

# Systems and servers for NGS data analysis

Rosangela Tozzoli

NGS Course,  
11-12 July 2019




Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# Data Analysis: A new syntax

Mobio\_16Run\_400\_hiq\_Pool17062015.fastq

```
@C9IBY:00426:00452
ATCAATTAATAATTTATCTAGCGCATTACATGCACTGATTTTATCCATTTTGCATTACCACCACATCGAGCAATTTTCCCAG
TCCGCATCGCTGGCAATATAGGCGAAGTTATTCCCTTACCCTAACCAGAACGGCGCACCGAGCATGCTCTTTAACGAAAGCAA
TCAATTGCTCGCCTCCACGTGGAACAATCAAATCGATCGGTTTCATCGGGGTTTTTCAGGAAGGCTCGCGTAGCCTCACGATCCAAC
GTGATGAGTTAATCCAATCTGTCGTCAATCCATTCTGCCGTAACGCGTC
+
BBCDAC6;;;/;CC8CCCCC?>??C@CCEEDDD;;;<<<</<<<<7<<<<<. ;BAA@CCCADDABCDCCCCCCC@@@. ; ;
1;;?CACC??CCCCACCA?;;:=9=>CACE@CC@CC>CCCC;>?CCACACAA=@>B;;@288888:@=888?,82::?
D<BBB=CB@5828=CBBC?CC?;;>:>7<9B>=@@A:@BBAA??;?
B@@8888*888*8<A=A2848888=::@=@@CB?;;8;?BBBB@@@ABBD3:2:0171777000:008700*//*///
828<<4:;@?87
@C9IBY:03696:02678
CTTGGTGGTAATGGTGGTGTGGTTGGGCGGAAGCCATCGCCGGGGTTGGCGCTATTCTTGAAATGCTGCTCGGTGCAGCAGGTG
CTGATGGCCAGGTTTTGACTCGCTGGGCGATACCATCTCTGGTCTGATGGGAGACACAGGTACCTCGGCGTCTGAAGCGGCCGA
TGCCATCACCGCCAGCCTGGGAGGAGCCCTGGATGGTAGCCTGGCGGGCCTGGGTATCGATGCTCCAGCCGAAGCCGTGATCGGA
GCCCTGTTGAGTGGTAAGGCGGGGGCGGTGATGCCGTCTTGAC
+
?@ACACCA????CACCA?;;:5:5:C/:5:<B=@<;>CCCACDD:DADADDCCCFAB@?::;/;BBB>?CD@CCCC>;@<??
ACCCCD@C@CA4888:08@CCC@; ;CC>CC@@@CACCCDDACBCCCCD=CCCC@@@AC@@;??>28888: ?
@ACCE@C@CDDDD>?>C@B?C?:CCCACDE?C??<888, :@4>?C?B@;>;>B7;?C=CAFCD<CC??>CC?@?
CADDACD9: :?CC??@A=??B;BBC?A@@@A=AB?E: ?>BBB3; ;?BBBB>B??>B??
@C9IBY:01239:00533
GGTATGGTGGTGGTTGGGTTGGGTTGGTGTGTGTCGGGCGTGTGGGGCGGGTTGCGGTTGGCTGGCGCGCATGGGCGCCCGGTCA
CGGTGAACGTACCATCCCCTGTCCACCCTGCTCGGCCTGGACGACCACCCGCG
+
D@DDCCC@CCADDAC>; ;1;; ;D?D>?>=::@;B??DDGEE5CCC>CAC;;8;>B>????A;; ;?>:=: ?CCF?
D@CCD??>?CCAF??=BA5:::/:88::B4::/:::AAC4:5;B><<000*00<*000
@C9IBY:02674:01311
TATTACCACAGGTAAGTGAATTAATGATATTAATATTTTTCAACGGTTAGCAGAAATTTATTCGACAGCTGGGGAATTTGAAAA
AGCAATTCATATTATGAACGTGTTTTAGAAGACAGACTTGATATTGACTGCTTATTTGGCTACGGGTTAACAGCATATCAAGCA
```



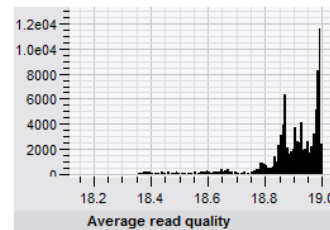
# Data Analysis: Software stand-alone

## Manage next generation sequencing data

The new sequence read sets experiment type offers an integrated environment for importing, preprocessing and analyzing sets of reads from high throughput sequencers or public repositories.

