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

The trendy “Omics” approach I

- Is changing the labs language
- Computers are becoming more and more visible in the labs
- Increasing importance of Data Storage and Sharing
- Structural re-think of the Labs management (Storage and back-up facilities, fast internet, compression, encryption and other data protection measures)
- Computing hardware become quickly obsolete

High throughput Sequence Data

High storage capacity

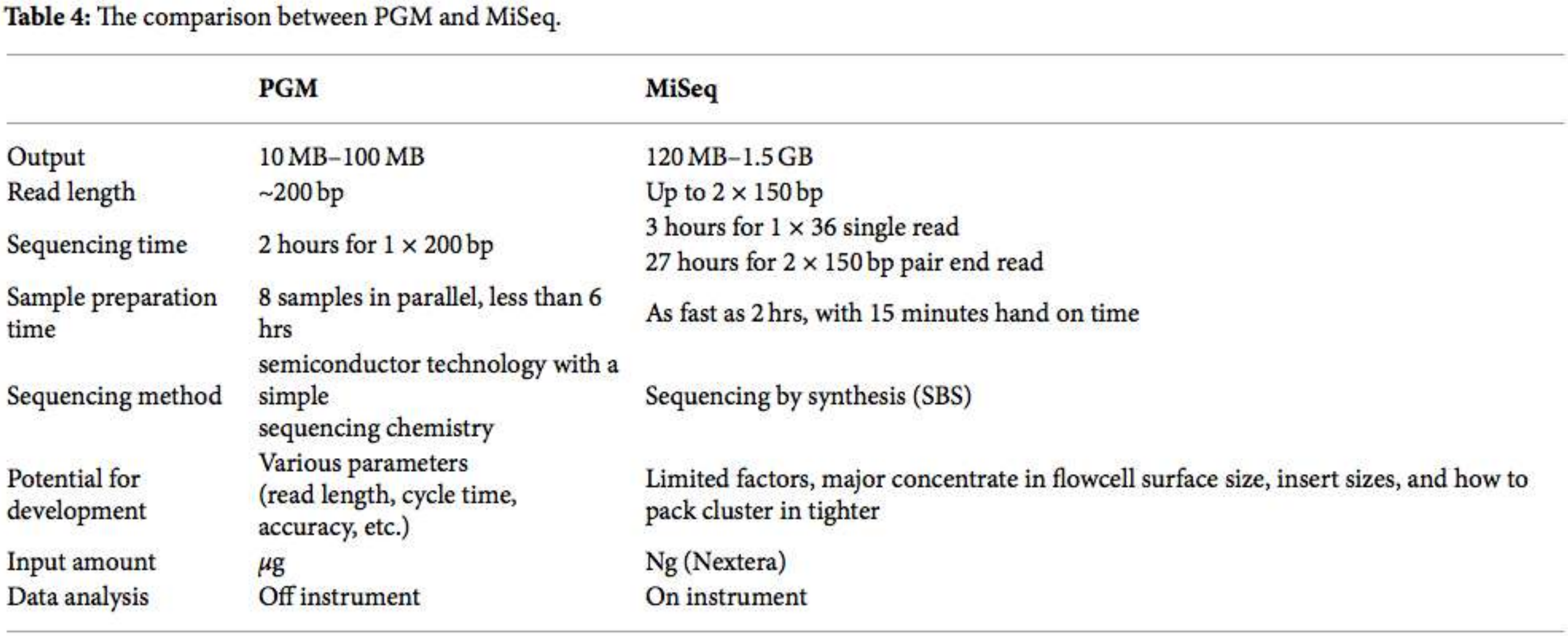
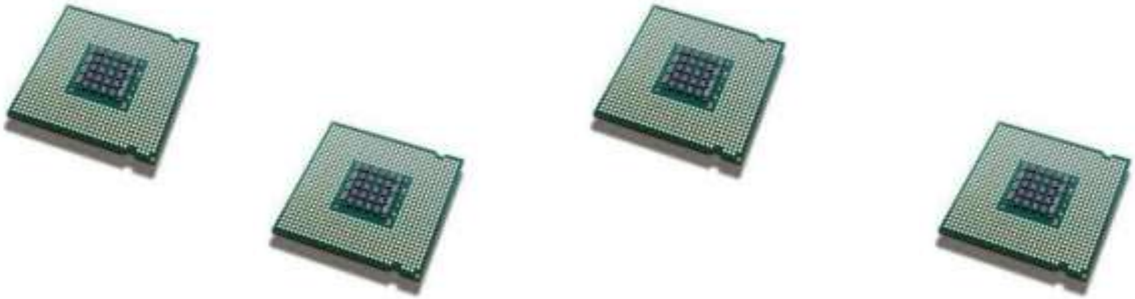


Table 4: The comparison between PGM and MiSeq.

