Session 2.1 – Secuenciación de genomas bacterianos: Aplicaciones

Isabel Cuesta

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

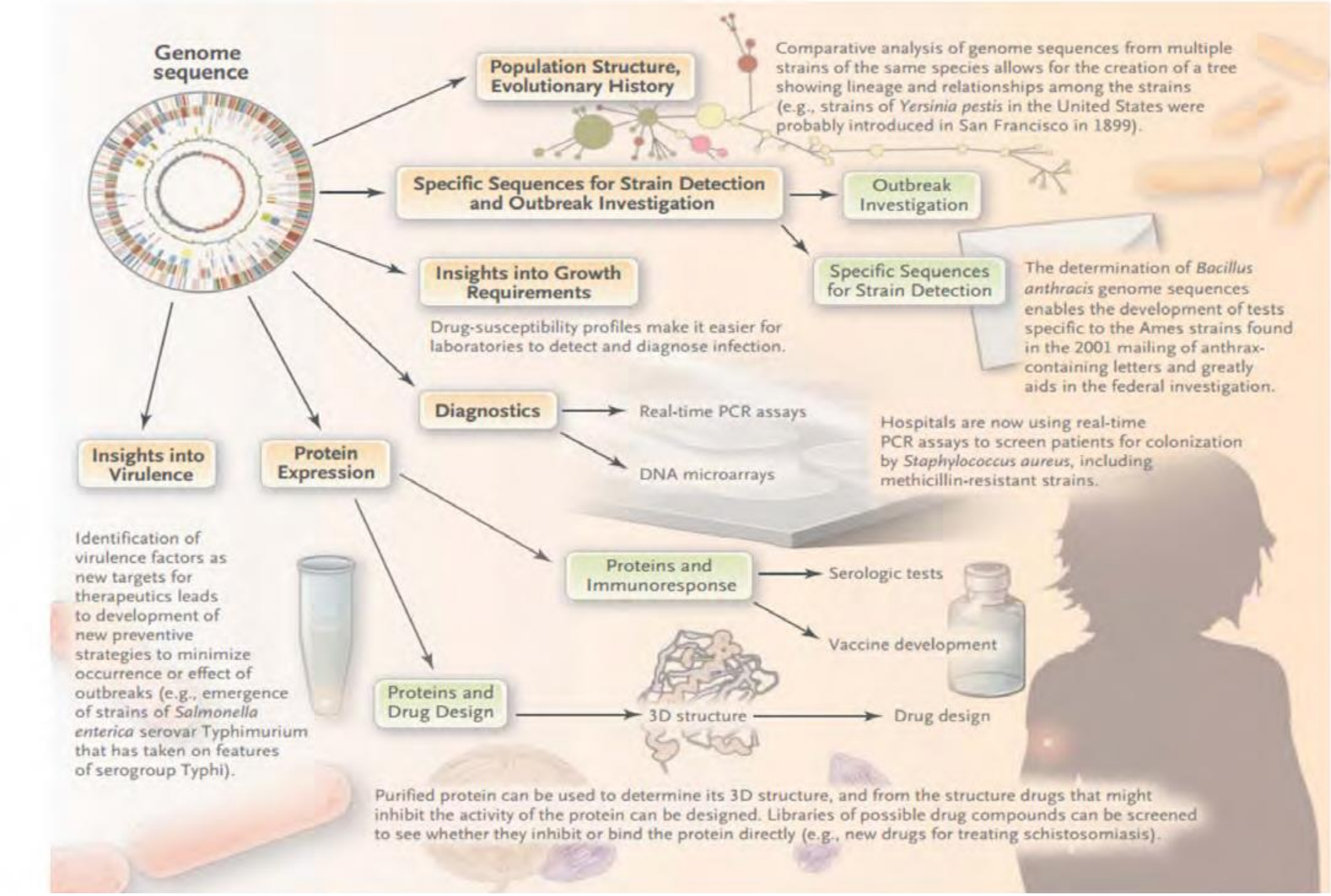
04-15 Noviembre 2019, 2ª Edición
Programa Formación Continua, ISCIII

Index

- Information obtained from WGS
- Microbial bioinformatics analysis
- Clinical use of WGS
- Considerations for implementing WGS in the laboratory routine

Use of microbial genomics for tool development

Report from The American Academy of Microbiology, 2015


Genome sequence
Population Structure, Evolutionary History
Specific Sequences for Strain Detection and Outbreak Investigation
Outbreak Investigation
Specific Sequences for Strain Detection
Insights into Growth Requirements
Diagnostics
Real-time PCR assays
DNA microarrays
Insights into Virulence
Protein Expression
Proteins and Immunoresponse
Serologic tests
Vaccine development
Proteins and Drug Design
3D structure
Drug design

Comparative analysis of genome sequences from multiple strains of the same species allows for the creation of a tree showing lineage and relationships among the strains (e.g., strains of *Yersinia pestis* in the United States were probably introduced in San Francisco in 1899).

The determination of *Bacillus anthracis* genome sequences enables the development of tests specific to the Ames strains found in the 2001 mailing of anthrax-containing letters and greatly aids in the federal investigation.