## An integrated NGS data analysis platform

- Fast import of sequence read sets from various next generation sequencing platforms, such as Roche 454, Illumina Solexa, IonTorrent, etc.
- Storage of large amounts of short sequences (including paired-end reads) and quality scores.
- Comprehensive data preprocessing and quality control settings for demultiplexing, splitting paired-end reads, primer removal, structural and quality trimming, chimera detection and cleaning up sequence read sets.
- Global statistics calculation of sequence reads: creation of read length histograms, revision of base distribution, and quality score distribution. Generation of reports in rich text, table and chart formats.
- Sequence read sets are *database objects*, meaning that they can be annotated using custom information fields and that *user privileges* determine who is allowed to access and/or modify the data.
- Create comparisons for Kmer based clustering of sequence read sets, using all available similarity coefficients and *hierarchical clustering methods*.



User-Friendly Interface, Processing, RAM needed



geneious

Transform biological data into knowledge and actionable insights



Biological sequence data made easy



Painfree analysis with an intuitive, user-friendly interface



Simply import and convert a vast range of data types



Cross-platform: Mac, Windows, Linux



Licensing options to suit your users and your budget cycle



Comprehensive support for technical issues, set up and training



Customize with your own algorithms, plugins or workflows



Increase process efficiency and improve data organisation



High interoperability with good API to link LIMS and other tools



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# Data Analysis: Software stand-alone

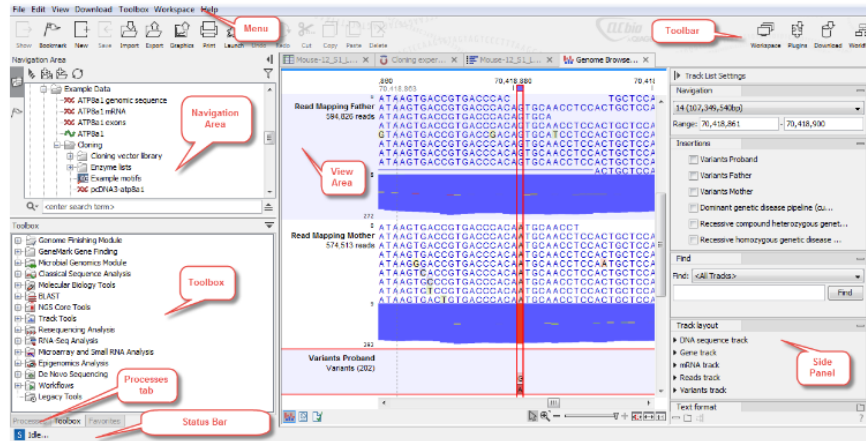
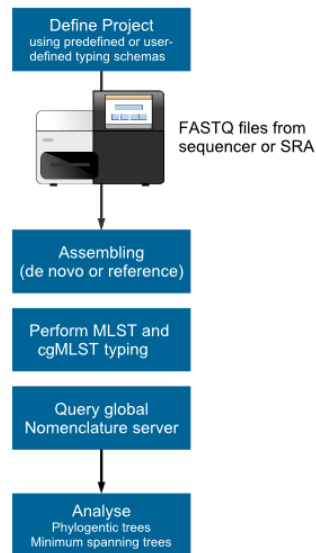


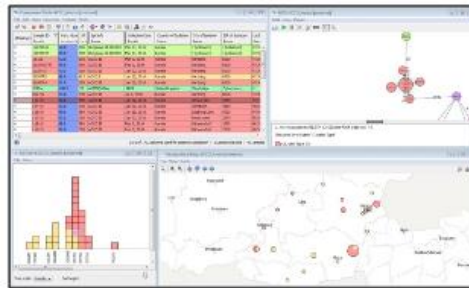
Figure 2.1: The user interface.