	PGM	MiSeq
Output	10 MB–100 MB	120 MB–1.5 GB
Read length	~200 bp	Up to 2 × 150 bp
Sequencing time	2 hours for 1 × 200 bp	3 hours for 1 × 36 single read 27 hours for 2 × 150 bp pair end read
Sample preparation time	8 samples in parallel, less than 6 hrs	As fast as 2 hrs, with 15 minutes hand on time
Sequencing method	semiconductor technology with a simple sequencing chemistry	Sequencing by synthesis (SBS)
Potential for development	Various parameters (read length, cycle time, accuracy, etc.)	Limited factors, major concentrate in flowcell surface size, insert sizes, and how to pack cluster in tighter
Input amount	μg	Ng (Nextera)
Data analysis	Off instrument	On instrument

Computationally intensive applications



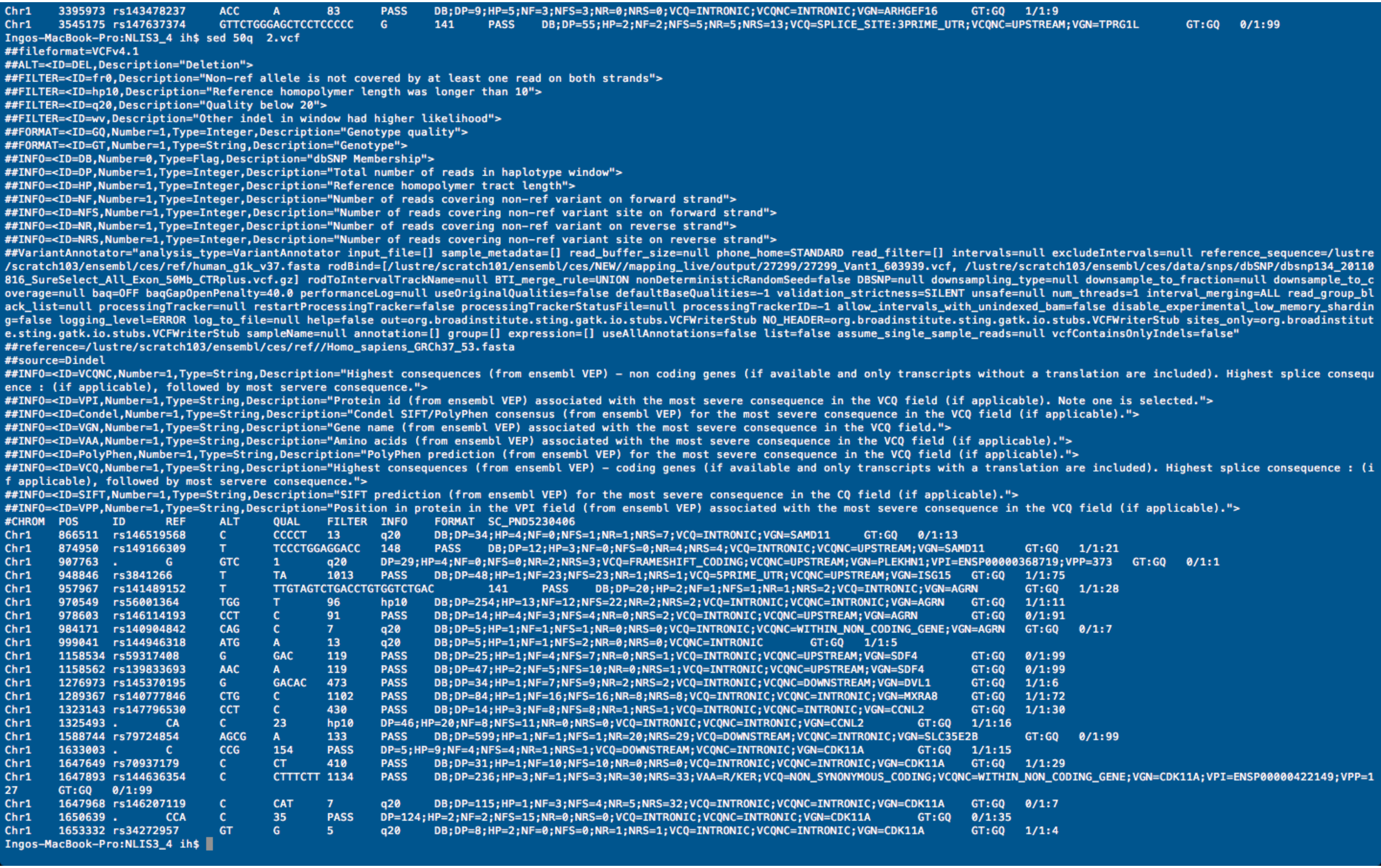
The assembly of a bacterial genome (approx 600 MB) takes approx 30mins to 5 hours and completely occupies the computation capacity of a processor (or a core of a quad-cores processor)