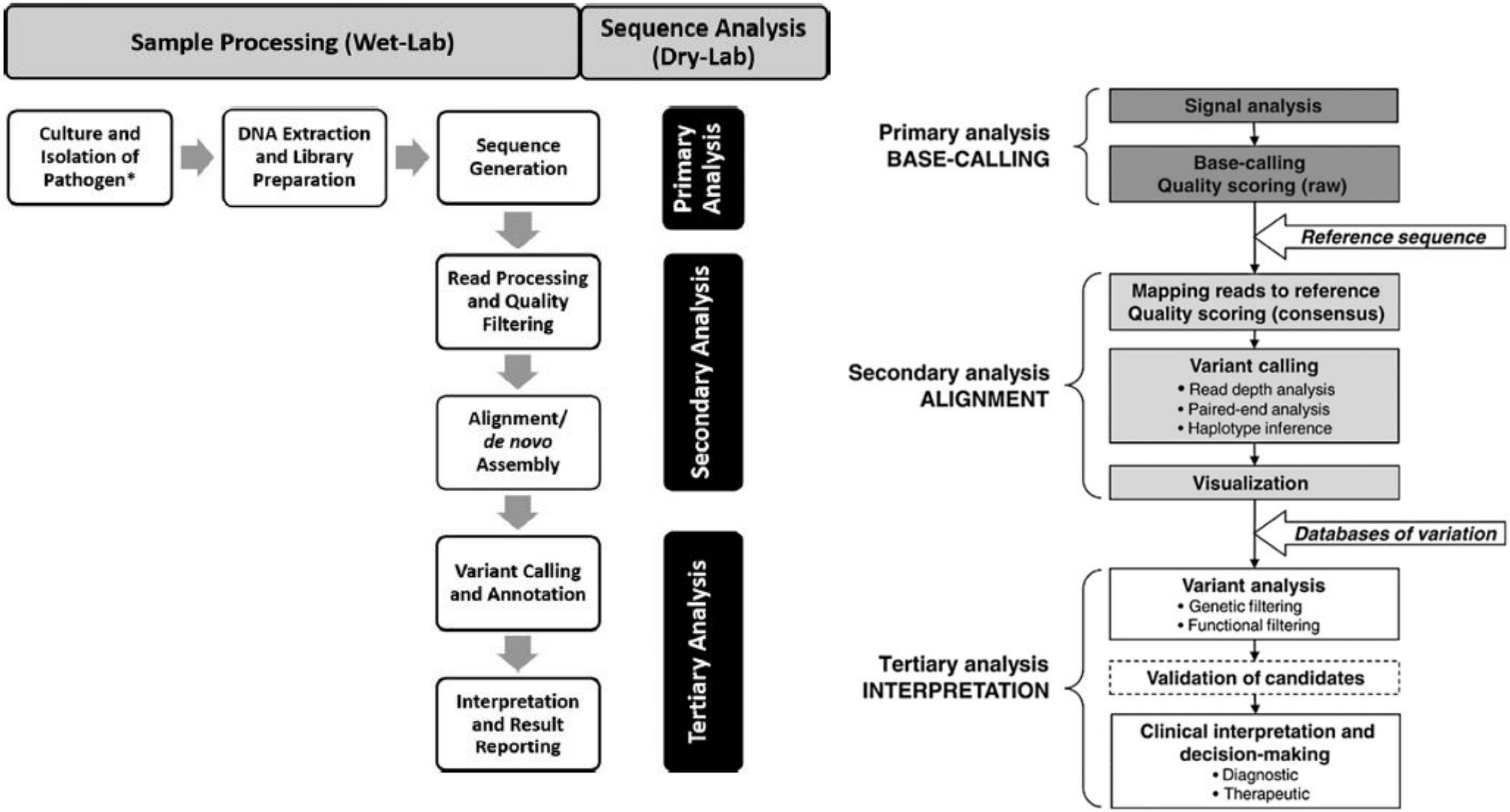
Drug-susceptibility profiles make it easier for laboratories to detect and diagnose infection.

Hospitals are now using real-time PCR assays to screen patients for colonization by *Staphylococcus aureus*, including methicillin-resistant strains.

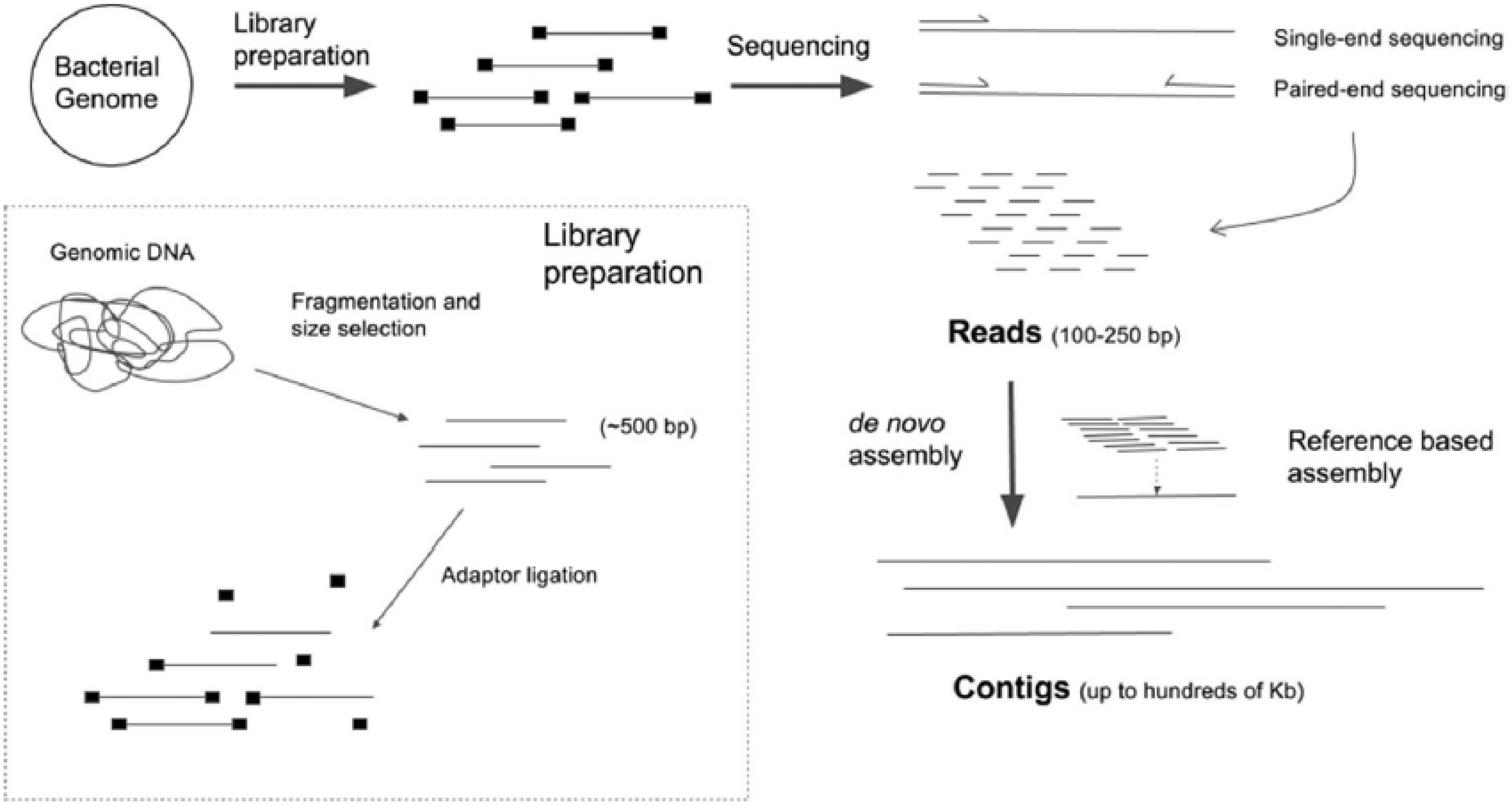
Identification of virulence factors as new targets for therapeutics leads to development of new preventive strategies to minimize occurrence or effect of outbreaks (e.g., emergence of strains of *Salmonella enterica* serovar Typhimurium that has taken on features of serogroup Typhi).