Ridom SeqSphere+



Pipeline for automated sequence analysis



Bacterial Genome characterization

Genome-wide allele and SNP calling



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# Data Analysis: Software stand-alone



[Our Customers](#) ▾

[Genomics Data Solutions](#) ▾

[Resources](#) ▾

[Per Sample Pricing](#)

[Sign in](#)

[Request A Demo](#)

**Bring your Genomics Data  
Analysis Workflow to Life.**

Fit-for-purpose. Clinical-grade  
security. Global data compliance.



# Data Analysis: Software stand-alone

---



Torrent Suite  
Software

**ion torrent**  
by *life* technologies™

- *de novo* assembly
- Search for interesting genes
- Alignment of sequences, production of VCF files

**BUILT IN THE ION TORRENT TECHNOLOGY PACKAGE**

# Data Analysis: Cloud-based Software

BaseSpace®  
Genomics Cloud Computing

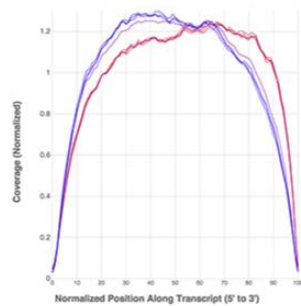
illumina

Sign up

Log in

Now Available on the BaseSpace AppStore

## RNA-Seq Workflow



Filters

$|\log_2(\text{ratio})|$   
0.0

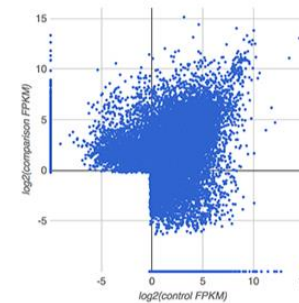
Significant

Choose a value...

Status

OK

Gene



TopHat Alignment



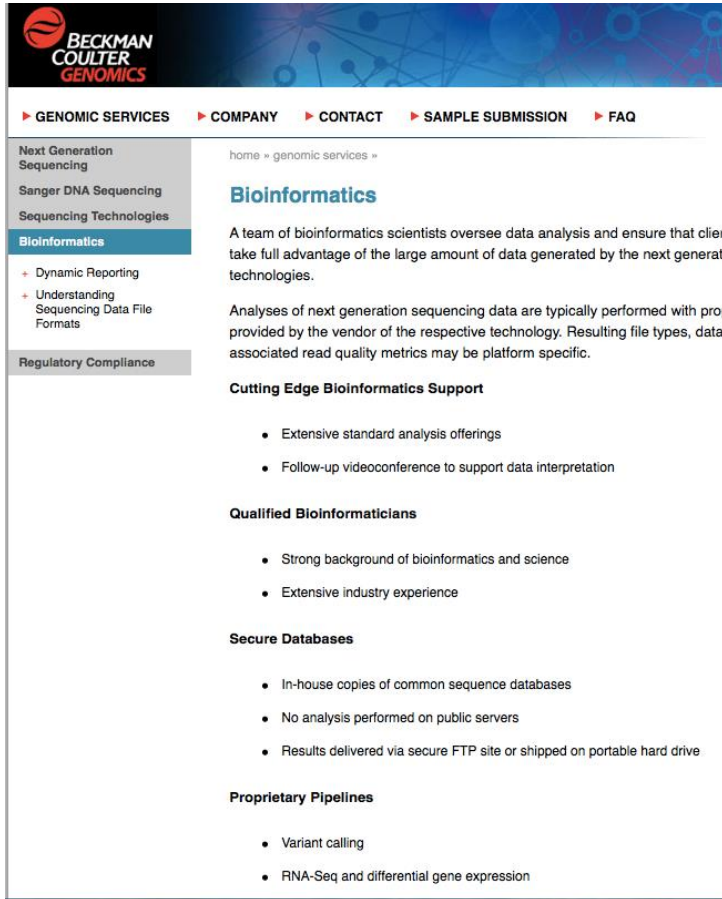
Cufflinks Assembly &  
Differential Expression



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# Data Analysis: Outsourcing



**BECKMAN COULTER GENOMICS**

► GENOMIC SERVICES ► COMPANY ► CONTACT ► SAMPLE SUBMISSION ► FAQ

home » genomic services »

## Bioinformatics

A team of bioinformatics scientists oversee data analysis and ensure that clients take full advantage of the large amount of data generated by the next generation technologies.

