Introduction to NGS data formats, basic tools and servers for analysis

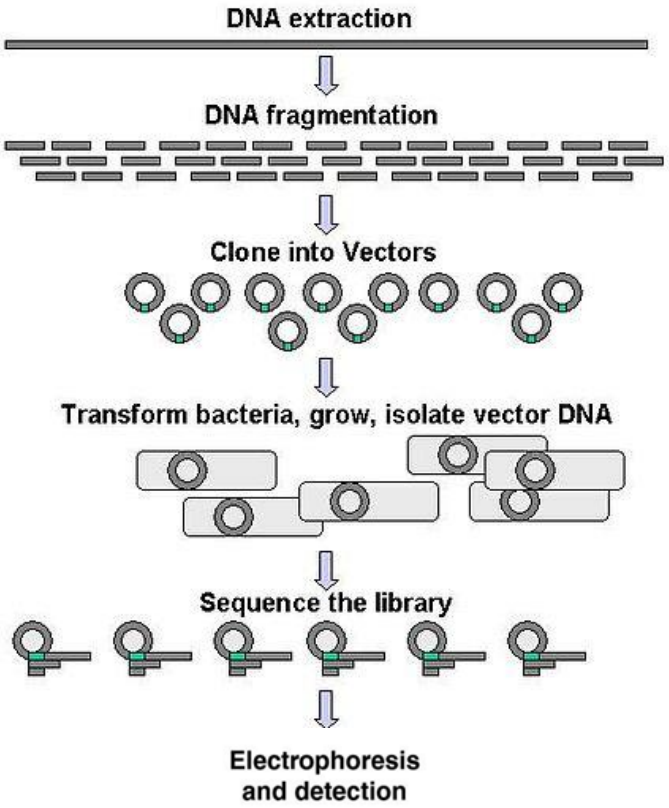
Valeria Michelacci

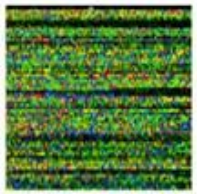
Bioinformatics training,
June 2018

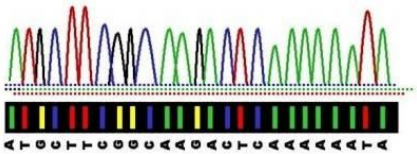
Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health
European Union and National Reference Laboratory for *E. coli*, Rome, Italy

Conventional sequencing vs NGS

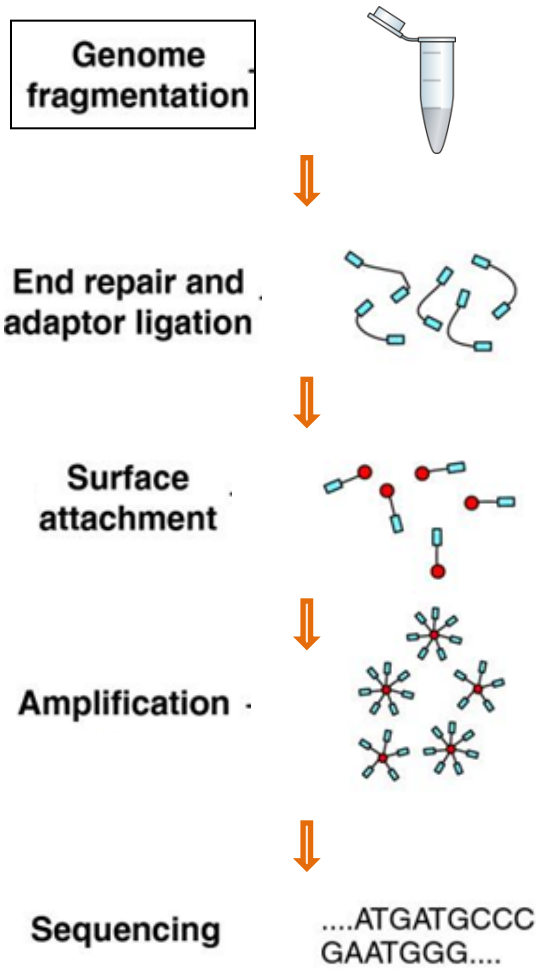
Conventional



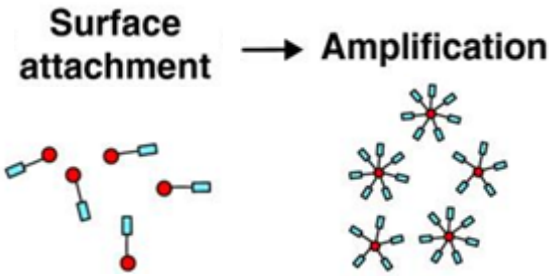




NGS Pipeline



Next generation sequencing



...

...

illumina MiSeq

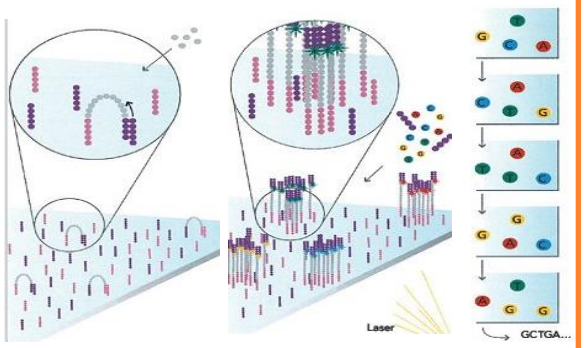
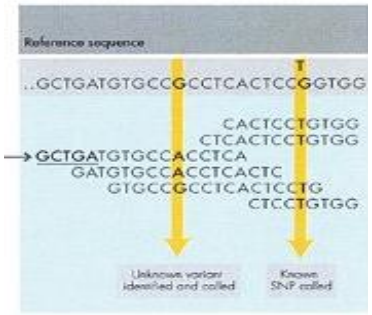
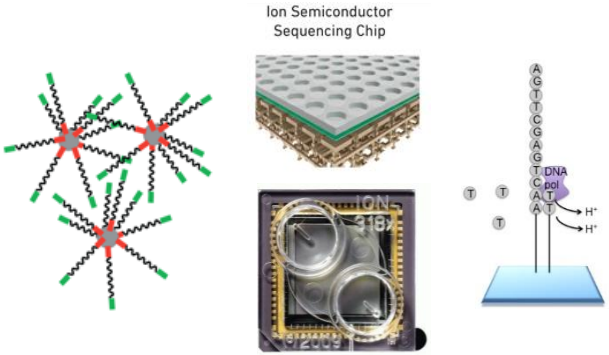