Purified protein can be used to determine its 3D structure, and from the structure drugs that might inhibit the activity of the protein can be designed. Libraries of possible drug compounds can be screened to see whether they inhibit or bind the protein directly (e.g., new drugs for treating schistosomiasis).

General NGS workflow


Sample Processing (Wet-Lab)
Sequence Analysis (Dry-Lab)
Culture and Isolation of Pathogen*
DNA Extraction and Library Preparation
Sequence Generation
Primary Analysis
Read Processing and Quality Filtering
Alignment/ *de novo* Assembly
Secondary Analysis
Variant Calling and Annotation
Interpretation and Result Reporting
Tertiary Analysis
Primary analysis BASE-CALLING
Signal analysis
Base-calling Quality scoring (raw)
Reference sequence
Secondary analysis ALIGNMENT
Mapping reads to reference Quality scoring (consensus)
Variant calling
• Read depth analysis
• Paired-end analysis
• Haplotype inference
Visualization
Databases of variation
Tertiary analysis INTERPRETATION
Variant analysis
• Genetic filtering
• Functional filtering
Validation of candidates
Clinical interpretation and decision-making
• Diagnostic
• Therapeutic

Common microbial bioinformatics analyses



Draft genome analysis

Carrico et al., CMI 2018

- Genome annotation is the process of identifying the location and biological role of genetic features present in a DNA sequence (e.g., CDS, tRNA, rTNA, operons, CRISPR elements and genomic islands.

  Software pipelines that use multiple external feature prediction algorithms

  NCBI prokaryotic genome annnotation pipeline (PGAP)

  RAST server

  PROKKA (local)
- Sequence-based microbial typing information (MLST, serotype, antimicrobial resistance or virulence genes)
- Gene-by-gene approaches: core genome = cgMLST, core + accesory genome = wgMLST

  PubMLST (https://pubmlst.org) hosts schemas

  Pasteur Institute (http://bigsdb.pasteur.fr) Listeria

  Enterobase (https://enterobase.warwick.ac.uk), Salmonella, Escherichia/Shigella, Yersinia.

  Pipelines: Genomic Profiler, ChewBBACA, TARANIS.

  Roary: presence or absence without requiring a predefined schema

  Neptune: identify differential abundance of genomic regions without requiring genome annotation information

Read mapping approaches

Carrico et al., CMI 2018

- Phylogenetic relation based on SNVs identification: if multiple strains are mapped against a single reference genome, the common variants can be used to produce a phylogenetic tree.

  Software: CFSAN SNP Pipeline, Snippy, Lyve-Set, SNVPhyl, WGSOutbraker

Challenges

- SNVs in recombinant regions or in multiple-copy regions of the genome may mask the true phylogenetic signal.
- Regions with high SNV density have a high likelihood of having a recombinant origin.

Read mapping approaches

- Reads to type: using a set of target genes or genomic regions instead os using a complete or draft genome as reference for mapping.
    - MLST, antimicrobial resistance genes, virulence genes

  Software: SRST2, ReMatCh

Advantages over gen-by-gene approach

The ability to analyse noncoding regions, which has the potential to reveal changes in gene regulatory regions, that can be translated in phenotypic differences → limited by the reference genome used

Challenges

- Phenotype inference from genotypic data needs to be validated on a case-by-case basis

Carrico et al., CMI 2018

Visualization software

Epidemiologic and genomic data integration

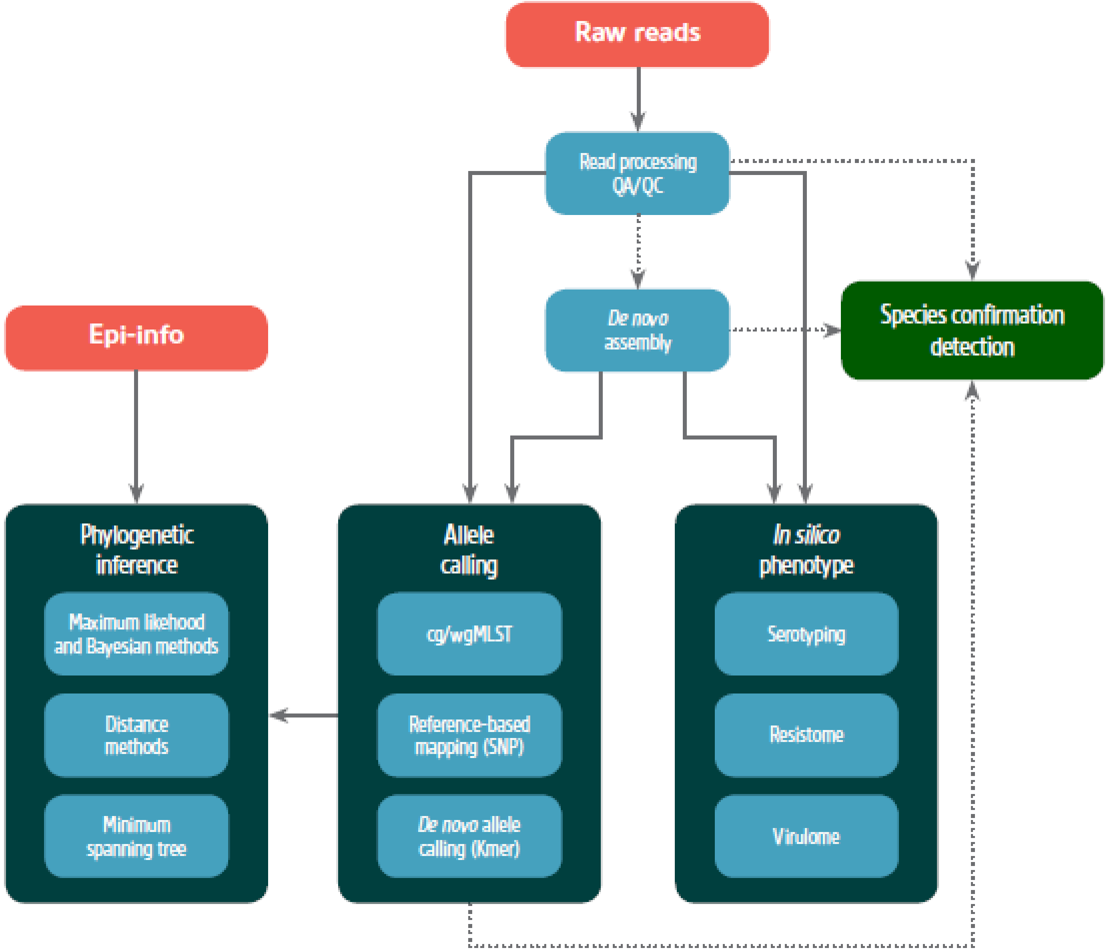
- Microreact: web service where phylogenetic tres and associated geographic, genetic and epidemiologic data can be uploaded, visualized and dynamically explored, which promotes the sharing of large data sets in the platform.
- GenGIS 2: geospatial data analysis that can use the tres and epidemiologic data provided by the user.

