data volumes)
- Full control over Galaxy instance(s)
- Fully customisable

(*) The Laniakea Supercluster (Laniakea; also called Local Supercluster or Local SCI or sometimes Lenakaeia) is the galaxy supercluster that is home to the Milky Way and approximately 100,000 other nearby galaxies [Wikipedia].



The Pulsar Network

The most innovative computing centers across Europe are currently interested to share their remote computation power to support the European Galaxy server UseGalaxy.eu load.



The funding model: ELIXIR and H2020 projects



propean Infrastructure for Biodiversity and Ecosystem Research

ELIXIR European Research Infrastructure for biological data which primary objective is to support research in the field of "life sciences" and their translational activities to medicine, environment, biotechnological industries and society.



[2015-2017]]Develop an open source computing and data platform, targeted at multi-disciplinary scientific communities, provisioned over public and private e-infrastructures.



[2019-2022] EOSC-Life brings together biological and medical RIs to create an open collaborative space for digital biology. It aims to publish FAIR life science data resources for cloud use creating an ecosystem of innovative tools in EOSC and enabling groundbreaking data-driven research in Europe by connecting life scientists to EOSC.



Empowering the infrastructure: PON R&I 2014-2020 Avviso 424/2018 Azione II. 1



[2019-2022] EOSC-Pillar will support the implementation of the European Open Science Cloud by leveraging national initiatives of the EU Member States and thematic initiatives developed by research communities working in national and European collaborations.



Bio-molecular data resources







Molecular Biodiversity Laboratory (MoBiLab) has a fully operative platforms based on:

- NGS technologies
- Data storage resources
- Computational analysis

In addition to their support to LifeWatch, the services hosted by MoBiLab contribute to the Italian node of the European infrastructure **ELIXIR**.



Bio-Molecular Databases

DIU-INICIECUIAL DALADASES

Curated specialized reference databases

- ITSoneDB: RNA Internal Transcribed Spacer 1 (ITS1) database
- ✓ COXI-DB: Cytochrome Oxidase subunit-I database







ITSoneDB

- A comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences
- Hidden Markov Model (HMM) profiles of ITS1 flanking -18S and 5.8S rRNA genes – are mapped on all the eukaryotic sequences in order to refine boundaries location of this region
- More than 1.1 million curated sequences



Availability: http://itsonedb.cloud.ba.infn.it/

Santamaria, M., et al. (2018), NAR, 46, D127-D132.

LW-ERIC, Porto: Feb 26-28 2020





COXI-DB on-going

on-going

- A collection of Animals' Cytochrome Oxidase subunit-I (COXI) sequences
- More than **5 million** curated sequences
- Sequences validated against PFAM (COXI profile)
- Updated records' Taxonomy according to NCBI taxonomy (whenever possible)







METAGENOMICS

Function-based Metagenomics

• Screen to identify functions of interest such as vitamins and antibiotic production

Extract data from microbial community in sampled environment

Sequence-based Metagenomics

•Determine the taxonomic composition of microbial communities;

Determine what genes are represented, i.e. identify genes and metabolic pathways



Essential Biodiversity Variables (EBVs)



Globis-B

Globis-B

- Global Biodiversity Monitoring
- Studying, reporting and managing Biodiversity change
- Harmonization & standardization of biodiversity data



Analysis resources



Shotgun Metagenomics

- identify species, genes and functional capabilities of mixed microbial communities;
- much more expensive in terms of sequencing and computational analysis.



Meta-Barcoding

- High sensitivity in species resolution and identification;
- Less expensive in terms of sequencing and computational analysis;
- Universal conditions of PCR;
- Specialized reference database (e.g. RDP for 16S, ITSoneDB for ITS1)
- may be biased due to the different efficiency of marker amplification in the different species;
- No functional information.



BIOINFORMATICS



 The critical bottleneck for NGS based projects is "Bioinformatics". The huge amount of sequence data generated by NGS platforms requires adequate computational infrastructures and bioinformatic resources for storage, retrieval and analysis of the data.

<u>The analysis of data requires advanced skills for</u> <u>establishing and running complex workflow</u> <u>including many steps.</u>



METASHOT

METASHOT is an automated pipeline designed for the identification of the microbial component in genomic (DNA-Seq) and transcriptomic (RNA-Seq) data.

Third party tools and *ad hoc* developed Python and BASH scripts are integrated to manage, analyze and taxonomically assign Illumina PE data.

Bioinformatics, 2017, 1–3 doi: 10.1093/bioinformatics/bb036 Advance Access Publication Date: 27 January 2017 Applications Note

Sequence analysis

MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data

B. Fosso¹, M. Santamaria¹, M. D'Antonio², D. Lovero¹, G. Corrado³, E. Vizza³, N. Passaro⁴, A.R. Garbuglia⁵, M.R. Capobianchi⁵, M. Crescenzi⁴, G. Valiente⁶ and G. Pesole^{1,7,*}





BENCHMARK

MetaShot has been benchmarked against **Kraken** and **MetaPhIAn2**. The Benchmark has been performed by using:

1. An in-silico generated human microbiome



2. A mock community consisting of 4 bacterial and 9 viral species (SRR3458569).

Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis

Nádia Conceição-Neto, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, Rob Lavigne, Piet Maes, Marc Van Ranst, Elisabeth Heylen & Jelle Matthijnssens 🏁



BENCHMARK

	Human (host)		Prokaryotes		Viruses				
	KR	MS	MP*	KR	MS	MP	KR	MS	MP
Precision	99.85	100.0	0	35.67	98.13	98.00	94.95	98.30	80.93
Recall	100.0	99.97	0	35.16	84.52	87.31	92.77	98.19	79.32
F-rate	99.92	100.0	0	35.36	86.79	90.72	92.82	98.07	79.93
Unclas	0.00	1.04	99.99	55.28	2.44	94.50	4.25	3.94	30.74







Expected relative abundances

MetaShot outperforms **Kraken** and **MetaPhIAn2** in terms of the overall accuracy of reads assignment for the Prokaryotes and Viruses at the Family, Genus and Species levels.



SAMPLES COLLECTION

METAGENOME EXTRACTION

BARCODE AMPLIFICATION

BIOINFORMATIC ANALYSIS

SEQUENCING

<u>Hypervariable regions of 16S rRNA</u> ITS1 of the ribosomal gene cluster

Meta-Barcoding

Taxonomic Genomic Markers

- Ubiquity in taxonomic range of interest (e.g. Bacteria, Fungi, etc)
- Reliable discrimination capacity at species level
- Hyper-variable regions flanked by highly conserved ones in taxonomic range under study
- Dimension fitting with the sequencing platforms read length





Benchmark

BioMaS classification performance has been compared to QIIME and Mothur





Benchmark: Bacteria



BioMaS assigned sequences
 QIIME assigned sequences
 Mothur assigned sequences
 Mothur correctly assigned sequences
 Mothur correctly assigned sequences



Benchmark: Fungi



 BioMaS assigned sequences BioMaS correctly assigned sequences 	 QIIME assigned sequences QIIME correctly assigned sequences 	 Mothur assigned sequences Mothur correctly assigned sequences
- Diolinao correctly assigned sequences	- Gime correctly assigned sequences	- Wollar concerty assigned sequences