Image capture
Fluoresence detectioin

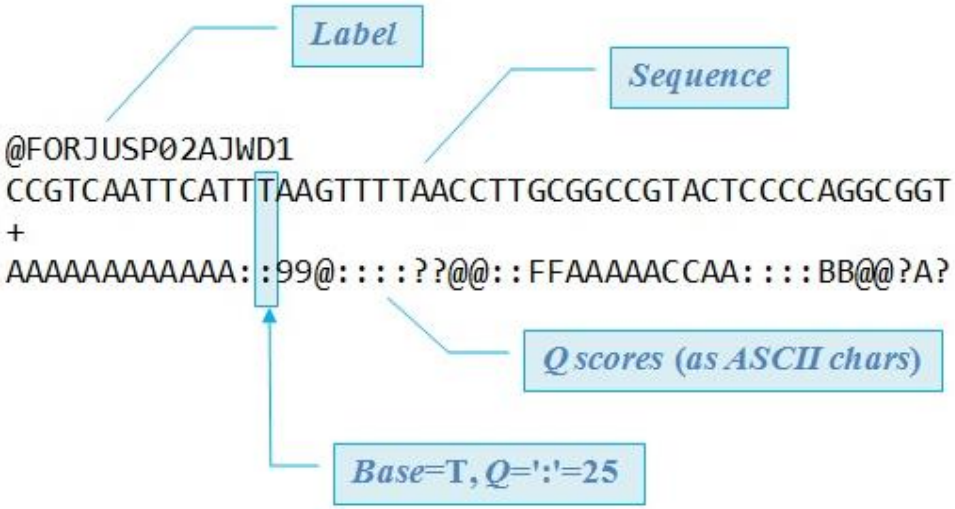


200bp-400bp short reads



pH variation when incorporating nucleotides in the growing strand

.fastq files



Each .fastq file covering a 5 Mb genome at 50X weights about **500 MB**

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Phred quality score

$$Q = -10 \log_{10} P$$

from 0 to 93 using ASCII characters 33 to 126

.fastq files

@

```
@X1L6C:01561:00672
AAATATCACCAAATAAAAAACGCCTTAGTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTG
GATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCAC
TAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCA
CCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAAGCCCGCA
CCTGACAGTGCGGGCTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCG
+
CC:9::FBC<CD7:88888(:>><C<CCCC<CCBBAAB/A@A8888,;<@;AABBB=?;B98992:B<
CGBBCGDCC??BCC;BB<ADEEED*CCCAAACCCBCABBDDBB>B??A;999;@8=>199A7>9::CBCH:B:>>>)999)
77037;<7==5=@@BBCC:C@BBB9B<E<D9>?><<6ADCBCBAABB@@@DDCCBA@@==+=.//?B<??AEB::6;DCD>
C:;;;-:9:BC<BBCCC9??<AA;AG<CB>GD@B;;;A<AE;AA<B??@9@C<BB<???BB;BBBAAAA:::BAB099/9>
@========(<<?)99997>>CCEBA>>=>2373333&3:99-33(3--717---43606704/47761
@X1L6C:01104:03031
AGAAGCTGCTATCAGACACTCTTTTTTTAATCCACACAGAGACATATTGCCCGTTGCAGTCAGAATGAAAAGCTGAAAAATA
CTTACTAAGGCGTTTTTTATTTGGTGATATTTTTTTCAATATCATGCAGCAAACGGTGCAACATTGCCGTGTCTCGTTGCTC
TAAAAGCCCCAGGCG
+
@AC=BCCC???B?@@CBB@???>>>>>*?8??>DAABEBCBABCAAA:@@>+9:8>;<;//.
98283988*44449;;9/88:?29:>>5;78333333&399298:6/./DCDDCC';>:ACBDAABB??9::+9<
1444@:?77-3<03368:8755888;:9833)3777'--'--
@X1L6C:03659:02717
GCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATA
GCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTA
CCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGACCCGCCACTGACCAGTGCG
+
???9?BB@<CAA;A8@?:?@@5::BCCCEC;C=CCC8CEJ8DE;AACF>CC?DDCCCBB:B@???9?;B=B=CAA@?;?BCG
CCCCCCBABBBBCCDDAA2:4;@???CAB@AAA9@@AB?C:;;C;CDCCC>ECCAA<AC<CB>DC<AB=CD=C9::A4::>
CC;@@@A?CI@DDAFKDDD:A@CBCDC::::99199+8;4746@CA?)<444/3:4934333-3888//
@X1L6C:02011:02071
TTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACA
CAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGTGACGCGTACAG
GAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGTAACG
+
=0>>>19;;,;;7=CCDADC;?::::::,5;;==4>273:<@BBCF=CDH;@;MMFEED@?>>>:::::*5/55<
;::@:;:BC=BCBB<B@@@D<@@B:;3:::9@<BB=BD=AC;@B:??3::CAC=CD;;;=BBAB>CC;AA;BAAAA9AD@>>
>>?955>4?9499985555558&4<>2:;661499888...88/56666666$;6/.5:8(..+'++
@X1L6C:01333:03005
GCAATGCCAGGCAGGGCATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGA
TGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACG
+
555/55/(///(////(/8:9:<=>?<?@:98A??676<:;;@:5555555554444;=4443333;383338<68>>
68=333118311111111113933644588?==<76992---2+++0/
```

@

@

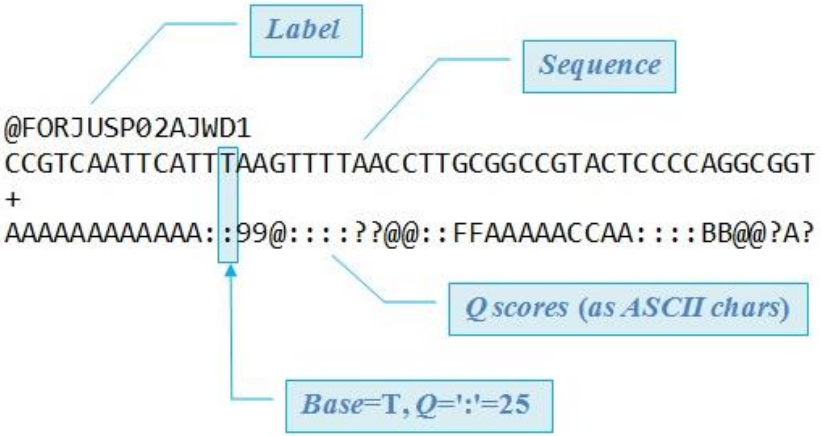
@

@

…and so on

Quality check

Output of NGS sequencers



.fastq file

Input for quality check

Sequencing errors would impact every following application

Unreliability of following results (and difficulty to detect the existence of problems!)