Allelic profiles derived from gene-by-gene methods or SNV analysis

- PHYLOVIZ 2.0: representation of mínimum spanning tres or hierarchical clustering together with associated metadata.
- PHYLOVIZ online: web browser, allows data sharing and visualization of large tres and metadata

Carrico et al., CMI 2018

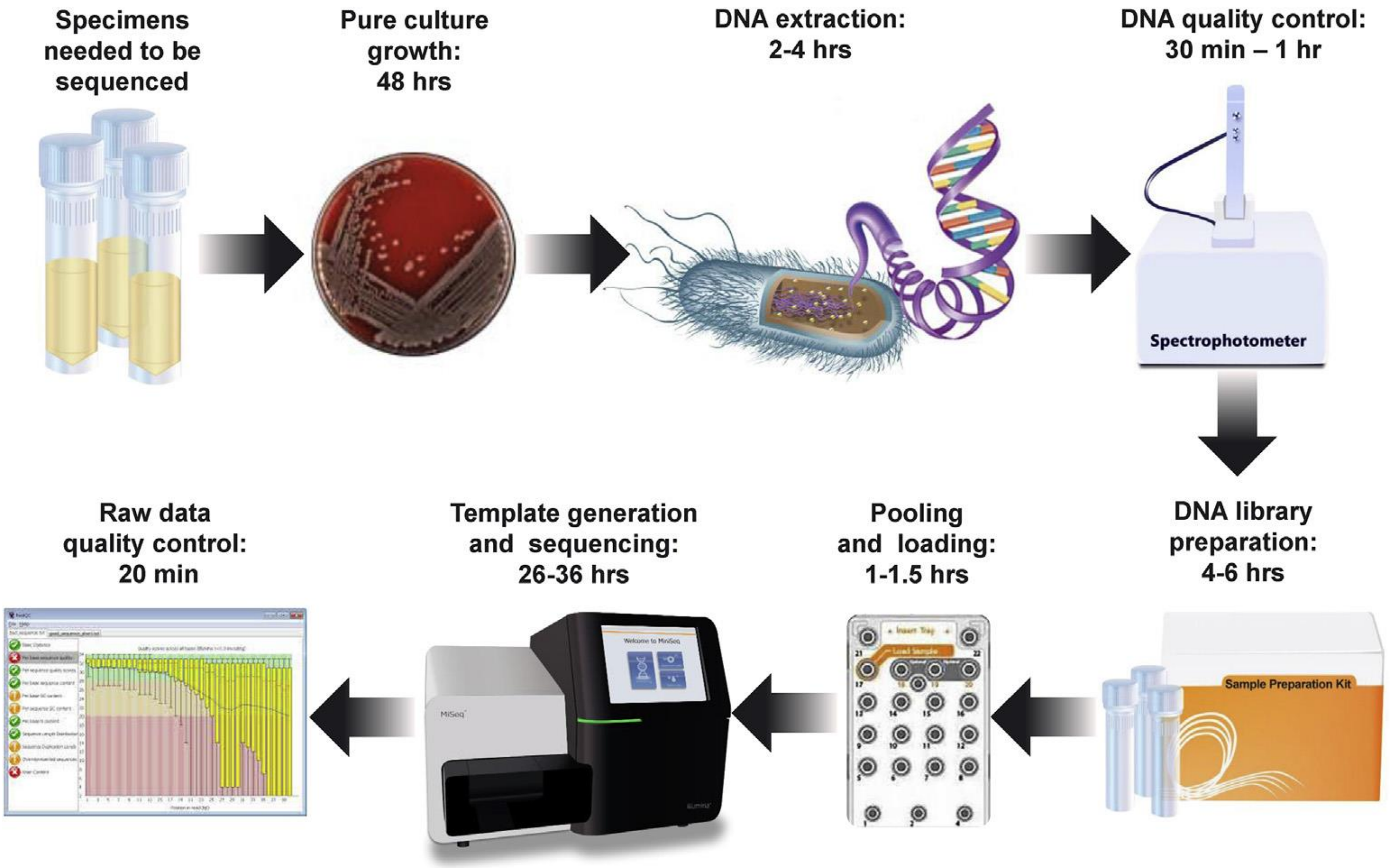
Schematic representation of WGS pipeline



Software available for NGS, pros & cons

	User access	OS	Large-scale	Data-adaptable	licencie	e.g.
Web Service	Web	Browser	Limited computation	Limited or not-configurable	free	Galaxy, Different webs
Commercial	Graphic interface	Windows	Not HPC	Limited or not-configurable	$ / €	Geneious, Bionumerics, Seqshere, Lasergene ….
Open-source	Comand line	Linux	HPC	Configurable	Open code $/ € ??	Most of them

Typical whole genome sequencing workflow in a clinical or public health laboratory



Specimens needed to be sequenced

Pure culture growth: 48 hrs

DNA extraction: 2-4 hrs

DNA quality control: 30 min – 1 hr

DNA library preparation: 4-6 hrs

Pooling and loading: 1-1.5 hrs

Template generation and sequencing: 26-36 hrs

Raw data quality control: 20 min

Besser et al., Clin Micr Infect, 2018

Applications (Clinical) of microbial genomics

1. **Those requiring bacterial isolate**

   1. Species identification
   (MALDI-TOF MS, rapid, but fails to identify unusual species)

   2. Bacterial typing (WGS vs PFGE, MLVA, MLST)
      2.1. Transmission pathways
      2.2. Outbreak monitoring
      i.e. E coli O104:H4, WGS is key to understanding the determinants and modelling the evolutionary events that may lead to a hypervirulent strain.