The assembly of a metagenomics sample (up to more than 3 GB) may take days and in some cases it will not be assembled at all (regardless the number of cores available)

The trendy “Omics” approach II

- Growth curve and colony morphology is going to be replaced by other concepts such as the Phred score or NG50
- Computing “omics” data requires expertise that is still scanty in the average NRL
- To get oriented in the blast of info proposing bioinformatics software might be a nightmare

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$ 
```

Data analysis: Software suites

BioNumerics

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

USER-(almost)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

BUILT IN THE ION TORRENT TECHNOLOGY PACKAGE
IT ADMIN BY LifeTech

Data analysis: Specialized web servers

DTU

Center for Genomic Epidemiology

- Species identification
- *de novo* assembly tools
- VirulenceFinder
- ResFinder
- MLST
- SNPs tree and newly deleveloped NGS-driven philogenetic tools
- Other useful molecular microbiology/epi tools

FREE, USER-FRIENDLY WEB INTERFACE, COMPLETELY CLOSED ENVIRONMENT, LIMTED POSSIBILITY TO INTERVENE FOR USERS

Data analysis: web servers for general “omics” analysis

Galaxy

- can be installed locally
- can run any commnd-line running scripts
- *de novo* assembly tools
- BLAST search of interesting genes
- Alignment of sequences, production of VCF files, production of dendrograms
- virtually unlimited possibilities……

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

What we had in mind…

Bring “omics” into the NRLs real life while keeping (almost) all those problems out

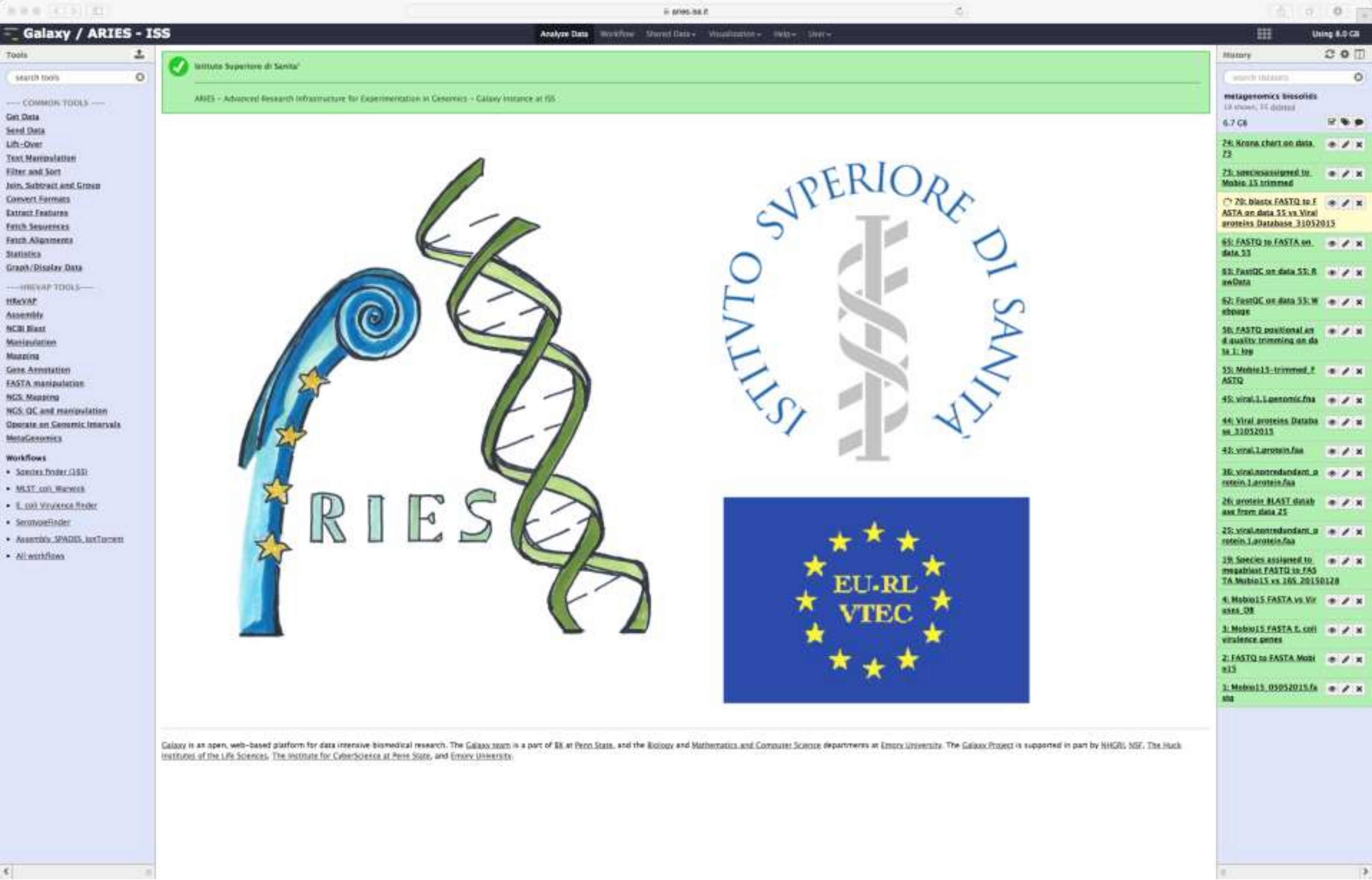
Build knowledge on these game-changing approaches in our network

Provide new analytical tools for *E. coli* detection and typing based on “omics”

Develop a flexible platform for Routine Work as well as Research

Keeping an eye on how the molecular surveillance will develop

ARIES: A shared workspace for intensive data analyses