Analyses of next generation sequencing data are typically performed with protocols provided by the vendor of the respective technology. Resulting file types, data associated read quality metrics may be platform specific.

### Cutting Edge Bioinformatics Support

- Extensive standard analysis offerings
- Follow-up videoconference to support data interpretation

### Qualified Bioinformaticians

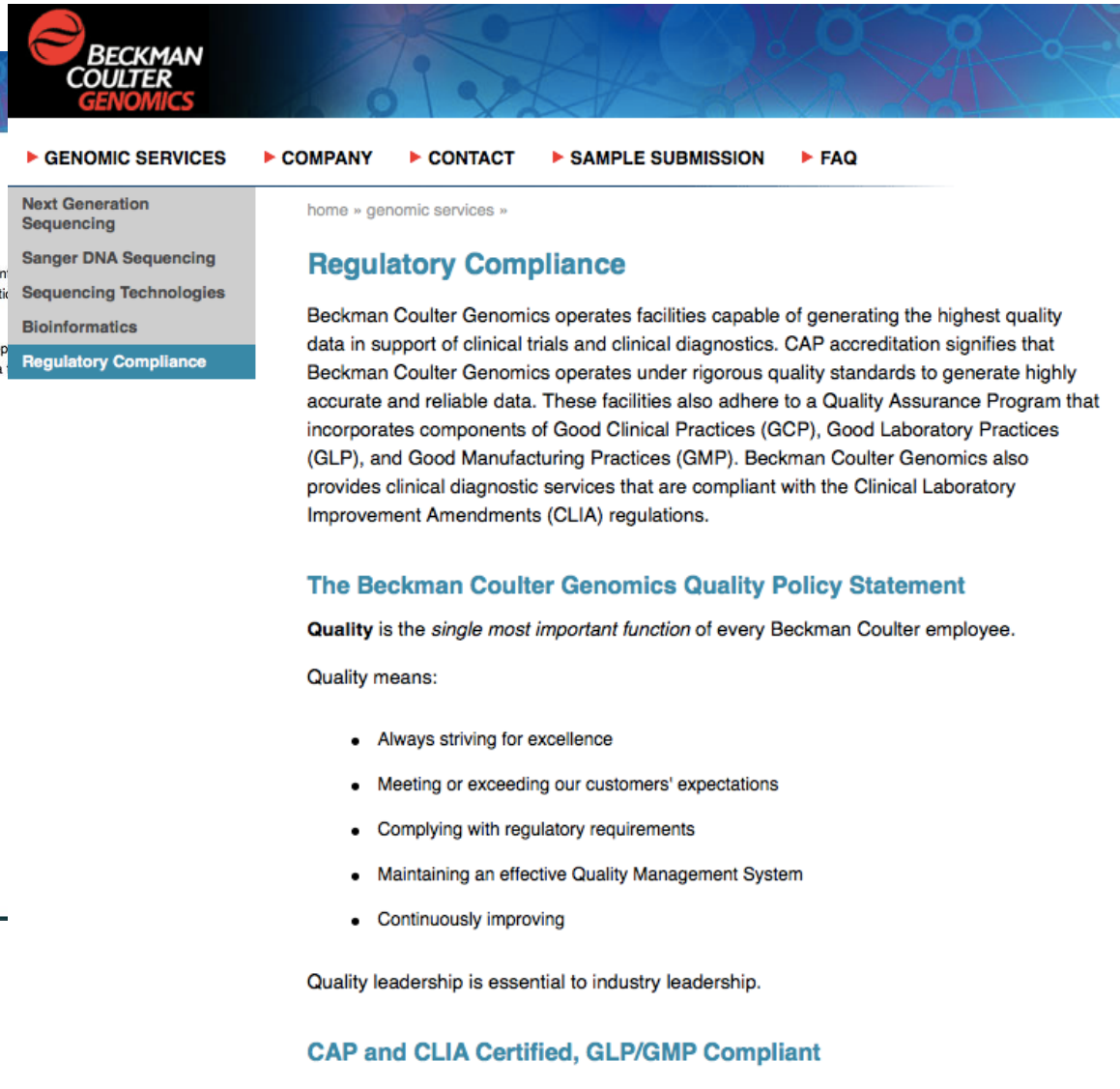
- Strong background of bioinformatics and science
- Extensive industry experience