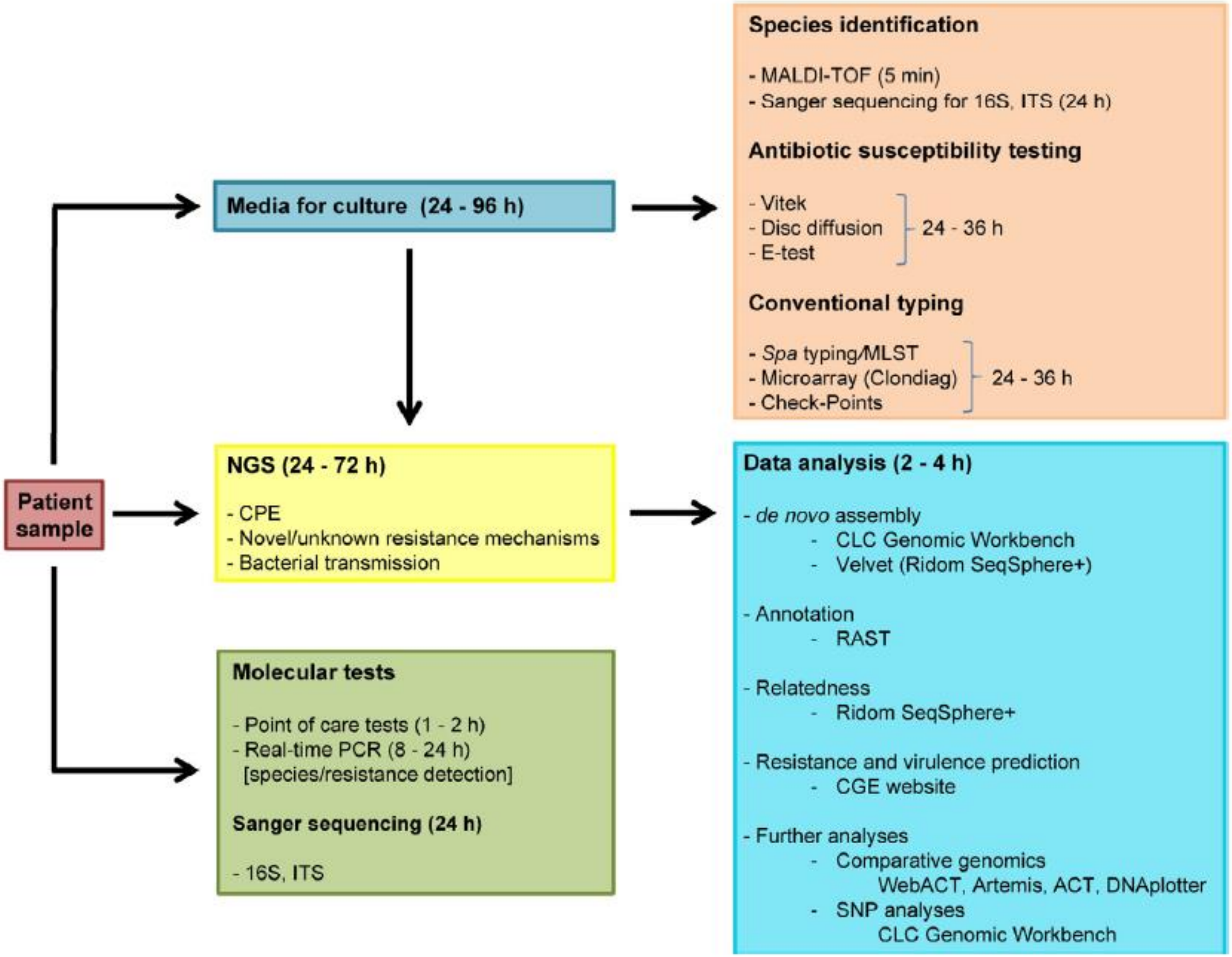
   3. Determination of virulence factors, resistance genes

2. **Those applied directly on the sample**
   1. Metagenomics
   2. Community profiling

   For direct detection of known or unknown disease-associated pathogens in clinical specimens; for investigation of microbial population diversity

A schematic overview of the general workflow of diagnostic procedures including NGS in a hospital from Netherlands

Deurenberg et al., Jbiotech, 2017



Patient sample

Media for culture (24 - 96 h)

Species identification

- MALDI-TOF (5 min)
- Sanger sequencing for 16S, ITS (24 h)

Antibiotic susceptibility testing

- Vitek
- Disc diffusion
- E-test

24 - 36 h

Conventional typing

- *Spa* typing/MLST
- Microarray (Clondiag)
- Check-Points

24 - 36 h

NGS (24 - 72 h)

- CPE
- Novel/unknown resistance mechanisms
- Bacterial transmission

Data analysis (2 - 4 h)

- *de novo* assembly
  - CLC Genomic Workbench
  - Velvet (Ridom SeqSphere+)
- Annotation
  - RAST
- Relatedness
  - Ridom SeqSphere+
- Resistance and virulence prediction
  - CGE website
- Further analyses
  - Comparative genomics
    WebACT, Artemis, ACT, DNAplotter
  - SNP analyses
    CLC Genomic Workbench

Molecular tests

- Point of care tests (1 - 2 h)
- Real-time PCR (8 - 24 h)
  [species/resistance detection]

Sanger sequencing (24 h)

