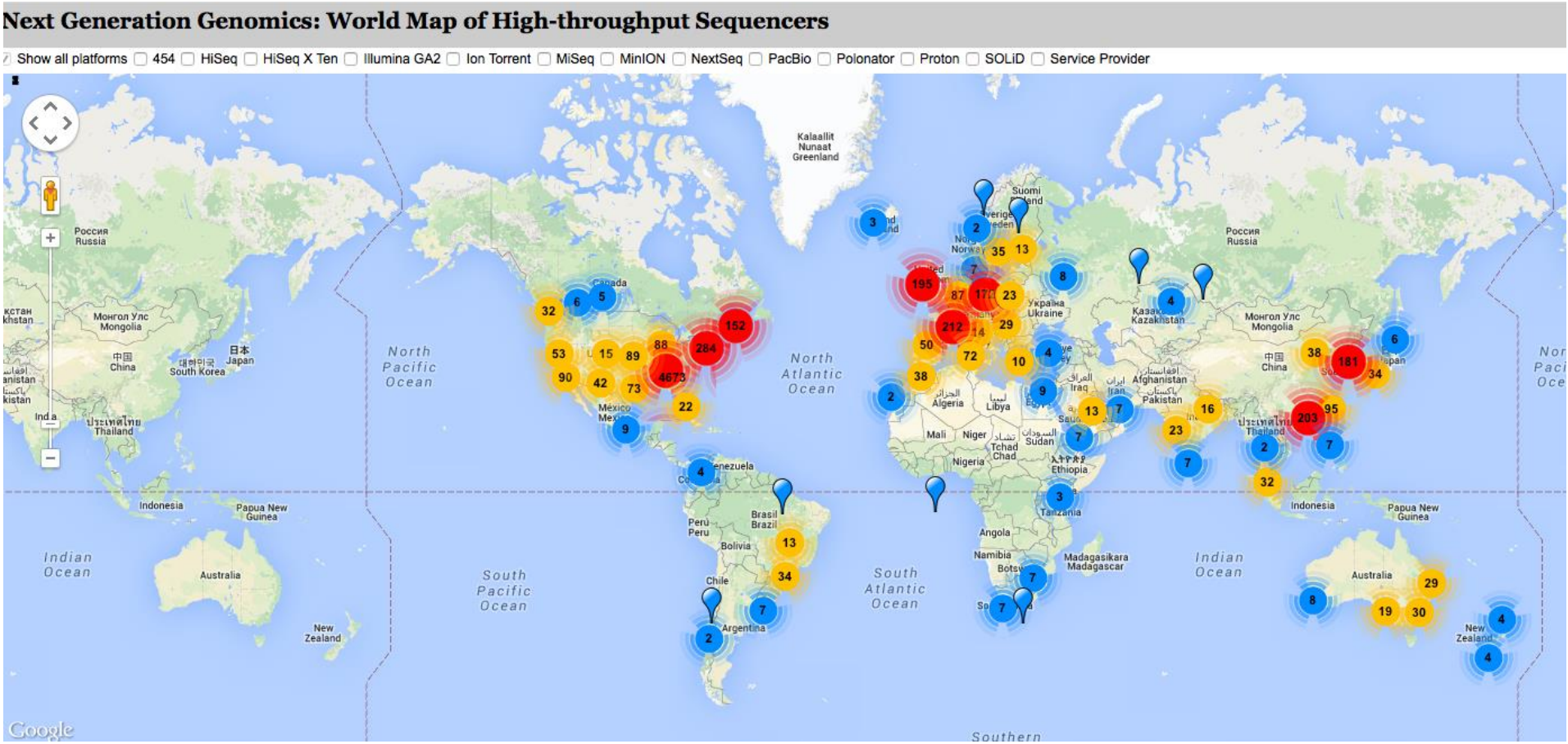
Data Intensive Biomedical Research: The EU RL VTEC efforts to take up the NGS challenge

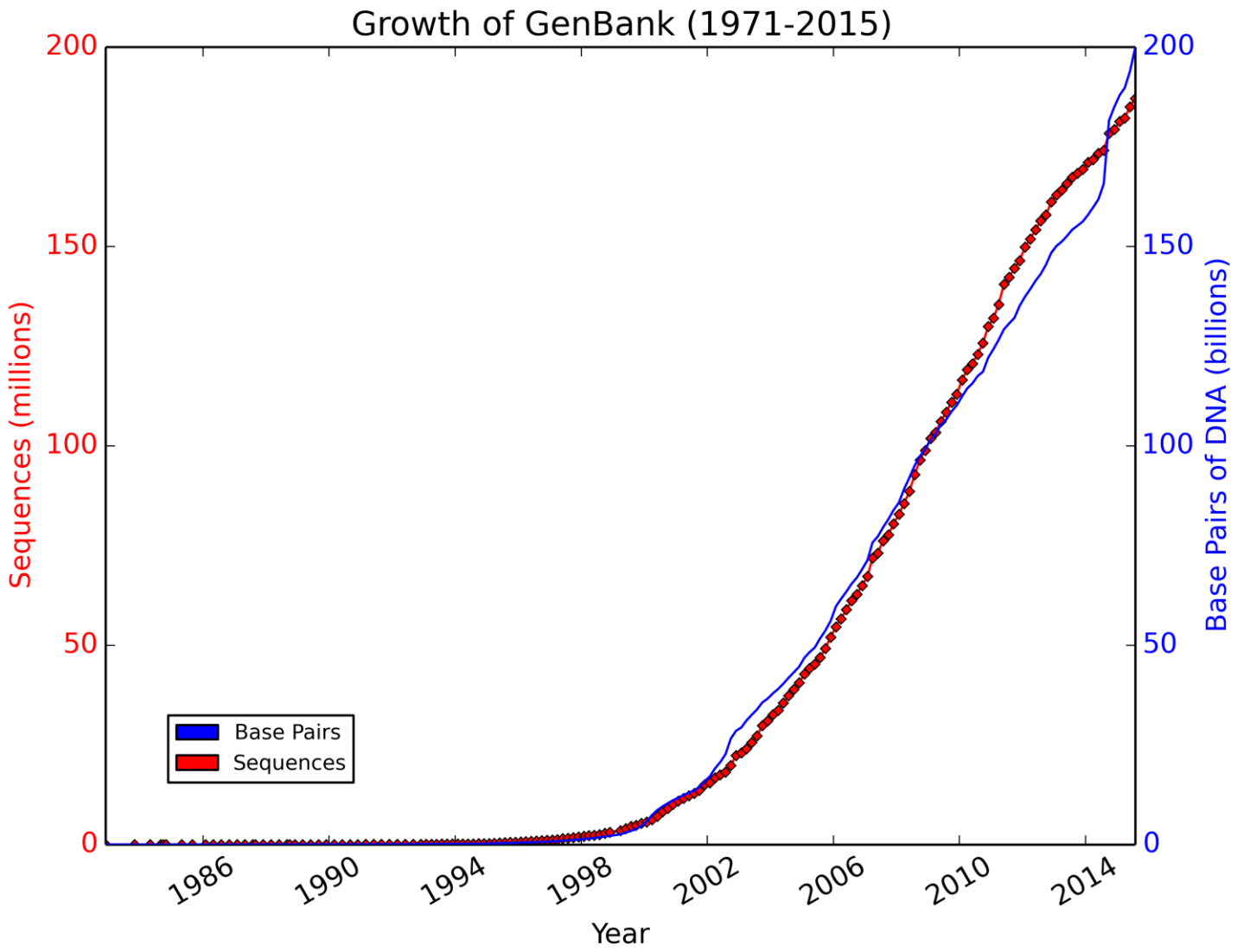
EU RL for *E. coli* Annual Workshop
2015