### Secure Databases

- In-house copies of common sequence databases
- No analysis performed on public servers
- Results delivered via secure FTP site or shipped on portable hard drive

### Proprietary Pipelines

- Variant calling
- RNA-Seq and differential gene expression



**BECKMAN COULTER GENOMICS**

► GENOMIC SERVICES ► COMPANY ► CONTACT ► SAMPLE SUBMISSION ► FAQ

home » genomic services »

## Regulatory Compliance

Beckman Coulter Genomics operates facilities capable of generating the highest quality data in support of clinical trials and clinical diagnostics. CAP accreditation signifies that Beckman Coulter Genomics operates under rigorous quality standards to generate highly accurate and reliable data. These facilities also adhere to a Quality Assurance Program that incorporates components of Good Clinical Practices (GCP), Good Laboratory Practices (GLP), and Good Manufacturing Practices (GMP). Beckman Coulter Genomics also provides clinical diagnostic services that are compliant with the Clinical Laboratory Improvement Amendments (CLIA) regulations.

### The Beckman Coulter Genomics Quality Policy Statement

**Quality** is the *single most important function* of every Beckman Coulter employee.

Quality means:

- Always striving for excellence
- Meeting or exceeding our customers' expectations
- Complying with regulatory requirements
- Maintaining an effective Quality Management System
- Continuously improving

Quality leadership is essential to industry leadership.

**CAP and CLIA Certified, GLP/GMP Compliant**





# Data Analysis: Outsourcing



COMPANY

SER

## BIOINFORMATICS RESEARCH & SOLUTIONS

COMPANY

### ABOUT US

SciBerg is a private research company located near the city of Heidelberg, a center of European bioinformatics and life sciences. We integrate a group of PhD-holding scientists working in the world's leading research institutions and having profound expertise in experimental design and analysis of various high-throughput sequencing data. Some of our specialists have a proven record of prior developing the novel NGS library preparation methods (such as CATS Technology) as well as various data analysis pipelines.

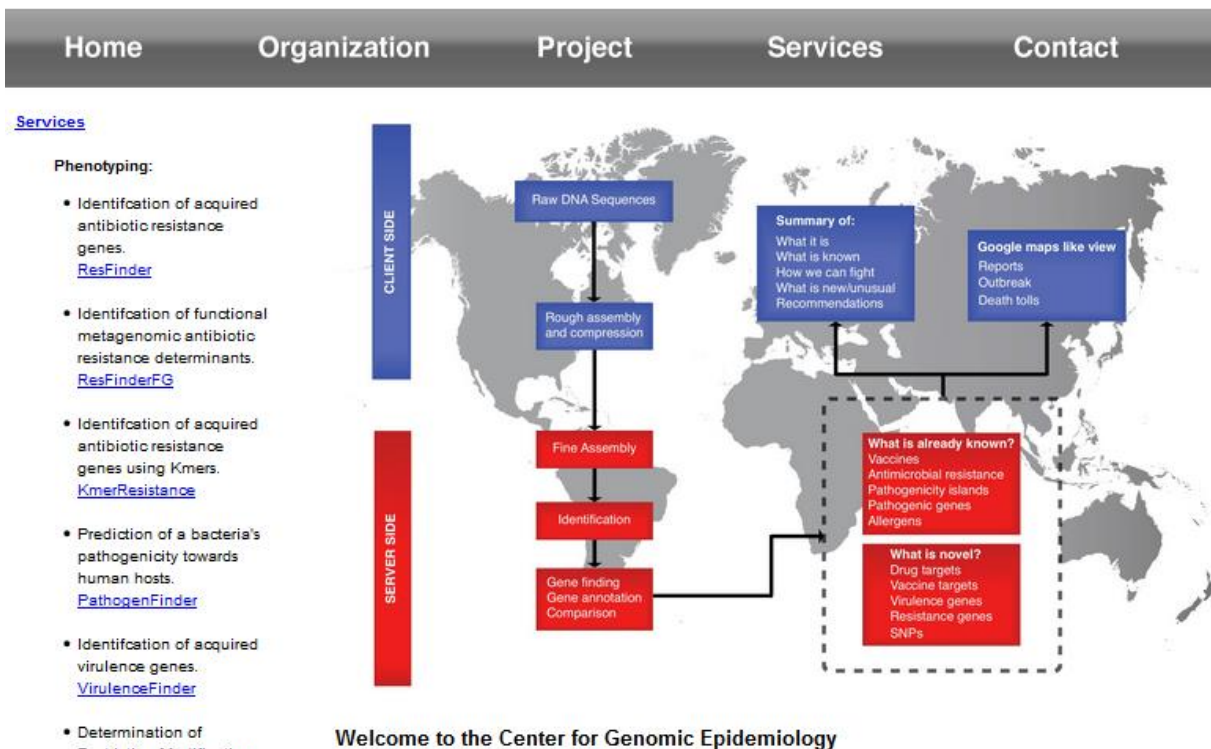
Besides providing commercial services, we participate in various fundamental and applied research projects together with academic and industrial partners on a not-for-profit basis. Our primary research interests are focused on developing novel experimental procedures and bioinformatic pipelines for precision medicine, as well as for non-invasive diagnosis/prognosis of human diseases using extracellular circulating nucleic acids.