- 16S, ITS

Practical issues in implementing wgs in routine diagnostic microbiology

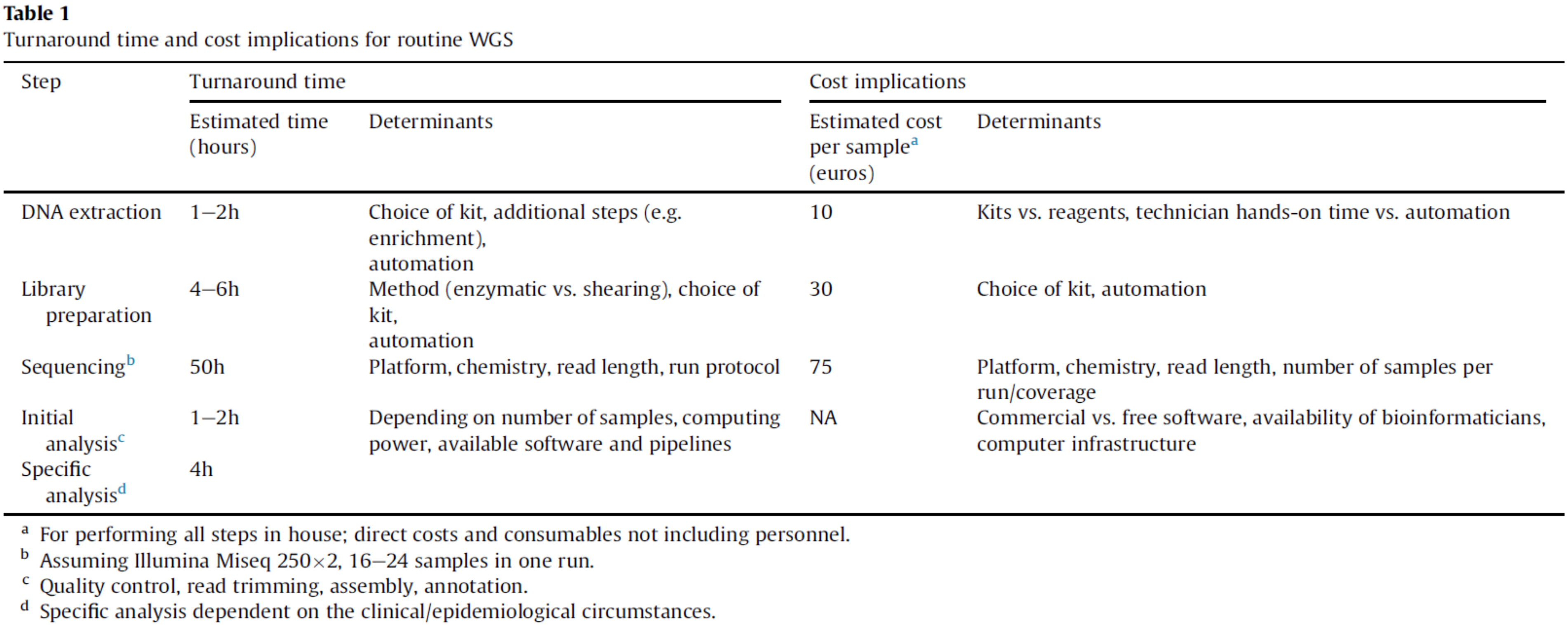


Table 1
Turnaround time and cost implications for routine WGS

Step	Turnaround time		Cost implications	
	Estimated time (hours)	Determinants	Estimated cost per sample[a] (euros)	Determinants
DNA extraction	1–2h	Choice of kit, additional steps (e.g. enrichment), automation	10	Kits vs. reagents, technician hands-on time vs. automation
Library preparation	4–6h	Method (enzymatic vs. shearing), choice of kit, automation	30	Choice of kit, automation
Sequencing[b]	50h	Platform, chemistry, read length, run protocol	75	Platform, chemistry, read length, number of samples per run/coverage
Initial analysis[c]	1–2h	Depending on number of samples, computing power, available software and pipelines	NA	Commercial vs. free software, availability of bioinformaticians, computer infrastructure
Specific analysis[d]	4h			

[a] For performing all steps in house; direct costs and consumables not including personnel.

[b] Assuming Illumina Miseq 250×2, 16–24 samples in one run.

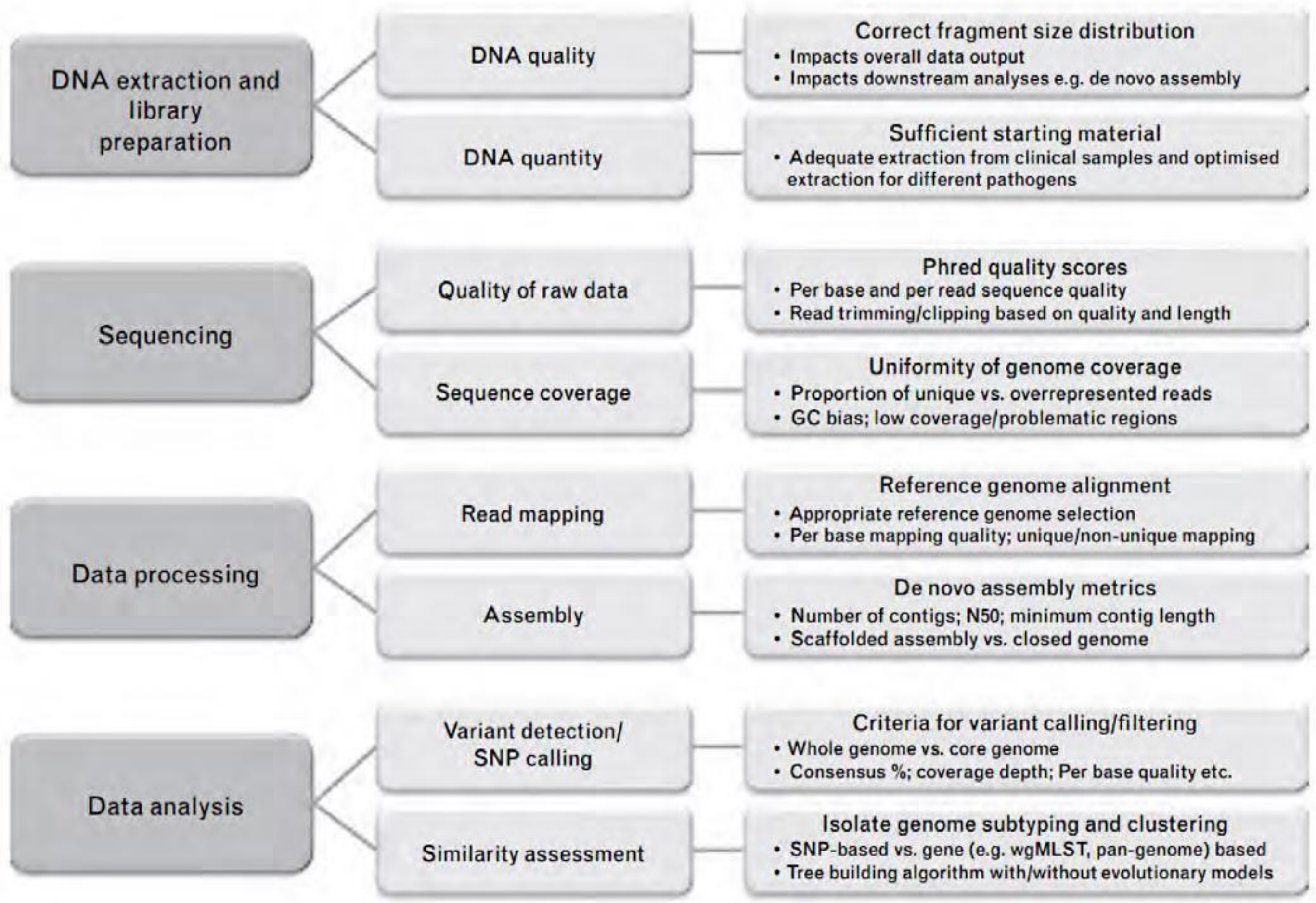
[c] Quality control, read trimming, assembly, annotation.

[d] Specific analysis dependent on the clinical/epidemiological circumstances.