NGS adoption: Worldwide



Source: Omicsmap.com November, 2015

Data Production rate



Source: http://genome.ist.unomaha.edu/cgi-bin/genbank.cgi

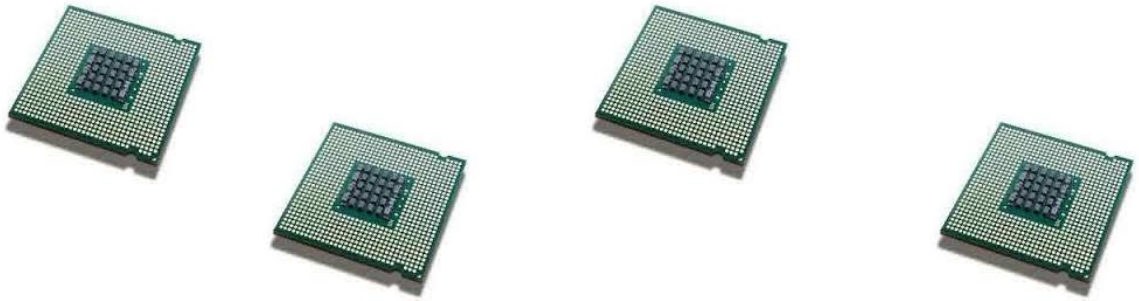
“The system has failed” paradox

Found on twitter:

Data analysis: The Black Hole

```
Chr1    3395973 rs143478237     ACC     A       83      PASS    DB;DP=9;HP=5;NF=3;NFS=3;NR=0;NRS=0;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=ARHGEF16     GT:GQ   1/1:9
Chr1    3545175 rs147637374     GTTCTGGGAGCTCCTCCCCC    G       141     PASS    DB;DP=55;HP=2;NF=2;NFS=5;NR=5;NRS=13;VCQ=SPLICE_SITE:3PRIME_UTR;VCQNC=UPSTREAM;VGN=TPRG1L       GT:GQ   0/1:99
Ingos-MacBook-Pro:NLIS3_4 ih$ sed 50q  2.vcf
##fileformat=VCFv4.1
##ALT=<ID=DEL,Description="Deletion">
##FILTER=<ID=fr0,Description="Non-ref allele is not covered by at least one read on both strands">
##FILTER=<ID=hp10,Description="Reference homopolymer length was longer than 10">
##FILTER=<ID=q20,Description="Quality below 20">
##FILTER=<ID=wv,Description="Other indel in window had higher likelihood">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total number of reads in haplotype window">
##INFO=<ID=HP,Number=1,Type=Integer,Description="Reference homopolymer tract length">
##INFO=<ID=NF,Number=1,Type=Integer,Description="Number of reads covering non-ref variant on forward strand">
##INFO=<ID=NFS,Number=1,Type=Integer,Description="Number of reads covering non-ref variant site on forward strand">
##INFO=<ID=NR,Number=1,Type=Integer,Description="Number of reads covering non-ref variant on reverse strand">
##INFO=<ID=NRS,Number=1,Type=Integer,Description="Number of reads covering non-ref variant site on reverse strand">
##VariantAnnotator="analysis_type=VariantAnnotator input_file=[] sample_metadata=[] read_buffer_size=null phone_home=STANDARD read_filter=[] intervals=null excludeIntervals=null reference_sequence=/lustre
/scratch103/ensembl/ces/ref/human_g1k_v37.fasta rodBind=[/lustre/scratch101/ensembl/ces/NEW//mapping_live/output/27299/27299_Vant1_603939.vcf, /lustre/scratch103/ensembl/ces/data/snps/dbSNP/dbsnp134_20110
816_SureSelect_All_Exon_50Mb_CTRplus.vcf.gz] rodToIntervalTrackName=null BTI_merge_rule=UNION nonDeterministicRandomSeed=false DBSNP=null downsampling_type=null downsample_to_fraction=null downsample_to_c
overage=null baq=OFF baqGapOpenPenalty=40.0 performanceLog=null useOriginalQualities=false defaultBaseQualities=-1 validation_strictness=SILENT unsafe=null num_threads=1 interval_merging=ALL read_group_bl
ack_list=null processingTracker=null restartProcessingTracker=false processingTrackerStatusFile=null processingTrackerID=-1 allow_intervals_with_unindexed_bam=false disable_experimental_low_memory_shardin
g=false logging_level=ERROR log_to_file=null help=false out=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub NO_HEADER=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub sites_only=org.broadinstitut
e.sting.gatk.io.stubs.VCFWriterStub sampleName=null annotation=[] group=[] expression=[] useAllAnnotations=false list=false assume_single_sample_reads=null vcfContainsOnlyIndels=false"
##reference=/lustre/scratch103/ensembl/ces/ref//Homo_sapiens_GRCh37_53.fasta
##source=Dindel
##INFO=<ID=VCQNC,Number=1,Type=String,Description="Highest consequences (from ensembl VEP) - non coding genes (if available and only transcripts without a translation are included). Highest splice consequ
ence : (if applicable), followed by most servere consequence.">
##INFO=<ID=VPI,Number=1,Type=String,Description="Protein id (from ensembl VEP) associated with the most severe consequence in the VCQ field (if applicable). Note one is selected.">
##INFO=<ID=Condel,Number=1,Type=String,Description="Condel SIFT/PolyPhen consensus (from ensembl VEP) for the most severe consequence in the VCQ field (if applicable).">
##INFO=<ID=VGN,Number=1,Type=String,Description="Gene name (from ensembl VEP) associated with the most severe consequence in the VCQ field.">
##INFO=<ID=VAA,Number=1,Type=String,Description="Amino acids (from ensembl VEP) associated with the most severe consequence in the VCQ field (if applicable).">
##INFO=<ID=PolyPhen,Number=1,Type=String,Description="PolyPhen prediction (from ensembl VEP) for the most severe consequence in the VCQ field (if applicable).">
##INFO=<ID=VCQ,Number=1,Type=String,Description="Highest consequences (from ensembl VEP) - coding genes (if available and only transcripts with a translation are included). Highest splice consequence : (i
f applicable), followed by most servere consequence.">
##INFO=<ID=SIFT,Number=1,Type=String,Description="SIFT prediction (from ensembl VEP) for the most severe consequence in the CQ field (if applicable).">
##INFO=<ID=VPP,Number=1,Type=String,Description="Position in protein in the VPI field (from ensembl VEP) associated with the most severe consequence in the VCQ field (if applicable).">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SC_PND5230406
Chr1    866511  rs146519568     C       CCCCT   13      q20     DB;DP=34;HP=4;NF=0;NFS=1;NR=1;NRS=7;VCQ=INTRONIC;VGN=SAMD11     GT:GQ   0/1:13
Chr1    874950  rs149166309     T       TCCCTGGAGGACC   148     PASS    DB;DP=12;HP=3;NF=0;NFS=0;NR=4;NRS=4;VCQ=INTRONIC;VCQNC=UPSTREAM;VGN=SAMD11      GT:GQ   1/1:21
Chr1    907763  .       G       GTC     1       q20     DP=29;HP=4;NF=0;NFS=0;NR=2;NRS=3;VCQ=FRAMESHIFT_CODING;VCQNC=UPSTREAM;VGN=PLEKHN1;VPI=ENSP00000368719;VPP=373   GT:GQ   0/1:1
Chr1    948846  rs3841266       T       TA      1013    PASS    DB;DP=48;HP=1;NF=23;NFS=23;NR=1;NRS=1;VCQ=5PRIME_UTR;VCQNC=UPSTREAM;VGN=ISG15    GT:GQ   1/1:75
Chr1    957967  rs141489152     T       TTGTAGTCTGACCTGTGGTCTGAC        141     PASS    DB;DP=20;HP=2;NF=1;NFS=1;NR=1;NRS=2;VCQ=INTRONIC;VGN=AGRN       GT:GQ   1/1:28
Chr1    970549  rs56001364      TGG     T       96      hp10    DB;DP=254;HP=13;NF=12;NFS=22;NR=2;NRS=2;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=AGRN    GT:GQ   1/1:11
Chr1    978603  rs146114193     CCT     C       91      PASS    DB;DP=14;HP=4;NF=3;NFS=4;NR=0;NRS=2;VCQ=INTRONIC;VCQNC=UPSTREAM;VGN=AGRN        GT:GQ   0/1:91
Chr1    984171  rs140904842     CAG     C       7       q20     DB;DP=5;HP=1;NF=1;NFS=1;NR=0;NRS=0;VCQ=INTRONIC;VCQNC=WITHIN_NON_CODING_GENE;VGN=AGRN   GT:GQ   0/1:7
Chr1    999041  rs144946318     ATG     A       13      q20     DB;DP=5;HP=1;NF=1;NFS=2;NR=0;NRS=0;VCQNC=INTRONIC       GT:GQ   1/1:5
Chr1    1158534 rs59317408      G       GAC     119     PASS    DB;DP=25;HP=1;NF=4;NFS=7;NR=0;NRS=1;VCQ=INTRONIC;VCQNC=UPSTREAM;VGN=SDF4        GT:GQ   0/1:99
Chr1    1158562 rs139833693     AAC     A       119     PASS    DB;DP=47;HP=2;NF=5;NFS=10;NR=0;NRS=1;VCQ=INTRONIC;VCQNC=UPSTREAM;VGN=SDF4       GT:GQ   0/1:99
Chr1    1276973 rs145370195     G       GACAC   473     PASS    DB;DP=34;HP=1;NF=7;NFS=9;NR=2;NRS=2;VCQ=INTRONIC;VCQNC=DOWNSTREAM;VGN=DVL1      GT:GQ   1/1:6
Chr1    1289367 rs140777846     CTG     C       1102    PASS    DB;DP=84;HP=1;NF=16;NFS=16;NR=8;NRS=8;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=MXRA8     GT:GQ   1/1:72
Chr1    1323143 rs147796530     CCT     C       430     PASS    DB;DP=14;HP=3;NF=8;NFS=8;NR=1;NRS=1;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CCNL2       GT:GQ   1/1:30
Chr1    1325493 .       CA      C       23      hp10    DP=46;HP=20;NF=8;NFS=11;NR=0;NRS=0;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CCNL2        GT:GQ   1/1:16
Chr1    1588744 rs79724854      AGCG    A       133     PASS    DB;DP=599;HP=1;NF=1;NFS=1;NR=20;NRS=29;VCQ=DOWNSTREAM;VCQNC=INTRONIC;VGN=SLC35E2B       GT:GQ   0/1:99
Chr1    1633003 .       C       CCG     154     PASS    DP=5;HP=9;NF=4;NFS=4;NR=1;NRS=1;VCQ=DOWNSTREAM;VCQNC=INTRONIC;VGN=CDK11A        GT:GQ   1/1:15
Chr1    1647649 rs70937179      C       CT      410     PASS    DB;DP=31;HP=1;NF=10;NFS=10;NR=0;NRS=0;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CDK11A    GT:GQ   1/1:29
Chr1    1647893 rs144636354     C       CTTTCTT 1134    PASS    DB;DP=236;HP=3;NF=1;NFS=3;NR=30;NRS=33;VAA=R/KER;VCQ=NON_SYNONYMOUS_CODING;VCQNC=WITHIN_NON_CODING_GENE;VGN=CDK11A;VPI=ENSP00000422149;VPP=1
27      GT:GQ   0/1:99
Chr1    1647968 rs146207119     C       CAT     7       q20     DB;DP=115;HP=1;NF=3;NFS=4;NR=5;NRS=32;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CDK11A    GT:GQ   0/1:7
Chr1    1650639 .       CCA     C       35      PASS    DP=124;HP=2;NF=2;NFS=15;NR=0;NRS=0;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CDK11A       GT:GQ   0/1:35
Chr1    1653332 rs34272957      GT      G       5       q20     DB;DP=8;HP=2;NF=0;NFS=0;NR=1;NRS=1;VCQ=INTRONIC;VCQNC=INTRONIC;VGN=CDK11A       GT:GQ   1/1:4
Ingos-MacBook-Pro:NLIS3_4 ih$
```