Parameters to control

- Phred score
- GC content distribution over all sequences
- Distribution of undetermined bases (N)
- Distribution of nucleotides
- Length of the reads
- Coverage

Adoption of corrective actions is possible to minimize some of these problems

Coverage (depth)

Reads mapped on a reference genome

Ref seq

COVERAGE

Mapped reads

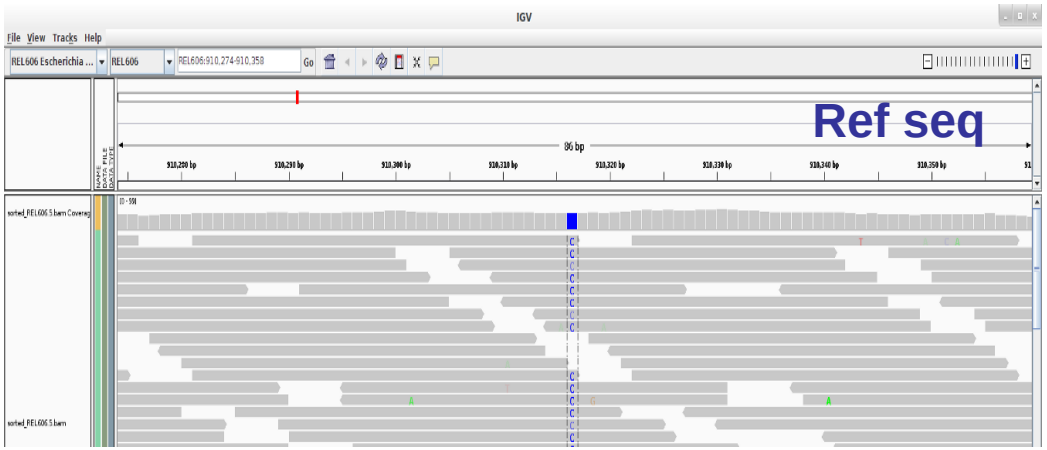
Coverage assessment:

Total length sequenced in Mb / expected genome size (5 Mb)

Count of reads mapping on housekeeping genes (e.g. MLST)

Alignment (mapping)

Alignment of the sequencing reads on a reference sequence or on a database of reference sequences



Possibility to directly inspect the presence/absence of a target sequence and the presence of SNPs at interesting positions by opening the bam file with a graphic viewer (e.g. IGV)

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR
ME2UT:01383:01267	0	gad:3:EF547388	1285	0	113M18I4M
ME2UT:02555:01592	16	gad:4:CP001925	1123	0	27M1I248M39I4M
ME2UT:02231:01820	0	gad:5:CP001846	87	1	138M
ME2UT:01605:00255	16	gad:5:CP001846	399	1	51M
ME2UT:01345:02031	16	gad:5:CP001846	685	1	176M
ME2UT:03330:02136	16	gad:5:CP001846	1050	1	6M1I38M
ME2UT:01475:02165	0	gad:6:BA000007	1	0	3M31I47M1D130M
ME2UT:01488:00709	16	gad:6:BA000007	1	0	4M32I55M1I149M
ME2UT:01943:01152	16	gad:6:BA000007	13	1	196M1I50M1I10M

Possibility to convert the output in a sam file (tabular) to extract interesting info and sequences

Assembly

Short sequencing reads

.fastq file

```
@HWI-ST700693:238:B0224ACXX:1:1101:1218:1982
NACACTTGCTTTGGTGACAGCGGGGCATCCTCAAGC
+
#1=DDDDDHAFF?GEFGIIIIIIIIIIIIIIIIIFI
@HWI-ST700693:238:B0224ACXX:1:1101:1161:1986
NGATTTTGACCTCTCCAGTTTCCTCTTAACACTTTG
+
#1:BDFFFGHHHGJJJIIJHIJJJJJJJJJJJJJJJ
@HWI-ST700693:238:B0224ACXX:1:1101:1193:1989
NTATCCAGCCTGCGGTGCTACTTGGTGGAAGAGGAT
+
#1=DDFFFHGHGGJJFGHJJIJJJIEGECHDFHCC?
@HWI-ST700693:238:B0224ACXX:1:1101:1440:1981
NTCAAGAATCCAAGTGGGGCCAGCATAATGTACGCT
+
#1=DDFFFHGHDFDAEGIIFGIICGGHGBFGEFDHI
@HWI-ST700693:238:B0224ACXX:1:1101:1367:1983
NATTAGAACAGATCGCTACTTCGCCCGAAGATACAT
+
#4BDFFFFFHHHHHJGIJIJJJIJJJJJJJJJJJJIJ
@HWI-ST700693:238:B0224ACXX:1:1101:1395:1988
NTGGAAACGTTTTTAAACGCGGAGACAGCGTGGAGT
+
#1=DDFFFHCFFHJJJIJJIJJIJJJJGGIFHIGI7
@HWI-ST700693:238:B0224ACXX:1:1101:1285:1994
NCTTTGCTGTATTGACCGTTTGTAGATTTGAATCTT
+
#4=DDFFFHBHHHHIGIJFHIJFGGGIGIHIJIJII
@HWI-ST700693:238:B0224ACXX:1:1101:1632:1989
NTCTATGAATGTTCAAGCGGTAGCTGAGGAGAGTCC
+
```

FastqSize ≈ GenomeSize x Coverage x 2

At least 0.5 GB per genome

Partially assembled genome (contigs)