Istituto Superiore di Sanita'

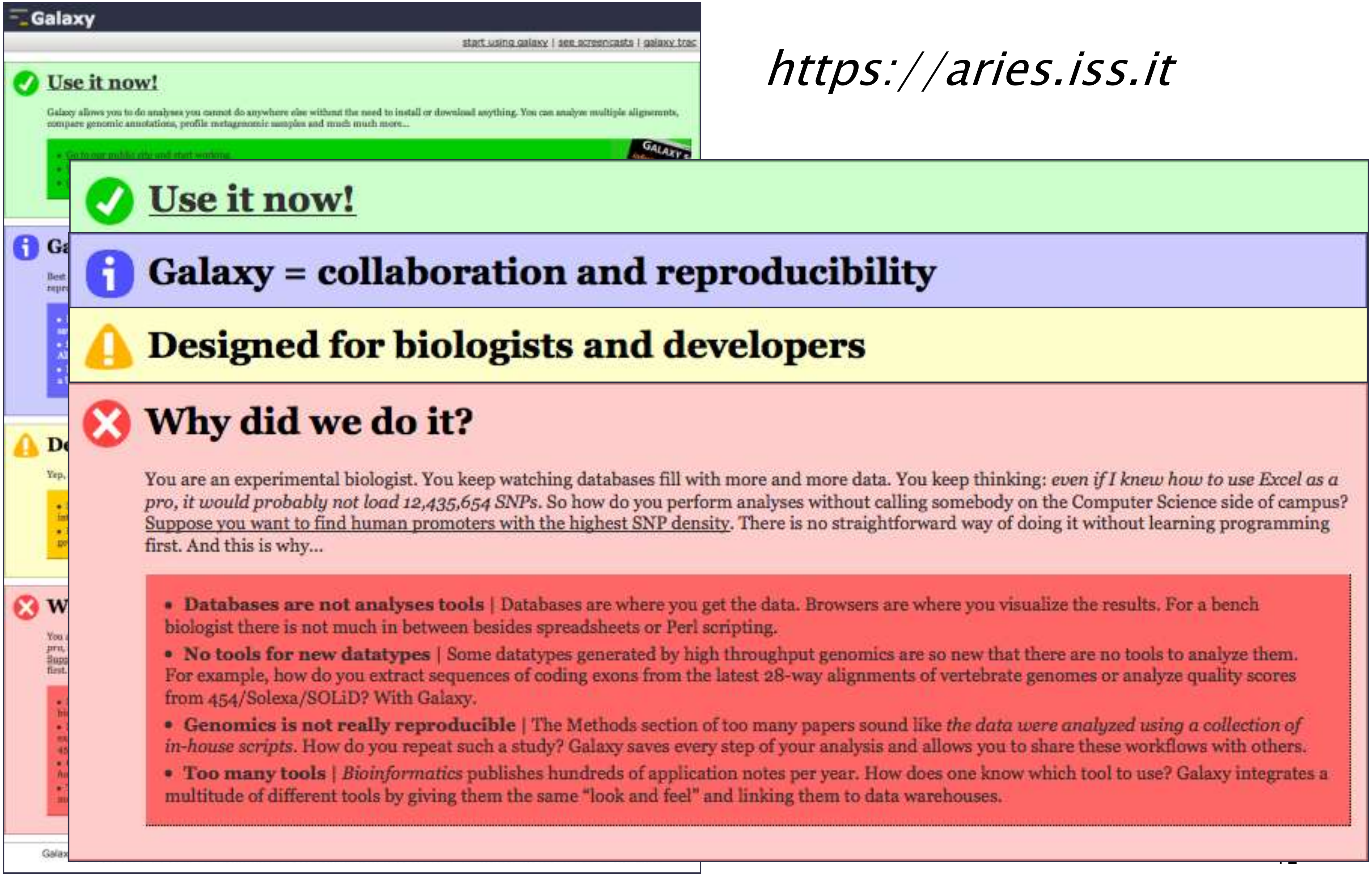
ARIES – Advanced Research Infrastructure for Experimentation in Genomics – Galaxy instance at ISS

- 74: Krona chart on data 73
- 73: speciesassigned to Mobio 15 trimmed
- 70: blastx FASTQ to FASTA on data 55 vs Viral proteins Database 31052015
- 65: FASTQ to FASTA on data 55
- 63: FastQC on data 55: RawData
- 62: FastQC on data 55: Webpage
- 56: FASTQ positional and quality trimming on data 1: log
- 55: Mobio15-trimmed FASTQ
- 45: viral.1.1.genomic.fna
- 44: Viral proteins Database 31052015
- 43: viral.1.protein.faa
- 36: viral.nonredundant_protein.1.protein.faa
- 26: protein BLAST database from data 25
- 25: viral.nonredundant_protein.1.protein.faa
- 19: Species assigned to megablast FASTQ to FASTA Mobio15 vs 16S 20150128
- 4: Mobio15 FASTA vs Viruses_DB
- 3: Mobio15 FASTA E. coli virulence genes
- 2: FASTQ to FASTA Mobio15
- 1: Mobio15_05052015.fastq

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. 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 Emory University.

ARIES: Get started

https://aries.iss.it



Galaxy

start using galaxy | see screencasts | galaxy trac

Use it now!

Galaxy allows you to do analyses you cannot do anywhere else without the need to install or download anything. You can analyze multiple alignments, compare genomic annotations, profile metagenomic samples and much much more...

- Go to our public site and start working

✓ Use it now!

Galaxy = collaboration and reproducibility

Designed for biologists and developers

Why did we do it?

You are an experimental biologist. You keep watching databases fill with more and more data. You keep thinking: *even if I knew how to use Excel as a pro, it would probably not load 12,435,654 SNPs*. So how do you perform analyses without calling somebody on the Computer Science side of campus? Suppose you want to find human promoters with the highest SNP density. There is no straightforward way of doing it without learning programming first. And this is why...