Computationally intensive applications



The assembly of a bacterial genome (approx 600 MB) takes approx 45mins and completely occupies the computation capacity of a processor (or a core of a quad-cores processor)

The assembly of a bacterial genome (approx 600 MB) through a complete pipeline (e.g. A5) takes up to three hours on a stand-alone workstation and takes approx 130% of the computation capacity of a core of a quad-cores processor

The assembly of a metagenomics sample (up to 3.2 GB) may take days and in some cases it will not be assembled at all (regardless the number of cores available)

Data analysis: Locally running softwares suites

BioNumerics

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

Private company

USER-FRIENDLY INTERFACE, Slow processing, RAM needed

Torrent Suite Software

ion torrent

by life technologies

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

Private company

BUILT IN THE ION TORRENT TECHNOLOGY PACKAGE

EU-RL VTEC

Data analysis: web servers

Center for Genomic Epidemiology

- Species identification
- *de novo* assembly tools
- VirulenceFinder
- ResFinder
- MLST
- SNPs tree and newly deleveloped NGS-driven philogenetic tools

Closed Public server

FREE, USER-FRIENDLY WEB INTERFACE

Galaxy

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

Open Public server

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

EU-RL VTEC

E. coli network and NGS: State of the Art

8 NRLs already access to benchtop sequencers

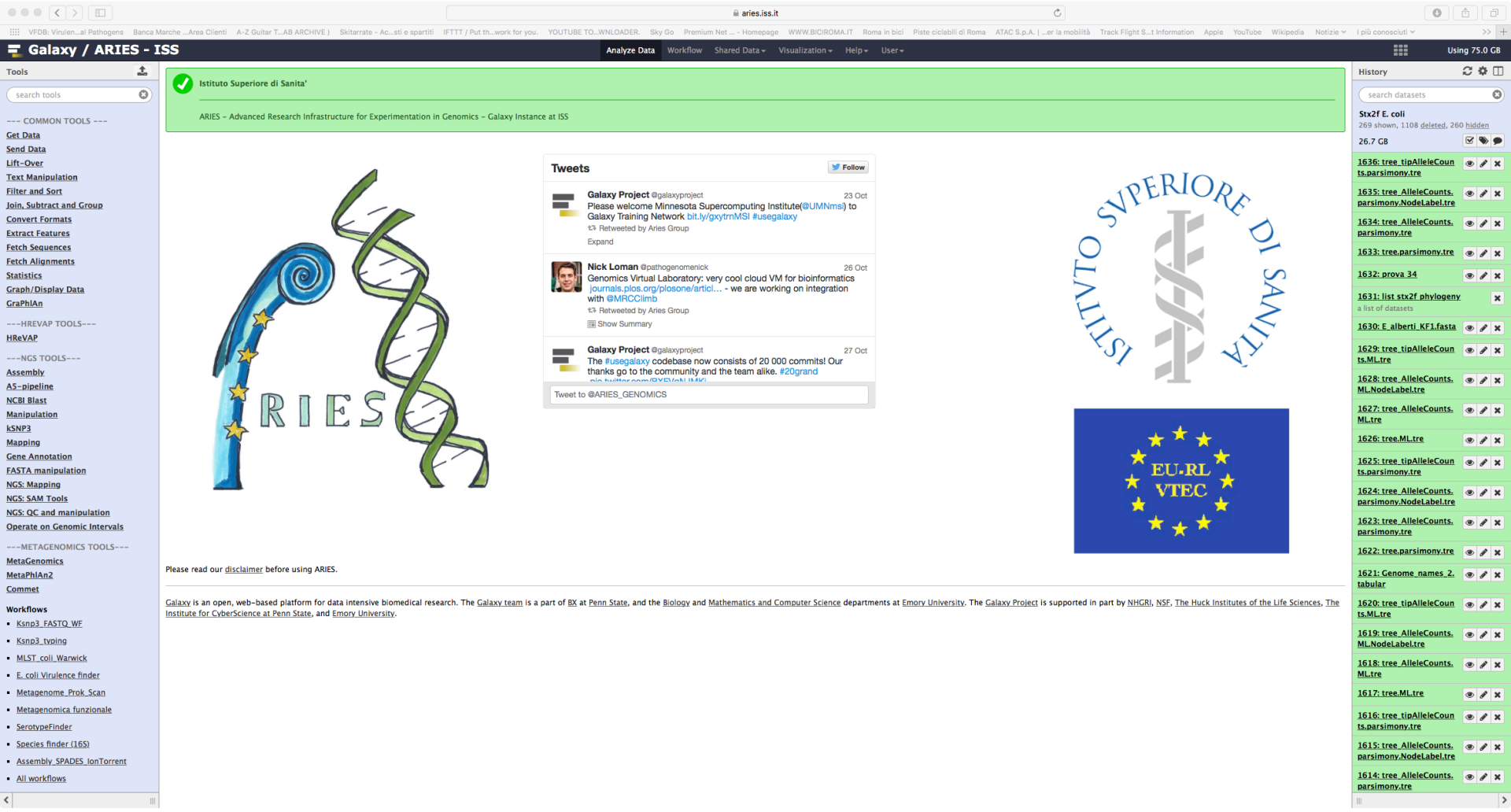
8 NRLs plan to outsource NGS data production

Capacity building is at the beginning

Many others will follow in the next few years!!