.fasta file

```
>NODE_1_length_449_cov_4.835189
ATCTTTCGCGCCTTCCAGCTCCAGCCATTCGGAACCGTTCGCCAGAAAACGGGCGTAATC
GGGTAAGACATAGCGCGGTTTGTACGGCGCATGACCTTCAAACATATCGCAGATTACACC
TTCATCCAGCGCGCGGCGGGCTTCGGCAGGAAGCTGTGGGTAAGGCAGATTGTTTTCTGC
TTCCAGTGCCAGAAAATGGCGCTTCTGCTCCGGGCTAAGCACTGGGCTGGTGACAATTTG
CTGGCAACGTTGTTGCAGTGCATTTTCATGAGAAGTGGGCATCTTCTTTTCCTTTTATGC
CGAAGGTGATGCGCCATTGTAAGAAGTTTCGTGATGTTCACTTTGATCCTGATGCGTTTG
CCACCACTGACGCATTCATTTGAAAGTGAATTATTTGAACCAGATCGCATTACAGTGATG
CAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGG
>NODE_2_length_309_cov_4.686084
ACTGGTCAGTGCGGGTATCCTTGGACAATGGCCGATTGGACGTCTGGCGGATAAGTTTGG
TCGACTGCTGGTGTTGCGTGTTCAGGTCTTTGTCGTCATTCTCGGCAGTATCGCGATGCT
TAGCCAGGCGGCGATGGCCCCAGCGTTATTCATCCTCGGTGCCGCTGGCTTTACGCTATA
TCCGGTGGCGATGGCATGGGCTTGCGAGAAAGTTGAACATCATCAACTGGTGGCGATGAA
CCAGGCCTTACTGTTGAGCTATACTGTGGGAAGTCTGCTTGGCCCGTCATTTACCGCTAT
GCTAATGCAGAATTTCTCCGATAATTTATTGTT
>NODE_3_length_101_cov_3.346535
AGCGCATGAGCGCGCAGCGCCGCGTTACGTGGTGCATCAGCATGATGTTGGCCGGAGAG
TACAGAGACTCCCCTTCATCCATGATGCCCTCTTTCACCAGCAGTTCTTCAATCATCACC
AGACC
>NODE_4_length_311_cov_3.610933
CATCAACGCTAAAAGCCAGATGACGCAGACCGCAAGCTTCCGGTCGGCTGGGTCGTTCCG
GCGGGAACGGAAATGAGAAAAGCTCAATCACATATTGCCCATTAAGCGCCAAATCCCCTT
TCCATGAGTCGCGCGCTTCGCGATAGACTTCGCTTTGCAGCGTGAAACCAAGAATATCGC
AGTAGAAAGCTTTGCTCACCGCATAATCCGTCGCAATAATCGCAATATGGTGAACCTGTT
TTAAACCCAGCATAACGTCTCCTTTATTTGTTAACAGCACGTTACTCGCCCGGAAGCCGC
TCTGGCAAGTTATCCCGCCATTTTTAGGACTCGTA
>NODE_5_length_186_cov_4.973118
CGAAGATATAAGAAAGCGAACCAGAAAGAATGCCGGAGAACTTCATCAATTCATCACCTG
CATTGAGCAGATTTTGCAGGTTCTCAATAACCGGTAATCCAGCCCCAACGTTGGTGTCAT
AGAGGAATTTACGCCGCGATTTTTCCGCCGCATAACGCAACTGATGGTAGTAATCCATCG
ACGAGGTGTTGGCCTTTTTGTTCGGCGTGA
```

FastaSize for *E. coli* contigs

~5.5 MB

Data Analysis: Software stand-alone

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

geneious

- *de novo* assembly
- Mapping and variant calling
- Alignment and tree building
- Tools for molecular cloning and chromatograms analysis

Private company

Data Analysis: Software stand-alone

CLC Genomics Workbench

QIAGEN

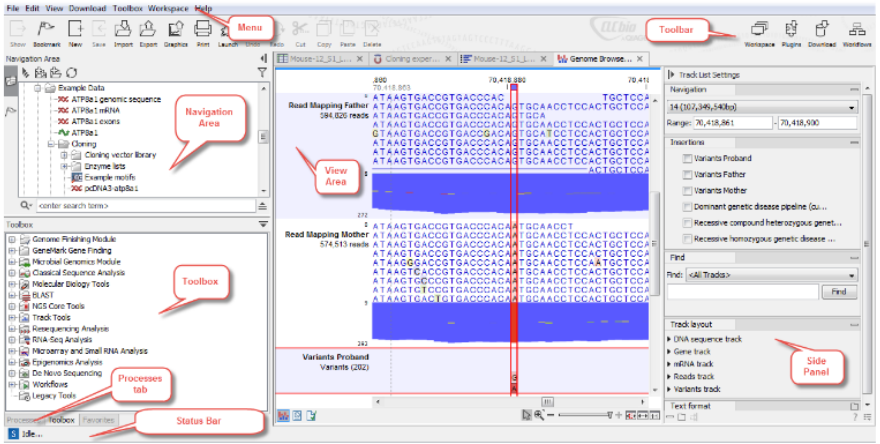
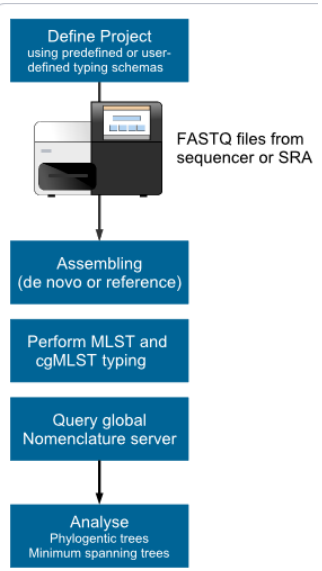