Rossen et al., CMI 2018

Quality assessment considerations for WGS analyses

Applications of Clinical Microbial NGS, AASM, 2016



DNA extraction and library preparation

- DNA quality
  - Correct fragment size distribution
    - Impacts overall data output
    - Impacts downstream analyses e.g. de novo assembly
- DNA quantity
  - Sufficient starting material
    - Adequate extraction from clinical samples and optimised extraction for different pathogens

Sequencing

- Quality of raw data
  - Phred quality scores
    - Per base and per read sequence quality
    - Read trimming/clipping based on quality and length
- Sequence coverage
  - Uniformity of genome coverage
    - Proportion of unique vs. overrepresented reads
    - GC bias; low coverage/problematic regions

Data processing

- Read mapping
  - Reference genome alignment
    - Appropriate reference genome selection
    - Per base mapping quality; unique/non-unique mapping
- Assembly
  - De novo assembly metrics
    - Number of contigs; N50; minimum contig length
    - Scaffolded assembly vs. closed genome

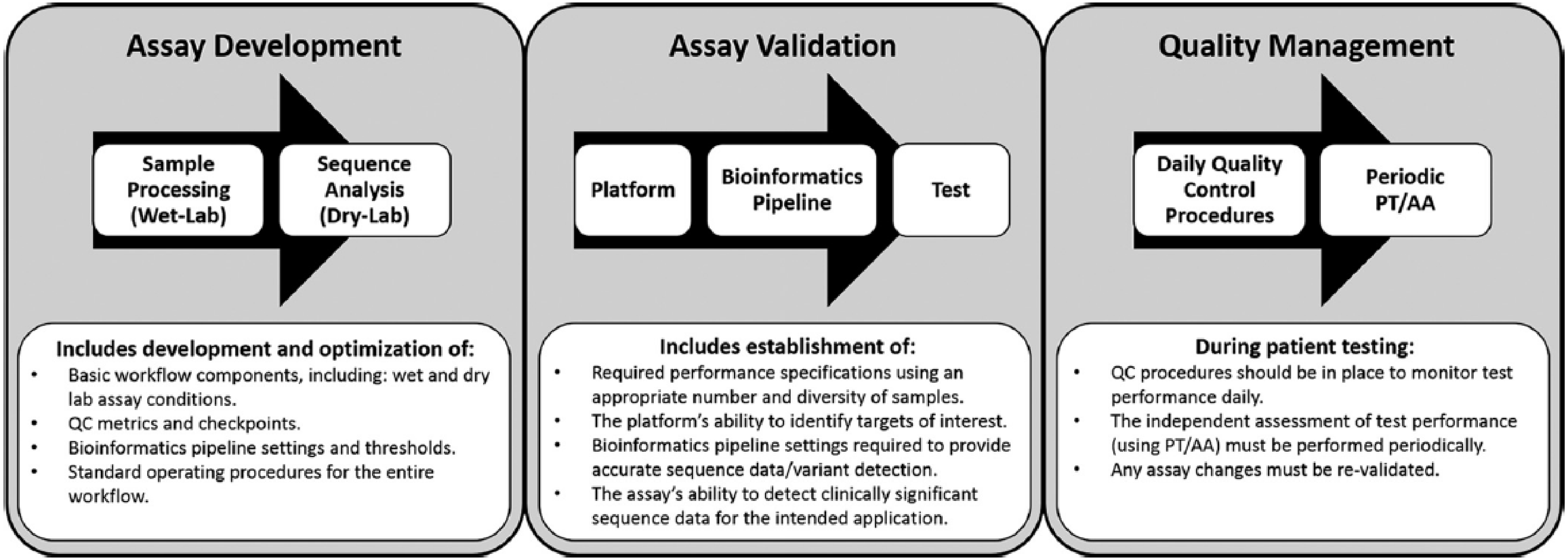
Data analysis

- Variant detection/ SNP calling
  - Criteria for variant calling/filtering
    - Whole genome vs. core genome
    - Consensus %; coverage depth; Per base quality etc.
- Similarity assessment
  - Isolate genome subtyping and clustering
    - SNP-based vs. gene (e.g. wgMLST, pan-genome) based
    - Tree building algorithm with/without evolutionary models

Proposed frameworks for clinical NGS implementation

Gargis et al., JCM 2016

Transition of NGS from research to the clinical and public health laboratory setting



Assay Development

Sample Processing (Wet-Lab) → Sequence Analysis (Dry-Lab)

Includes development and optimization of:

- Basic workflow components, including: wet and dry lab assay conditions.
- QC metrics and checkpoints.
- Bioinformatics pipeline settings and thresholds.
- Standard operating procedures for the entire workflow.

Assay Validation

Platform → Bioinformatics Pipeline → Test

Includes establishment of:

- Required performance specifications using an appropriate number and diversity of samples.
- The platform's ability to identify targets of interest.
- Bioinformatics pipeline settings required to provide accurate sequence data/variant detection.
- The assay's ability to detect clinically significant sequence data for the intended application.

Quality Management

Daily Quality Control Procedures → Periodic PT/AA

During patient testing:

- QC procedures should be in place to monitor test performance daily.
- The independent assessment of test performance (using PT/AA) must be performed periodically.
- Any assay changes must be re-validated.

No clinical microbiology NGS tests have been approved by the FDA, and the limited number of clinical infectious disease NGS-bases assays currently offered are being performed as LDTs

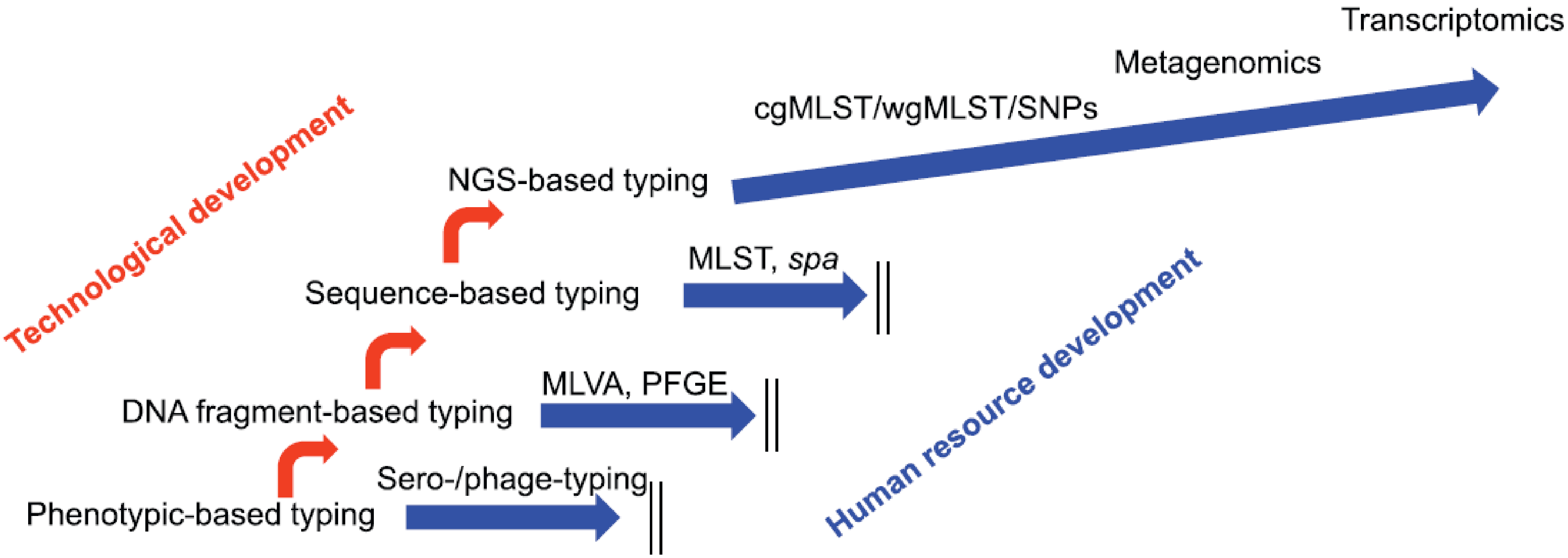
Practical issues in implementing wgs in routine diagnostic microbiology