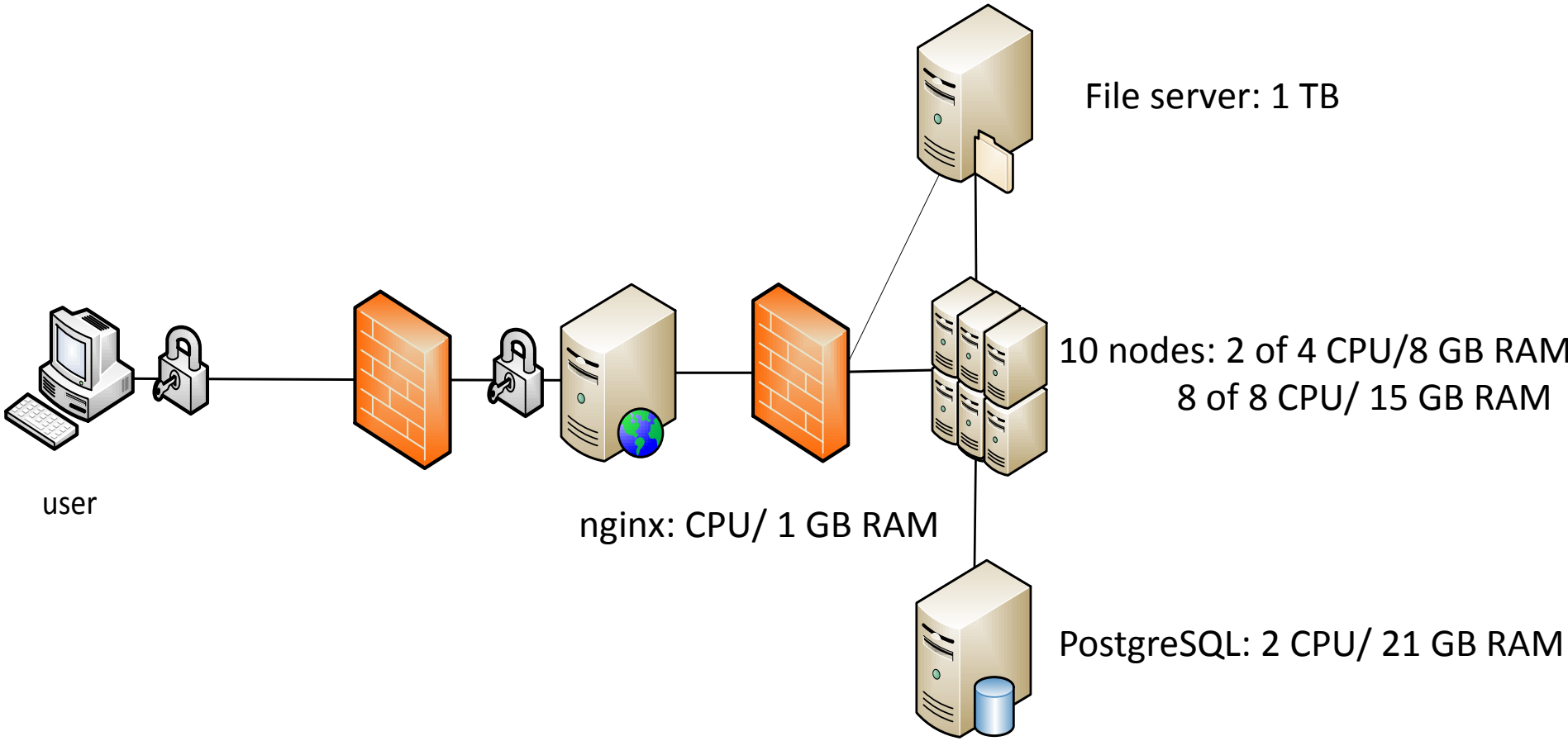
All the NRLs expressed the need for education in genomics data analysis by replying to a questionnaire

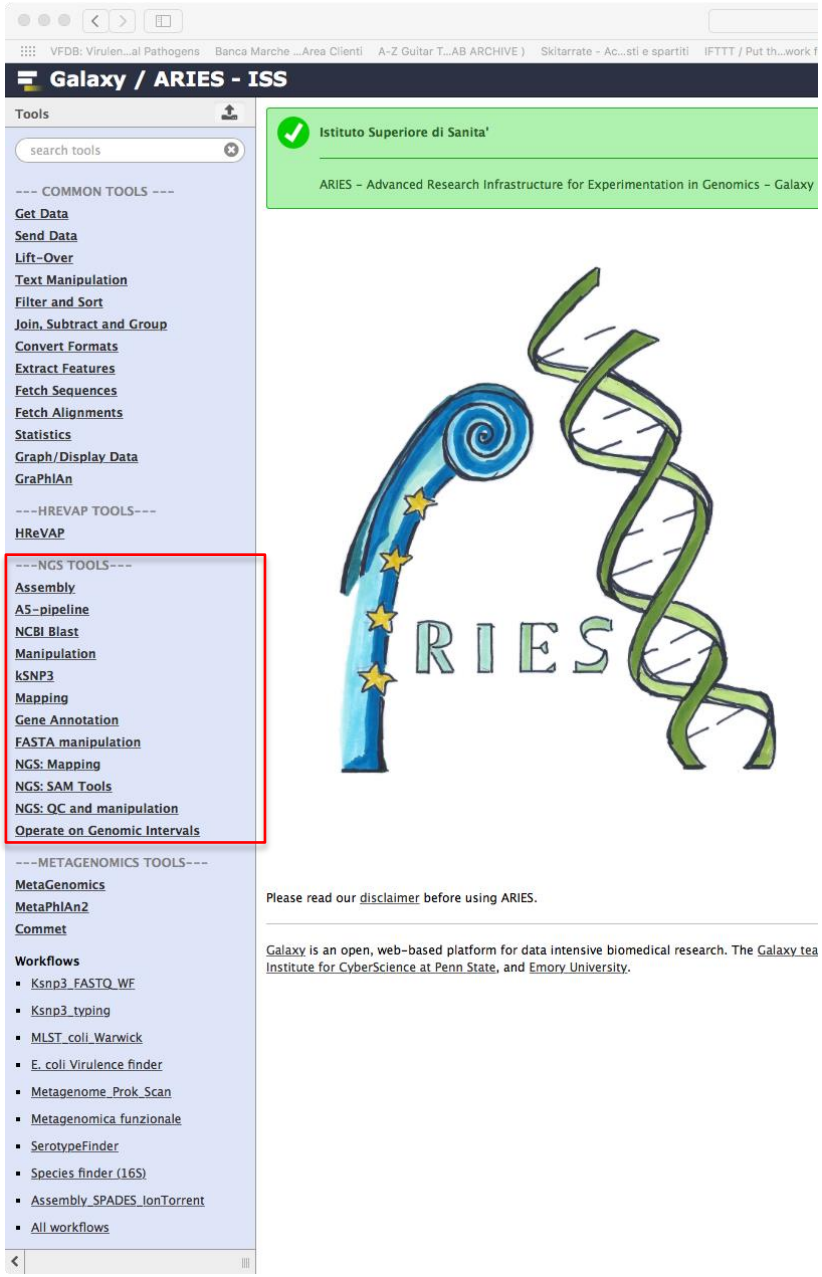
ARIES: A Galaxy-based workspace for intensive data analyses