# Data Analysis: Public servers

- Species identification
- *de novo* assembly tools
- VirulenceFinder
- SerotypeFinder
- ResFinder
- MLST
- SNPs tree and newly developed NGS-driven phylogenetic tools

**FREE, USER-FRIENDLY WEB INTERFACE**

## Center for Genomic Epidemiology



**Closed Public server**



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# Data Analysis: Public servers



- *de novo* assembly tools
- BLAST search of genes of interest
- Alignment of sequences, typing tools, production of dendrograms



ARIES - ISS

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors. The Galaxy Project is supported in part by UNISRI, ISE, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

OPEN SOURCE, USER-FRIENDLY WEB INTERFACE, OPEN FOR INTRODUCTION OF CUSTOMIZED TOOLS, ELECTION PLATFORM FOR DEVELOPING AND SHARING OF NEW TOOLS




Open Public server



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# ARIES: A Galaxy-based workspace for intensive data analyses

 **Istituto Superiore di Sanita'**


ARIES - Advanced Research Infrastructure for Experimentation in Genomics - Galaxy Instance at ISS



Tweet di @ARIES\_GENOMICS



Please read our [disclaimer](#) before using ARIES.

 - FTP is now available for data upload at [ariesftp.iss.it](http://ariesftp.iss.it) (explicit FTP over TLS)

- Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#). The [Galaxy Project](#) is supported in part by [NHGRI](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Johns Hopkins University](#).