Figure 2.1: *The user interface.*

Private company

ridom BIOINFORMATICS



Pipeline for automated sequence analysis

Bacterial Genome characterization

Genome- wide allele and SNP calling

Private company

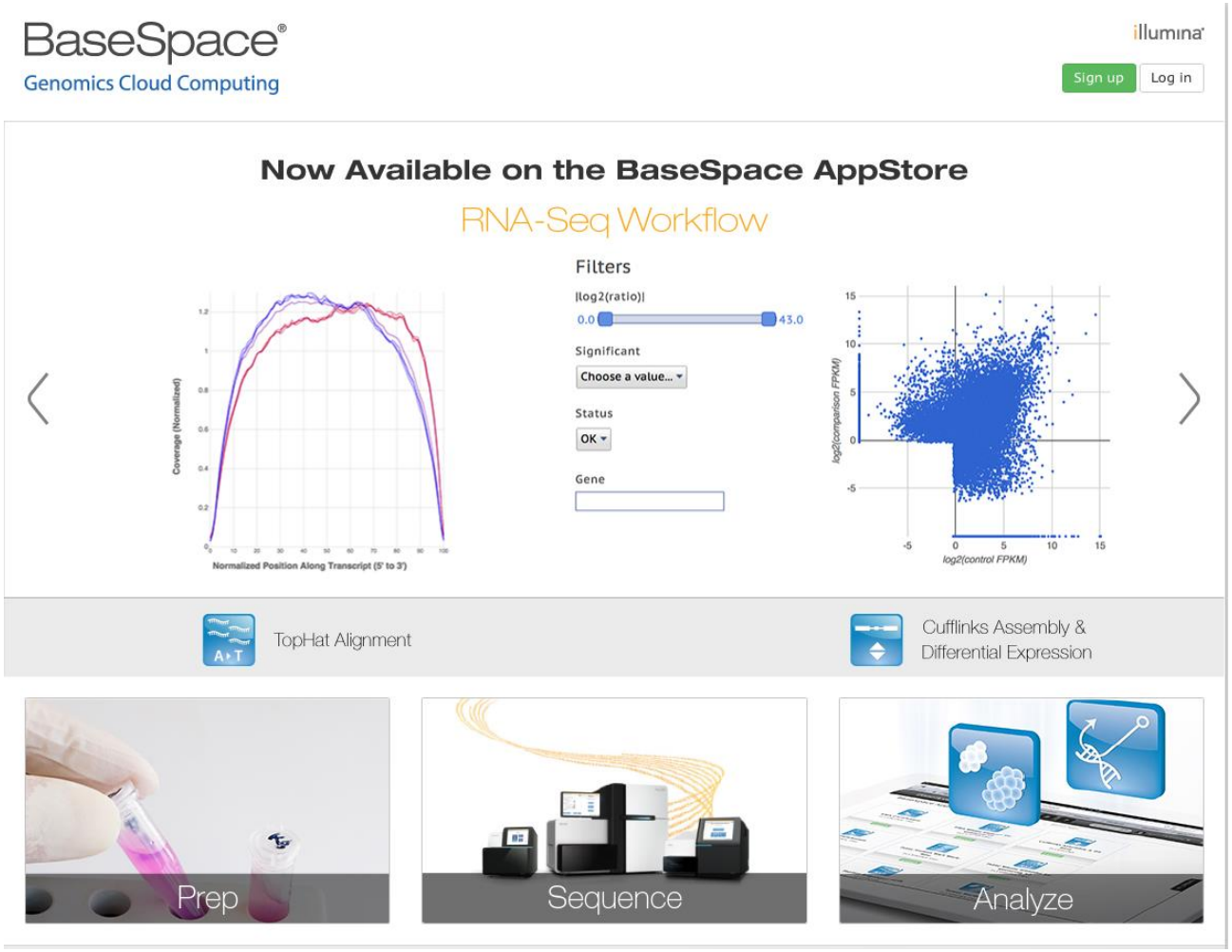
Data Analysis: Software stand-alone

Private company

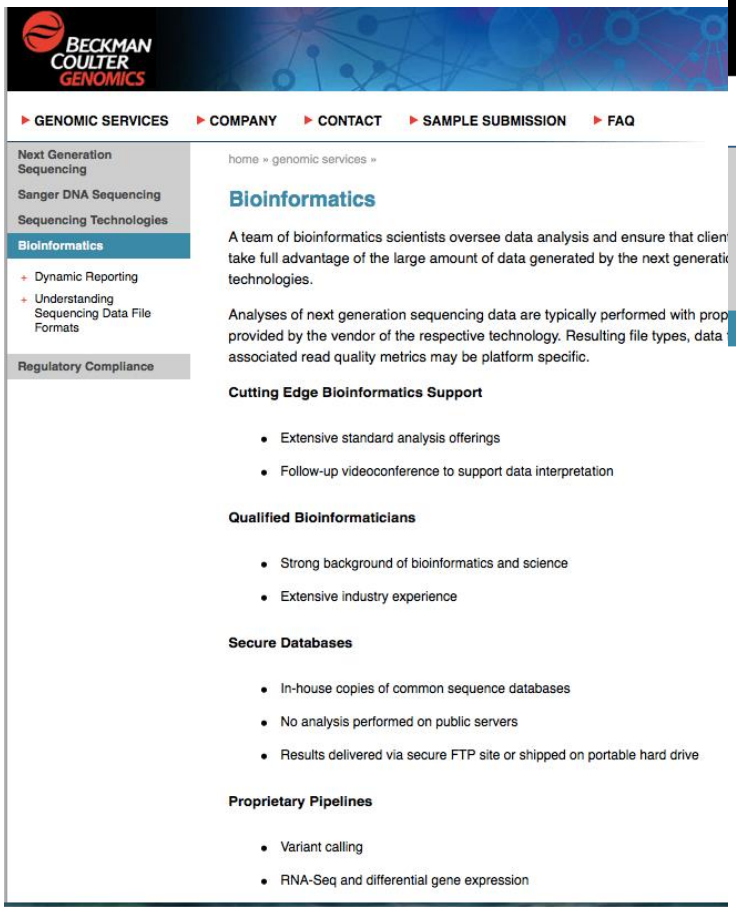
- *de novo* assembly
- Search for interesting genes
- Alignment of sequences, production of VCF files

BUILT IN THE ION TORRENT TECHNOLOGY PACKAGE

Data Analysis: Cloud-based Software



Data Analysis: Outsourcing



BECKMAN COULTER GENOMICS

GENOMIC SERVICES | COMPANY | CONTACT | SAMPLE SUBMISSION | FAQ

- Next Generation Sequencing
- Sanger DNA Sequencing
- Sequencing Technologies
- Bioinformatics
  - Dynamic Reporting
  - Understanding Sequencing Data File Formats
- Regulatory Compliance

home » genomic services »

Bioinformatics

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

Analyses of next generation sequencing data are typically performed with prop 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