Opened to the public on October the 1°, 2015

ARIES Under the hood





QC:
Reads scan
Contigs stats
Trim (quality and positional)

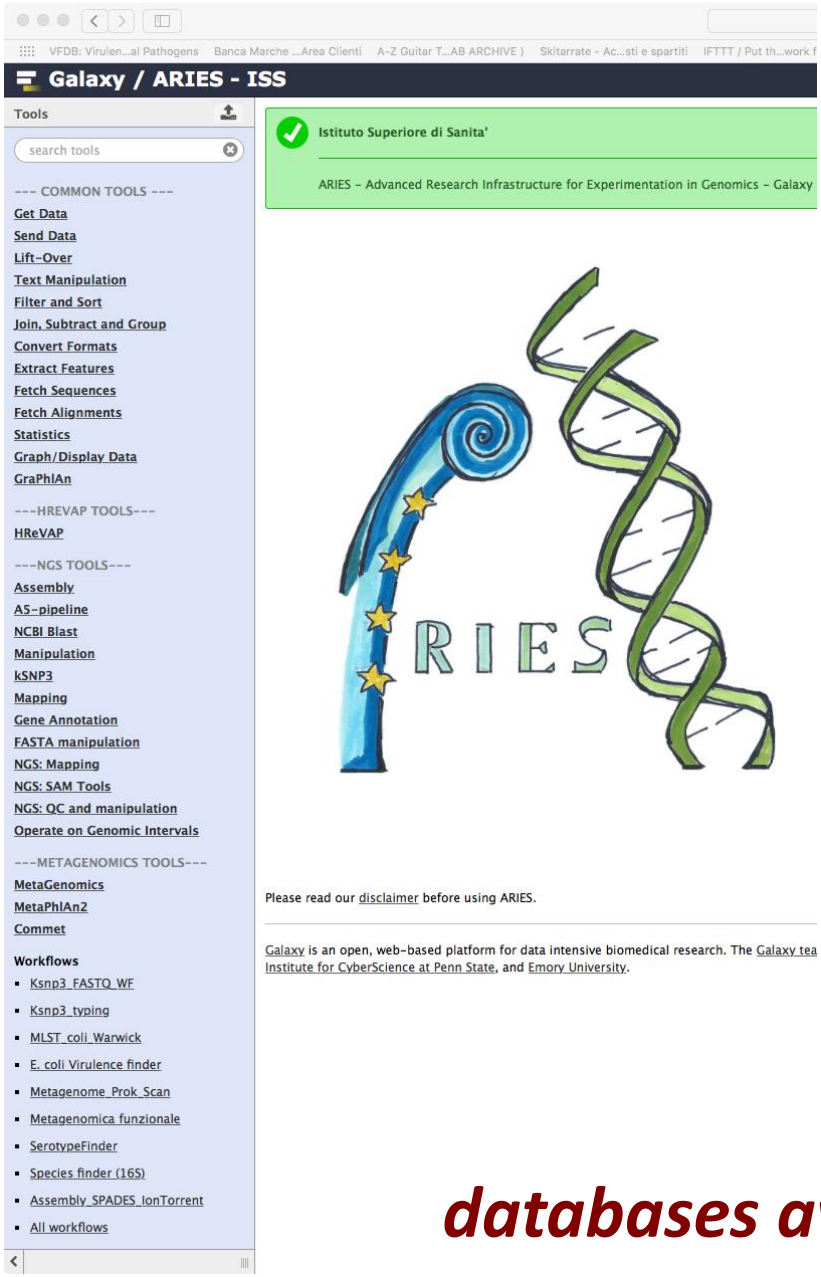
Assembly de novo:
SPADES
Velvet
VELVET/Optimizer
EDENA

Mapping:
Bowtie2
BWA-MEM

Microbial genome annotation:
PROKKA

Genomes searches:
Full BLAST suite

Basic NGS data management tools



Genomes BLAST searches: pre-installed Dbases:

NCBI Databases

- 16S rRNA
- Bacterial
- Plasmids
- Viruses

Databases shared with CGE/SSIs

- Enteroaggregative E. coli
- E. coli Virulence genes
- O/H antigens E. coli genes