History

search datasets

**O26comparison**  
1179 shown, 2160 deleted, 378 hidden  
35 GB

1st of 3 pages

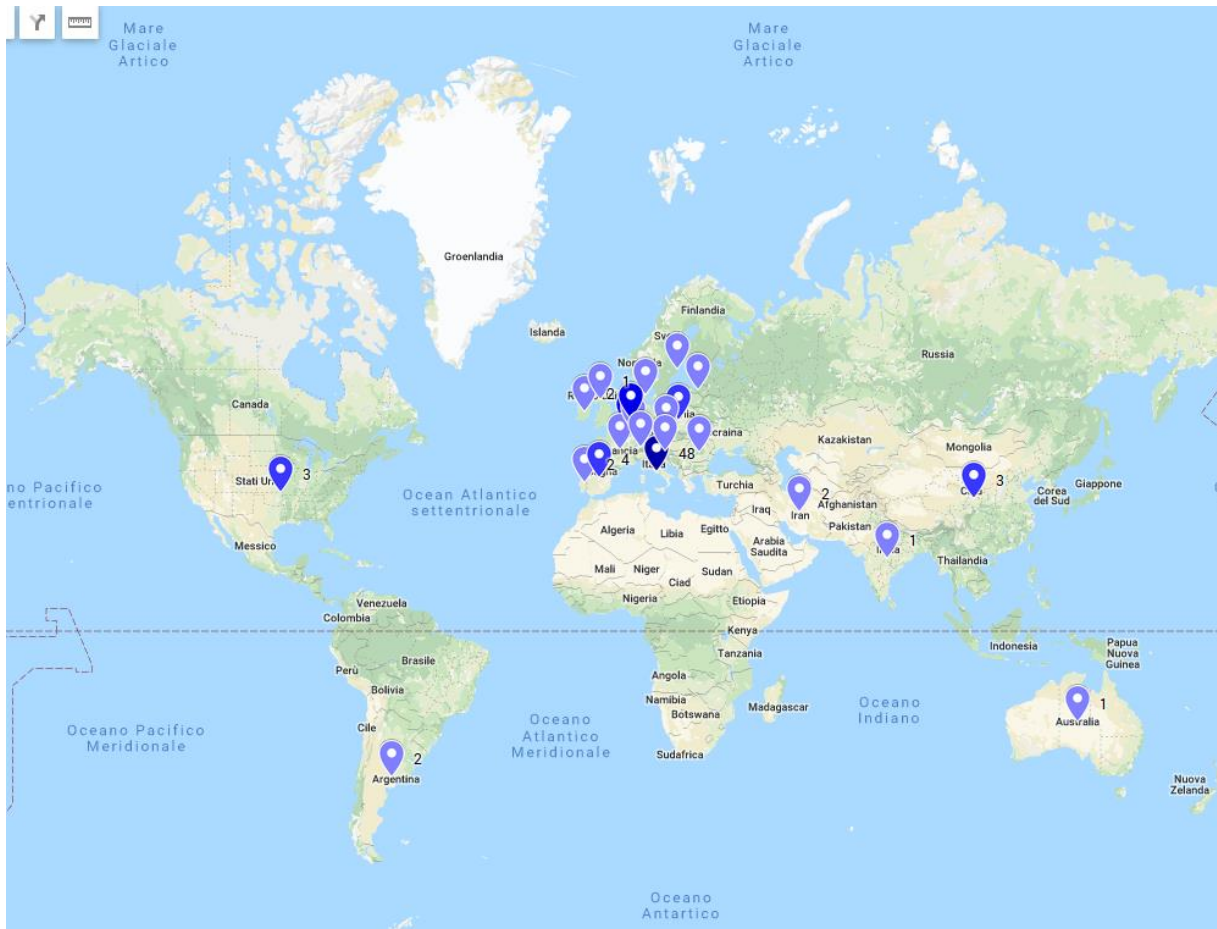
- 3565: [SNPs\\_all\\_matrix.fasta](#)
- 3564: [tree\\_tipAlleleCounts.ML.tre](#)
- 3563: [tree\\_AlleleCounts.ML.NodeLabel.tre](#)
- 3562: [tree\\_AlleleCounts.ML.tre](#)
- 3561: [tree.ML.tre](#)
- 3560: [tree\\_tipAlleleCounts.parsimony.tre](#)
- 3559: [tree\\_AlleleCounts.parsimony.No deLabel.tre](#)
- 3558: [tree\\_AlleleCounts.parsimony.tre](#)
- 3557: [tree.parsimony.tre](#)
- 3556: [O26\\_paper\\_Acilia](#)
- 3555: [aqMLST Log File](#)
- 3554: [aqMLST New Alleles File](#)
- 3553: [aqMLST](#)



Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health  
European Union and National Reference Laboratory for *E. coli*, Rome, Italy