- **Databases are not analyses tools** | Databases are where you get the data. Browsers are where you visualize the results. For a bench biologist there is not much in between besides spreadsheets or Perl scripting.
- **No tools for new datatypes** | Some datatypes generated by high throughput genomics are so new that there are no tools to analyze them. For example, how do you extract sequences of coding exons from the latest 28-way alignments of vertebrate genomes or analyze quality scores from 454/Solexa/SOLiD? With Galaxy.
- **Genomics is not really reproducible** | The Methods section of too many papers sound like *the data were analyzed using a collection of in-house scripts*. How do you repeat such a study? Galaxy saves every step of your analysis and allows you to share these workflows with others.
- **Too many tools** | *Bioinformatics* publishes hundreds of application notes per year. How does one know which tool to use? Galaxy integrates a multitude of different tools by giving them the same "look and feel" and linking them to data warehouses.

ARIES: Roadmap towards a Common analytical bioinformatics interface

July 1, 2015: ARIES Opens to ISS Users

- Accounts distributed
- New sections may be created
- Open to collaborations

July 1, 2015:

- More Molecular epidemiology tools
- Metagenomics

October 31, 2015: ARIES publicly exposed (Beta)

- Accounts available at National Level
- Accounts available for the *E. coli* network
- Stress test for the architecture

October 31, 2015:

- More tools for NGS data Mining
- More tools for NGS data Visualization

January 1, 2016: ARIES on the World Wide Web (Beta)

Accounts granted to international users

Requests addressed to the IT administrator

January 2016 on…

- Other "Omics"…..

ARIES: Credits

The Galaxy ARIES core group:

- 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, Galaxy tools integration, contact person

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

Tools design

- Massimiliano Orsini (IZS AM): m.orsini@izs.it

Tools design, code-writing