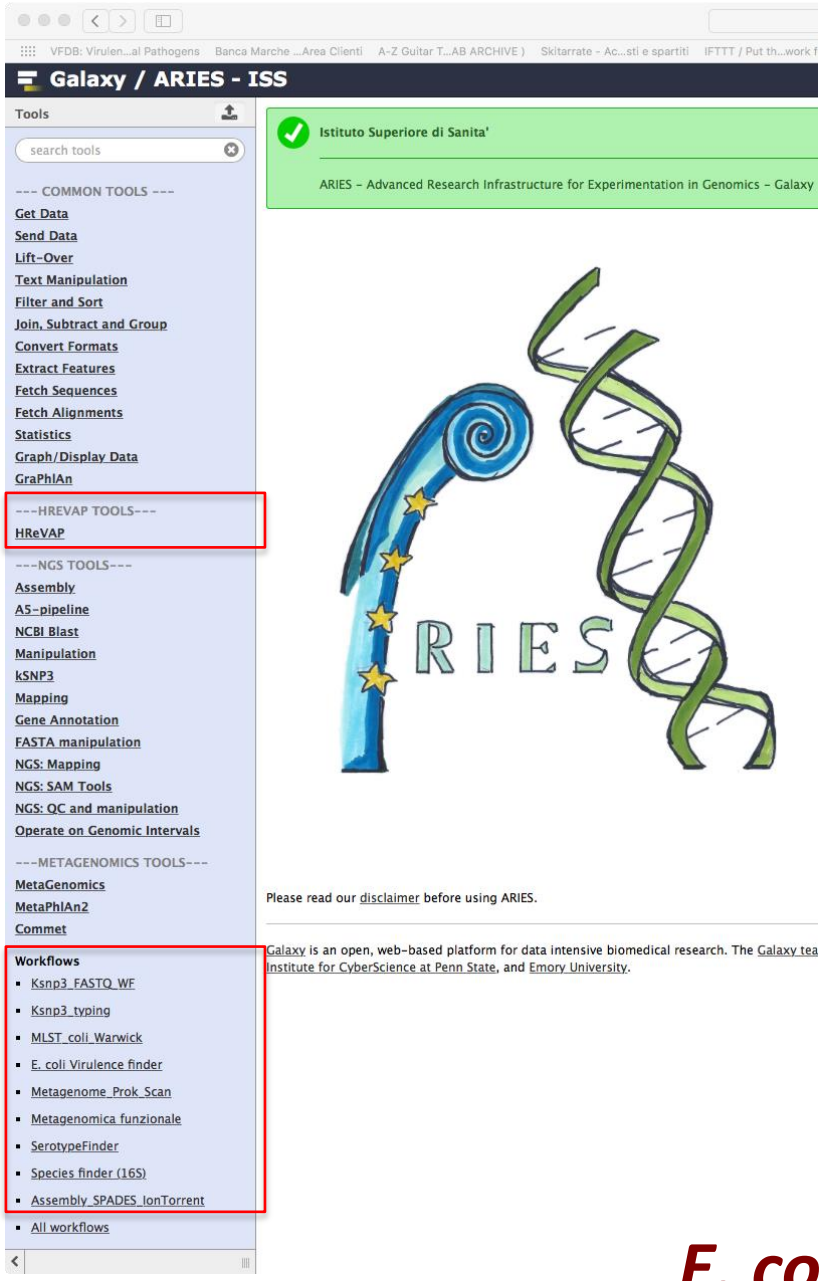
Databases shared with Warwick University

- E. coli MLST

Custom Databases

- Shigella/EIEC Virulence genes
- HReVAP Alleles
- Clermont Phylogrouping

databases available (counting….)



Custom workflows:

E. coli Serotyper

E. coli Virulotyper

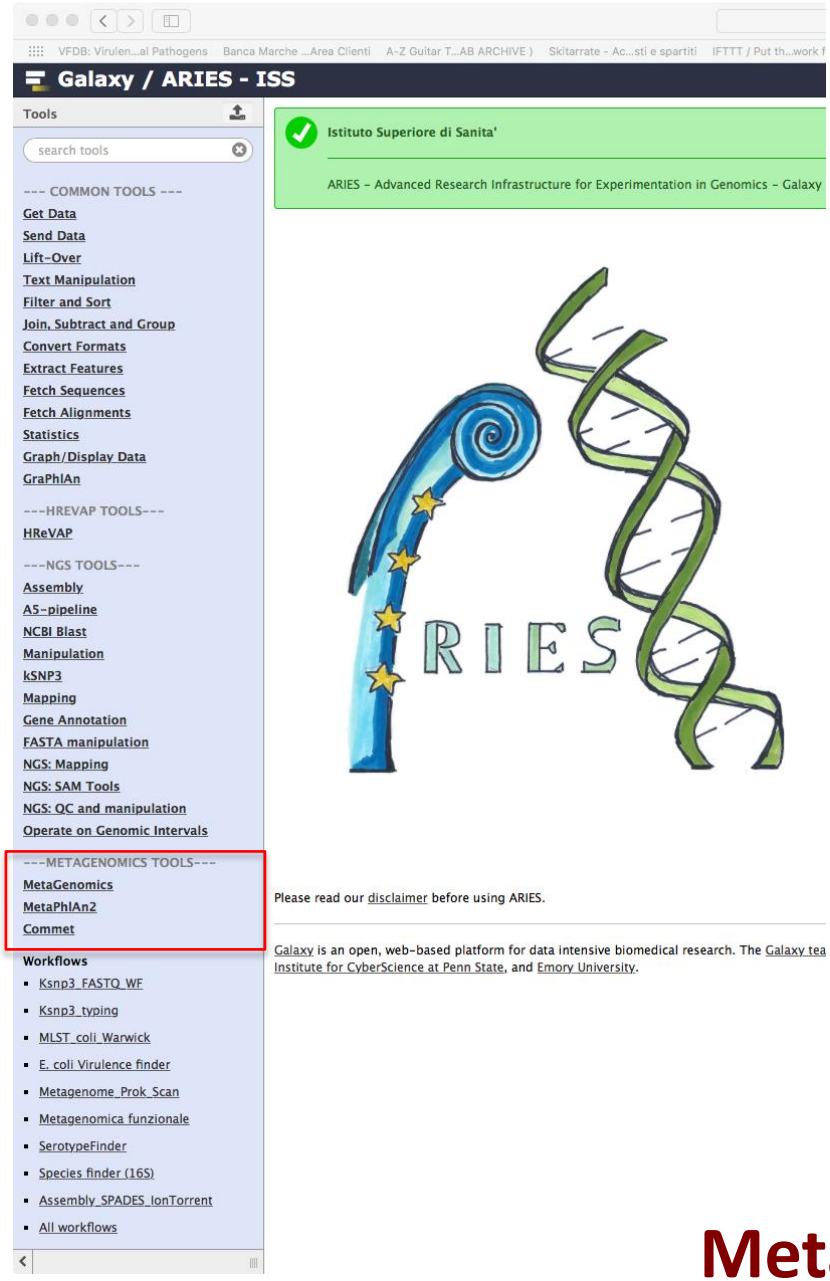
MLST:

Warwick University DB

HReVAP:

Completely new. Developed at the EU RL for *E. coli* on ARIES. (It will be shared on the Galaxy toolshed upon validation)

E. coli typing



Already on the menu:

MetaPhlAn

Qiime

Krona Tools

Blast wokflows (16S, Viral proteins, Viral genomes)

Coming soon:

MetaPhlAn2 (viral metage

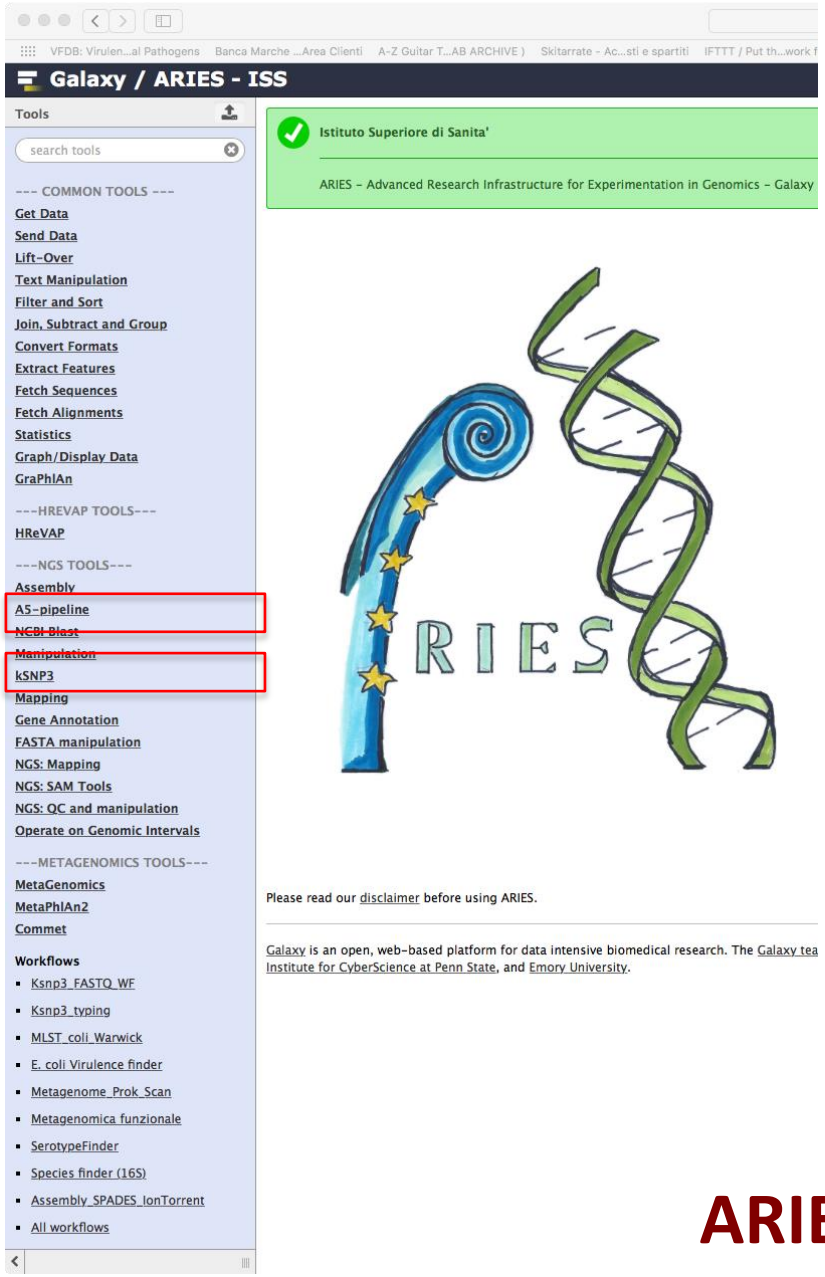
DONE!

Newly develped (under coc writing):

Functional annotation

DONE!

Metagenomics



Assembly pipeline specific for Illumina redas:

A5

SNPs analysis:

Ksnp3 Pipeline

SNPs analysis:

FDA pipeline

ARIES new features

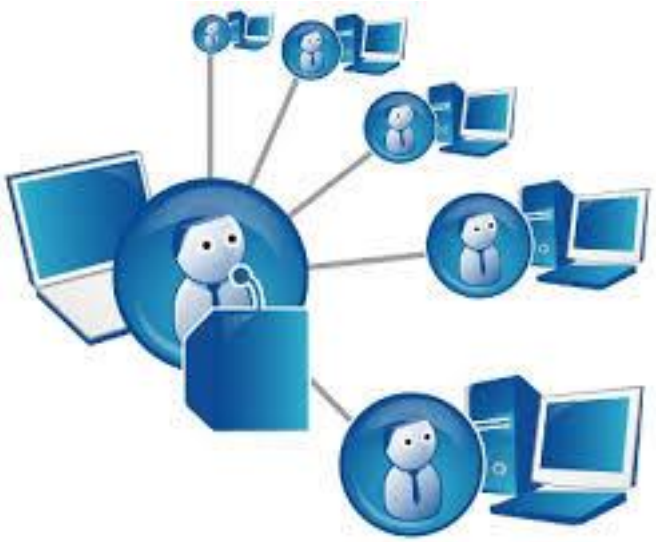
Training at the EU RL for *E. coli* on NGS data anlysis:

EU Reference Laboratory for *E. coli*
Department of Veterinary Public Health and Food Safety
Unit of Foodborne Zoonoses
Istituto Superiore di Sanità

Basic Course on Bioinformatics tools
for Next Generation Sequencing data mining

11-12 June, 2015

SIDBAE Training Room
(*Building 1, Floor B*)
Istituto Superiore di Sanità
Viale Regina Elena, 299 – Rome, Italy



Web-based training modules in the following years

A residential training course on NGS-based *E. coli* typing in 2016

ARIES: Policy

Ask for an account:

- E-mail aries@iss.it

Collaborative projects on *E. coli* :

- E-mail stefano.morabito@iss.it
- E-mail valeria.michelacci@iss.it

ARIES is a collaborative platfrom. It is a Beta. It is open and based on the social science concept

We protect your data from hacking (better than how you would on your local machine)
We don't backup your data
We don't see your data (unless you're asking us to do so)
Read the Disclaimer at https://w3.iss.it/site/aries/

ARIES: Credits

The Galaxy ARIES core group at the EU RL for *E. coli*:

- Stefano Morabito (EU-RL VTEC): stefano.morabito@iss.it

ARIES Scientific coordination, tools design, contact person

- Arnold Knijn (SIDBAE): arnold.knijn@iss.it

ARIES Administrator, tools integration, contact person

- Valeria Michelacci (EU-RL VTEC): valeria.michelacci@iss.it

Tools design