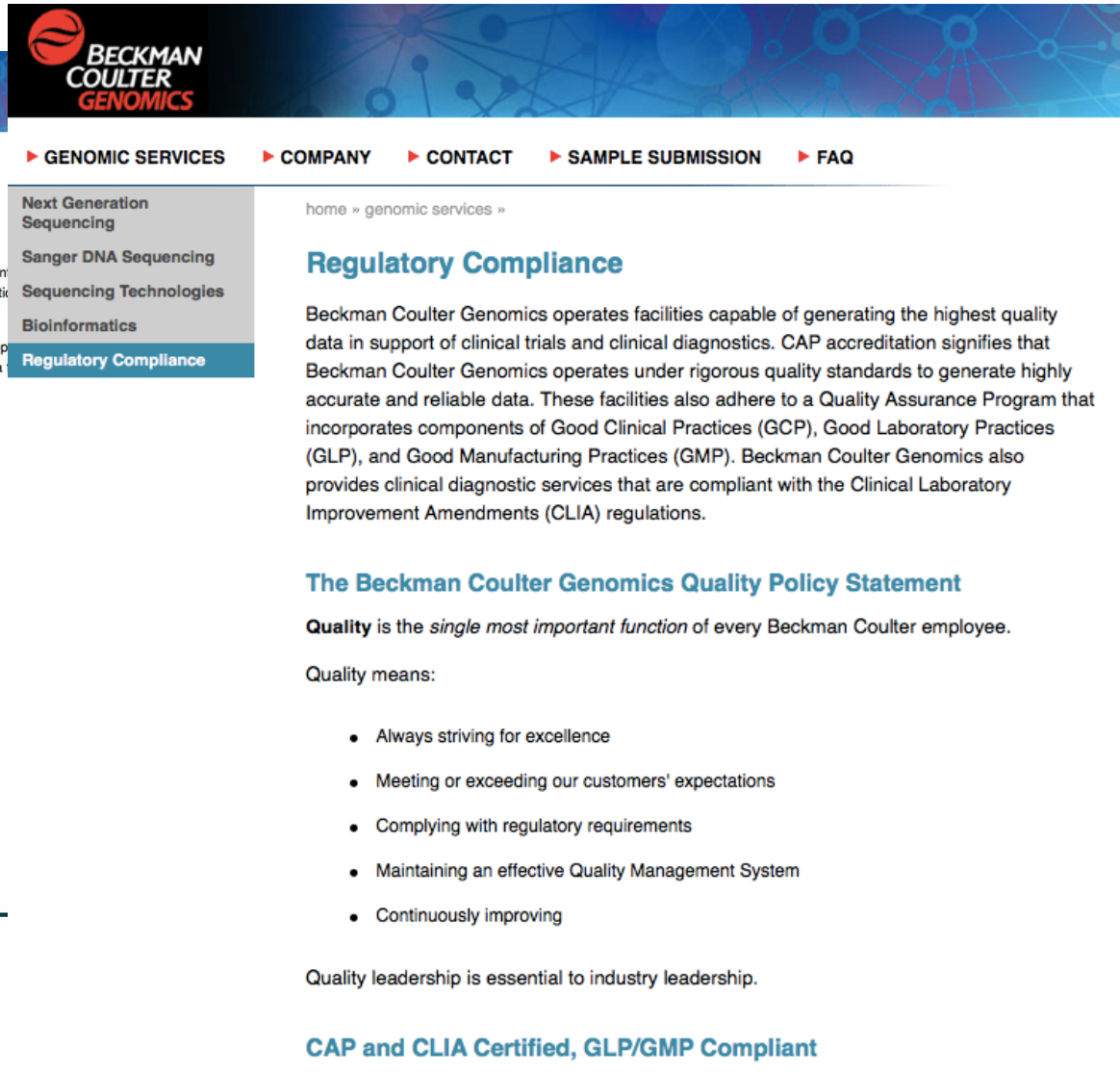
- Strong background of bioinformatics and science
- Extensive industry experience

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

- Next Generation Sequencing
- Sanger DNA Sequencing
- Sequencing Technologies
- Bioinformatics
- Regulatory Compliance

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

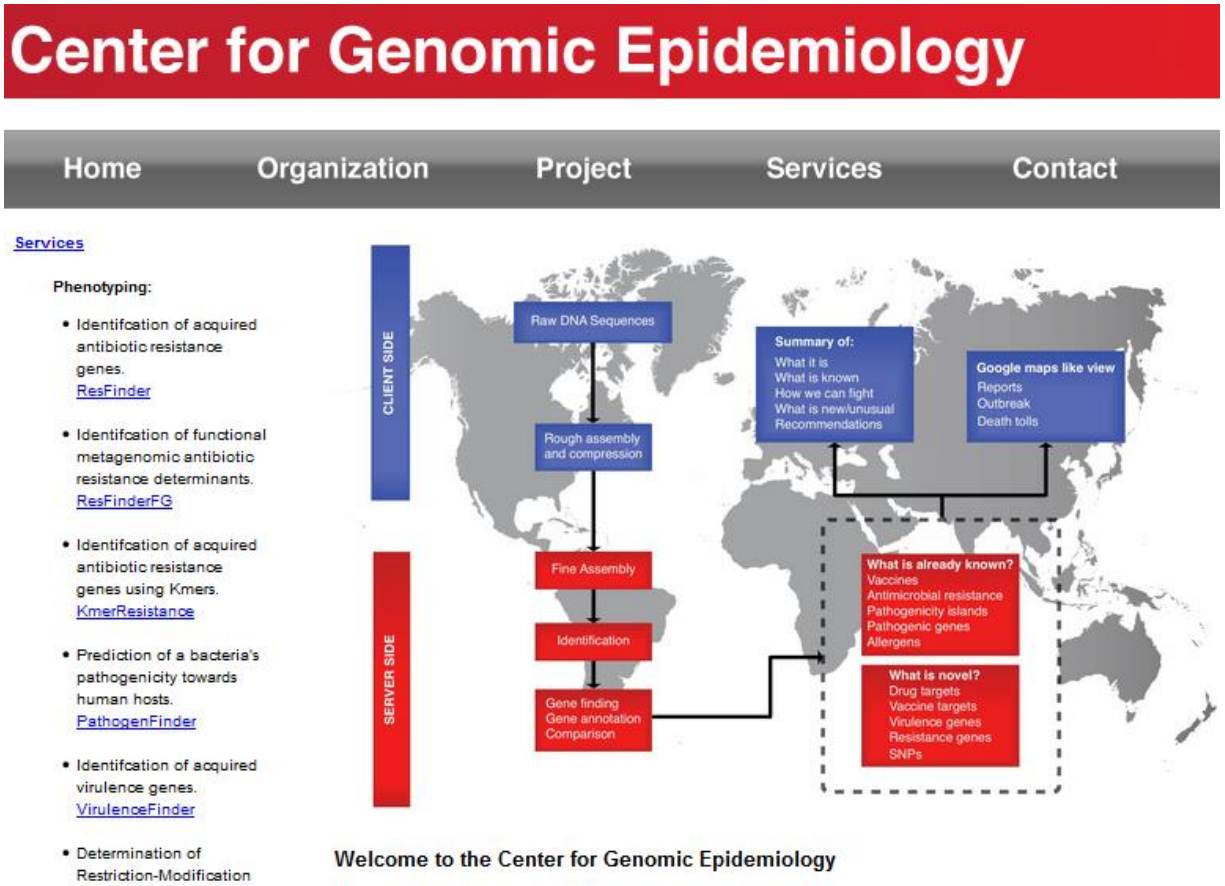
Quality leadership is essential to industry leadership.