# ARIES geographic spread

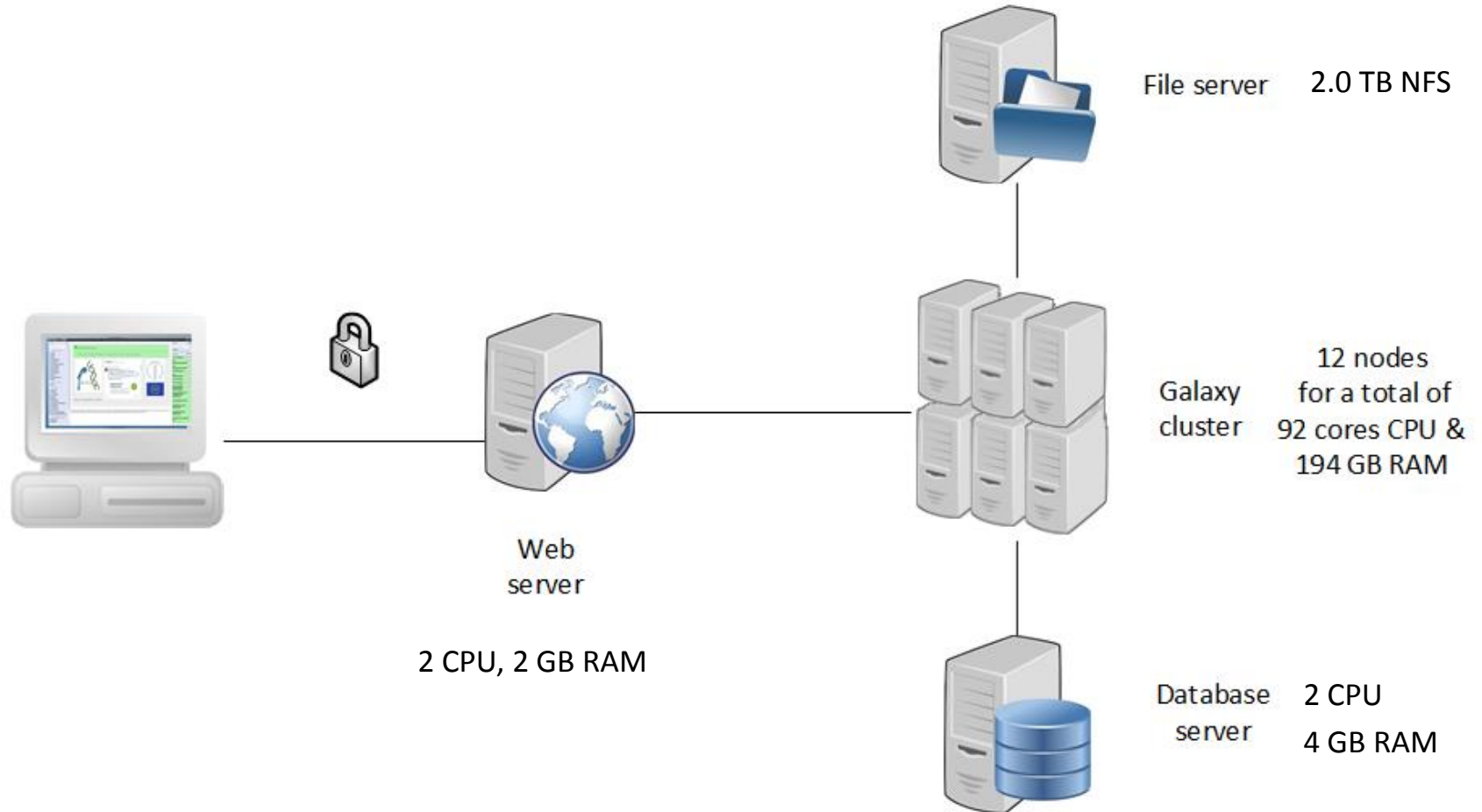


**102 total Users**

**90 European  
Users (15 NRLs)**

**12 Users from  
outside EU**

# ARIES Under the hood



Tools

--- COMMON TOOLS ---

[Get Data](#)[Send Data](#)[Lift-Over](#)[Text Manipulation](#)[Filter and Sort](#)[Join, Subtract and Group](#)[Convert Formats](#)[Extract Features](#)[Fetch Sequences](#)[Fetch Alignments](#)[Statistics](#)[Graph/Display Data](#)[GraPhlAn](#)

---HREVAP TOOLS---

[HReVAP](#)

---NGS TOOLS---

[In Silico PCR](#)[E coli typing](#)[NGS: Assembly](#)[NCBI Blast](#)[Manipulation](#)[Gene Annotation](#)[FASTA/FASTQ manipulation](#)[NGS: Mapping](#)[NGS: SAM Tools](#)[NGS: BED Tools](#)[NGS: QC and manipulation](#)[Operate on Genomic Intervals](#)

---METAGENOMICS TOOLS---



Istituto Superiore di Sanita'

ARIES - Advanced Research Infrastructure for Experi

Please read our [disclaimer](#) before using ARIES.

QC

Assembly de novo

Mapping

Microbial genome annotation

NCBI Databases

Databases shared with  
CGE/SSIs

Custom Databases

***E. coli* typing:**

Virulotyping

Serotyping

Clermont phylogrouping

HReVAP

MLST

ksnp3 for ref-free wgSNPs

cgMLST