CAP and CLIA Certified, GLP/GMP Compliant

Data Analysis: Public servers

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

FREE, USER-FRIENDLY WEB INTERFACE



Closed Public server

Data Analysis: Public servers

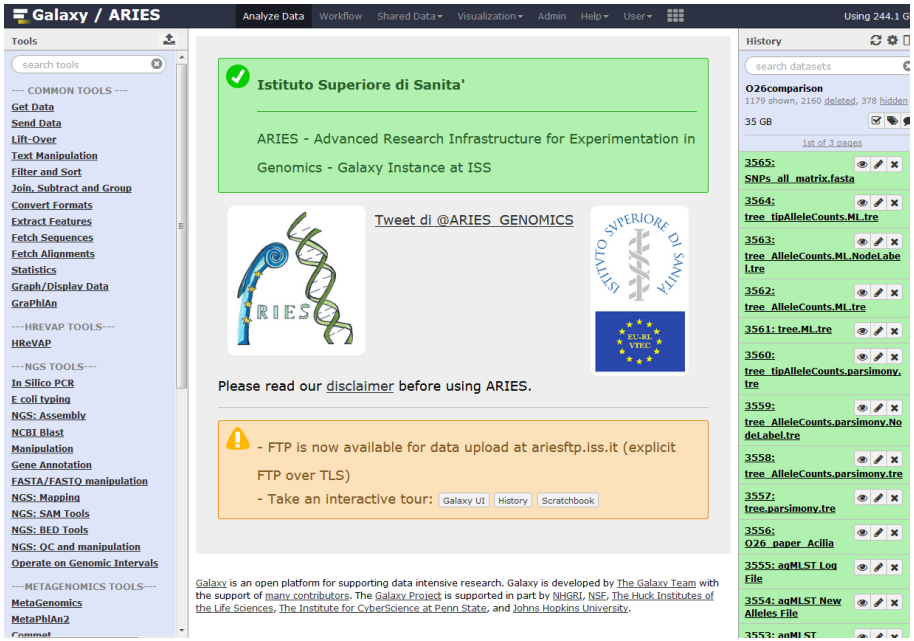
Galaxy

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

090+

Public Galaxy Servers
and *still* counting

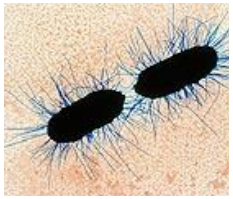
ARIES - ISS



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

Open Public server

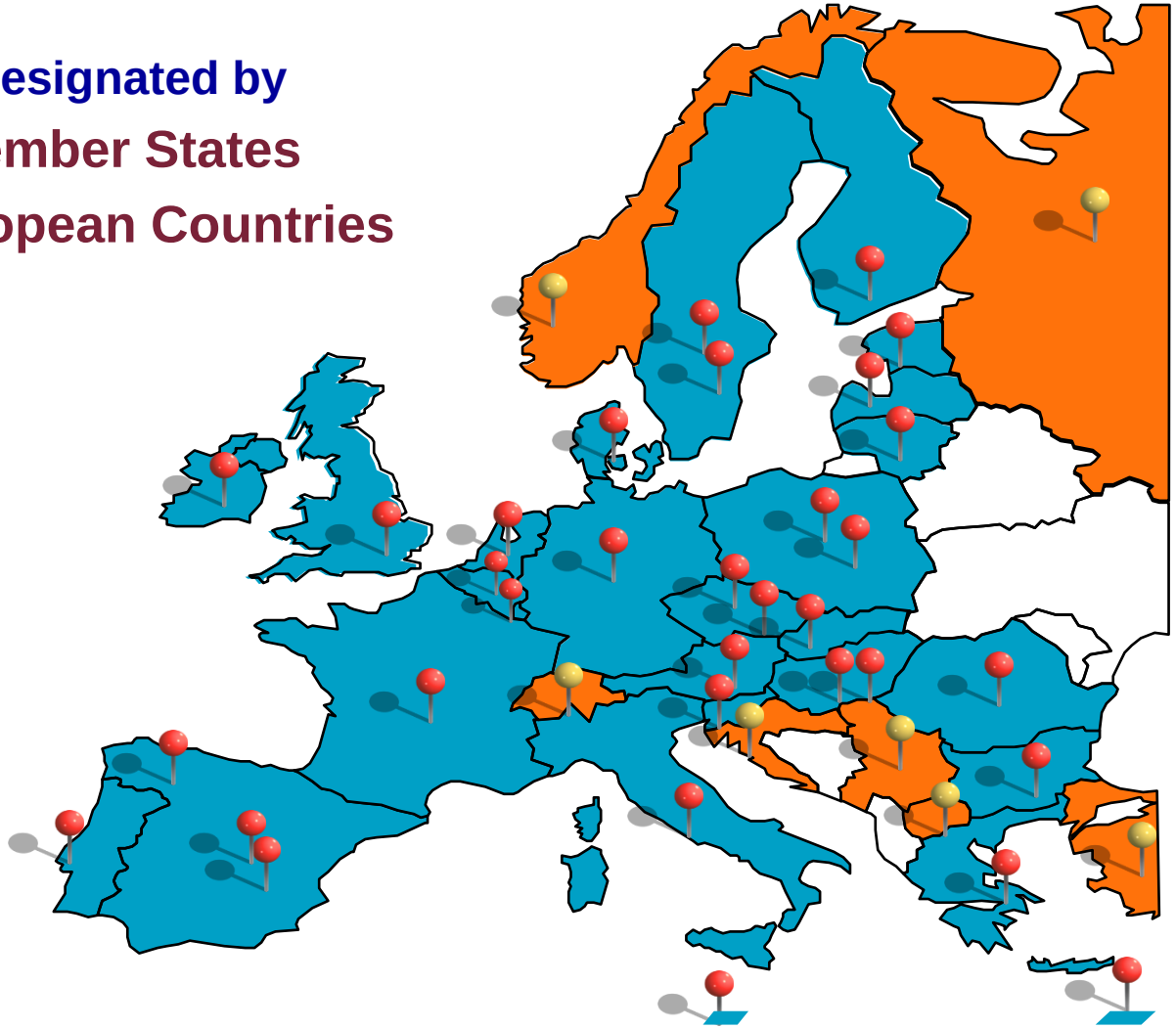
E. coli network and NGS: State of the Art



41 NRLs designated by

27 EU Member States

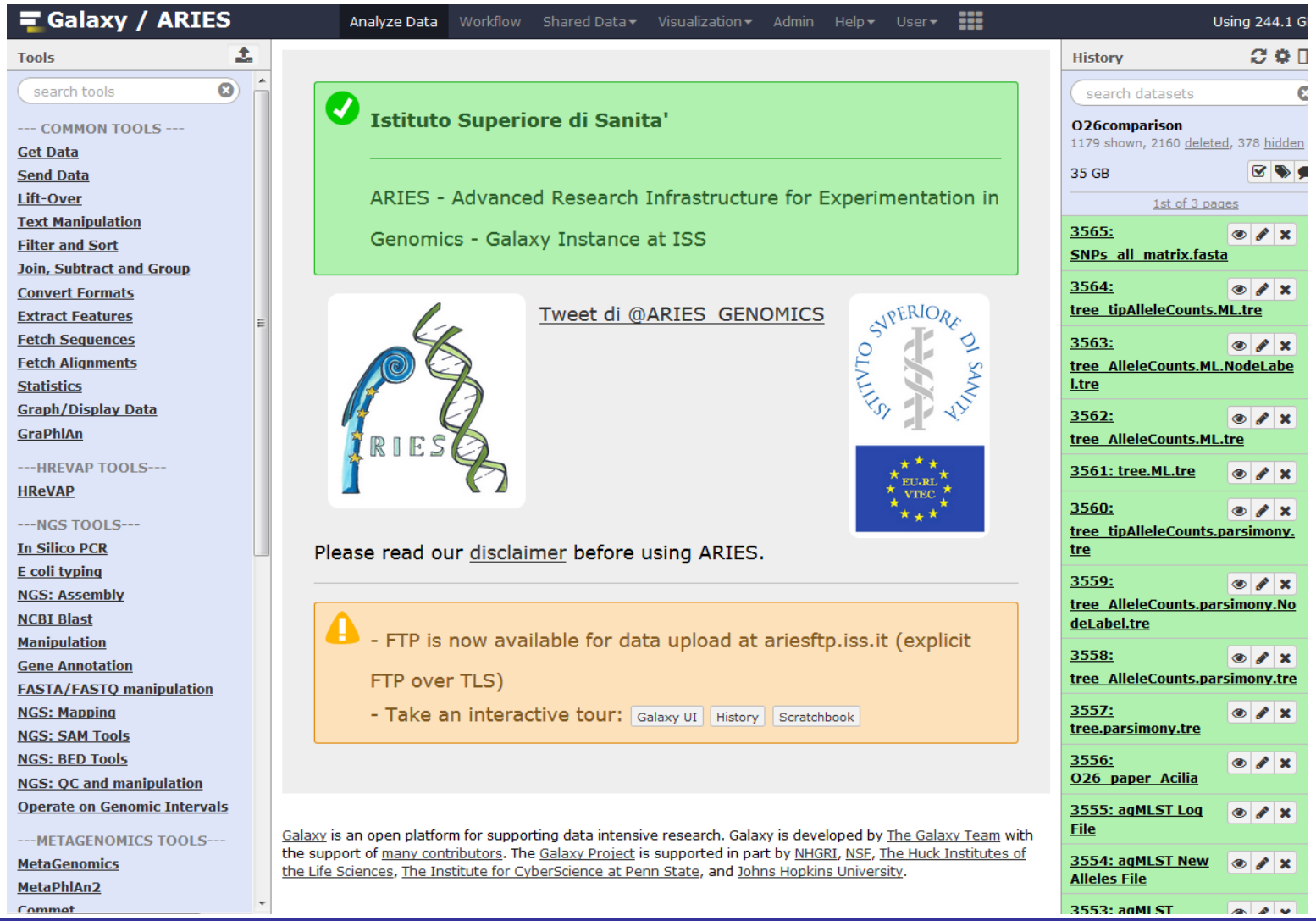
+ 7 other European Countries



In 2014 all the NRLs expressed the need for education in genomics data analysis by replying to a questionnaire

In 2015 we launched ARIES

ARIES: A Galaxy-based workspace for intensive data analyses



NGS data analysis vision: user - friendly UI

a5_pipeline.pl

```
Getopt::Long::Configure(qw{no_auto_abbrev no_ignore_case_always pass_through});
my $start = 1;
my $end = 5;
my $preproc = 0;
my $debug = 0;
my $metagenome = 0;
my $threads = 4;
my $adapter = dirname(abs_path($0))."/../adapter.fasta";
GetOptions( 'begin=i' => \$start,
            'end=i' => \$end,
                      'preprocessed' => \$preproc,
                      'debug' => \$debug,
                      'metagenome' => \$metagenome,
                      'threads=i' => \$threads,
                      'adapter=s' => \$adapter);

die $usage if (@ARGV < 2);

$AVAILMEM = get_availmem();
```

A5 pipeline A5 is a pipeline for assembling DNA sequence data generated on the Illumina sequencing platform. (Galaxy Version 20150522) Options

First read file in fastq format

379: Bowtie2 on data 135 and data 109: aligned reads (sorted BAM) (as FASTQ)

Second read file in fastq format

379: Bowtie2 on data 135 and data 109: aligned reads (sorted BAM) (as FASTQ)

Execute

A5 is a pipeline for assembling DNA sequence data generated on the Illumina sequencing platform. There are many situations where A5 is not the right tool for the job. In order to produce accurate results, A5 requires Illumina data with certain characteristics. A5 will likely not work well with Illumina reads shorter than around 80nt, or reads where the base qualities are low in all or most reads before 60nt. A5 assumes it is assembling homozygous haploid genomes. Use a different assembler for metagenomes and heterozygous diploid or polyploid organisms. Use a different assembler if a tool like FastQC reports your data quality is dubious. You've been warned!

```
        $OUTBASE = $ARGV[2];
        open(TMPLIBFILE, ">$OUTBASE.tmplibs");
        print TMPLIBFILE "[LIB]\n";
        print TMPLIBFILE "p1=$ARGV[0]\n";
        print TMPLIBFILE "p2=$ARGV[1]\n";
        close TMPLIBFILE;
        $libfile = "$OUTBASE.tmplibs";
} else {
        $OUTBASE = $ARGV[1];
        my $file = $ARGV[0];
        my $file_type = `file $file`;
        my $first_line = "";
        if ($file_type =~ /gzip/){
                $first_line = `gunzip -c $file | head -n 1`;
        } elsif ($file_type =~ /bzip2/) {
                $first_line = `bunzip2 -c $file | head -n 1`;
        } else {
                $first_line = `head -n 1 $file`;
        }
        if ($first_line =~ /^@/){ # assume interleaved
                open(TMPLIBFILE, ">$OUTBASE.tmplibs");
                print TMPLIBFILE "[LIB]\n";
                print TMPLIBFILE "shuf=$ARGV[0]\n";
                close TMPLIBFILE;
                $libfile = "$OUTBASE.tmplibs";
        } elsif ($first_line =~ /^\[LIB\]/) {
                $libfile = $ARGV[0];
        } else {
                print STDERR "$file is neither a library file nor a fastq file.\n";
                exit;
```