- When and how to integrate NGS in the routine workflow?
- The place of NGS in the diagnostic hierarchy of microbiology
- Quality control issues for using NGS in microbiology
- Proficiency testing for WGS in microbiology
- Maintaining backwards compatibility with WGS

Rossen et al., CMI 2018

When and how to integrate NGS in the routine workflow?

Integrating interventional genomics in clinical microbiology settings requires development of NGS technology going hand in hand with human resource development



Rossen et al., CMI 2018

When and how to integrate NGS in the routine workflow?

Research or Surveillance:

WGS fits best in a batch-wise approach for analysing samples

Routine diagnostics:

Case-by-case approach, a balance should be kept between costs, quality, speed and complexity of the wet and dry processes.

more samples tested in parallel: costs decrease

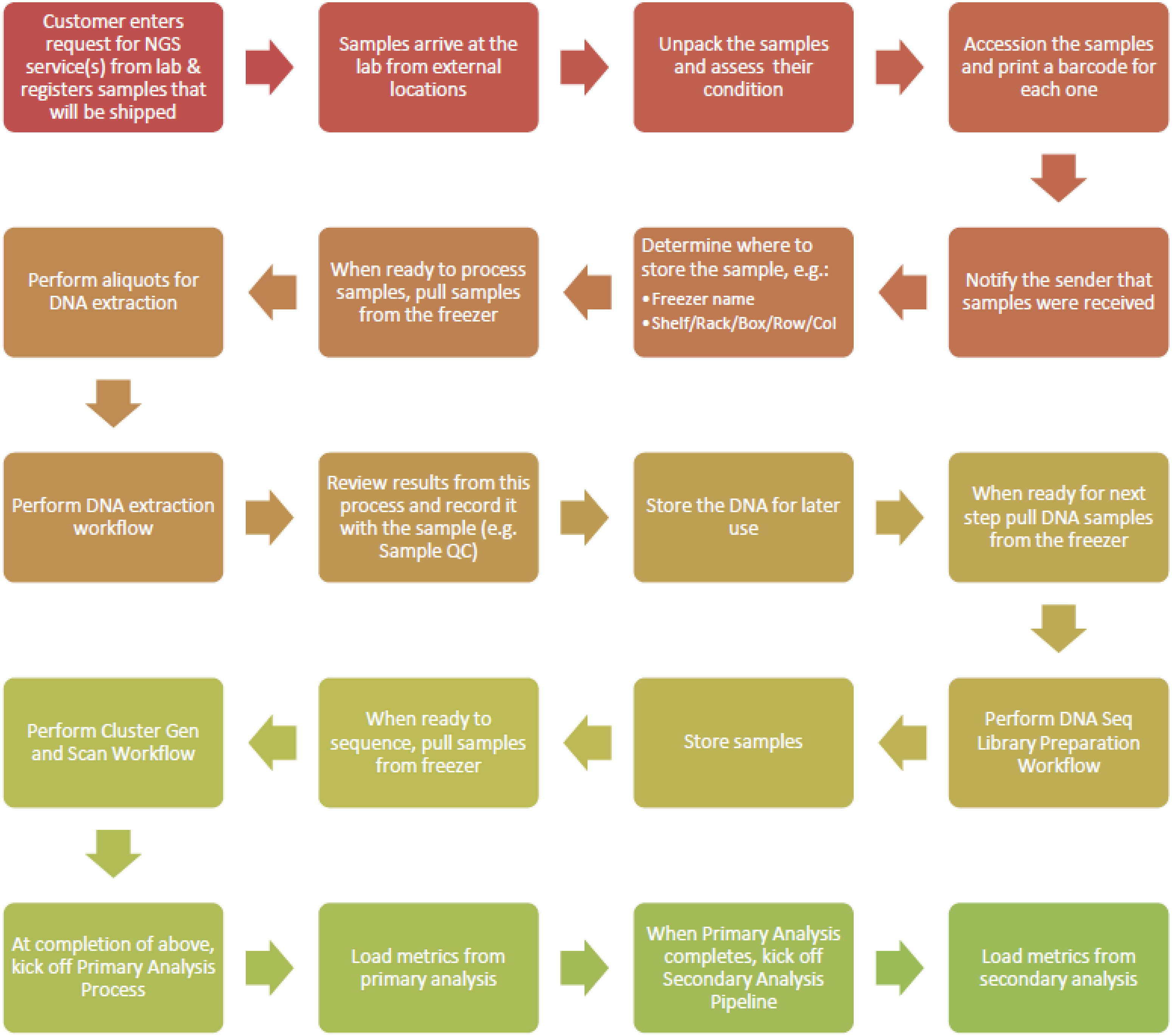
longer reads may facilitate the downstream analysis: longer running times and higher costs

Rossen et al., CMI 2018

Practical issues in implementing wgs in routine diagnostic microbiology

- When and how to integrate NGS in the routine workflow?
- The place of NGS in the diagnostic hierarchy of microbiology
- Quality control issues for using NGS in microbiology
- Proficiency testing for WGS in microbiology
- Maintaining backwards compatibility with WGS

Rossen et al., CMI 2018



Customer enters request for NGS service(s) from lab & registers samples that will be shipped

Samples arrive at the lab from external locations

Unpack the samples and assess their condition

Accession the samples and print a barcode for each one

Notify the sender that samples were received

Determine where to store the sample, e.g.:

- Freezer name
- Shelf/Rack/Box/Row/Col

When ready to process samples, pull samples from the freezer

Perform aliquots for DNA extraction

Perform DNA extraction workflow

Review results from this process and record it with the sample (e.g. Sample QC)

Store the DNA for later use

When ready for next step pull DNA samples from the freezer

Perform DNA Seq Library Preparation Workflow

Store samples

When ready to sequence, pull samples from freezer

Perform Cluster Gen and Scan Workflow

At completion of above, kick off Primary Analysis Process

Load metrics from primary analysis

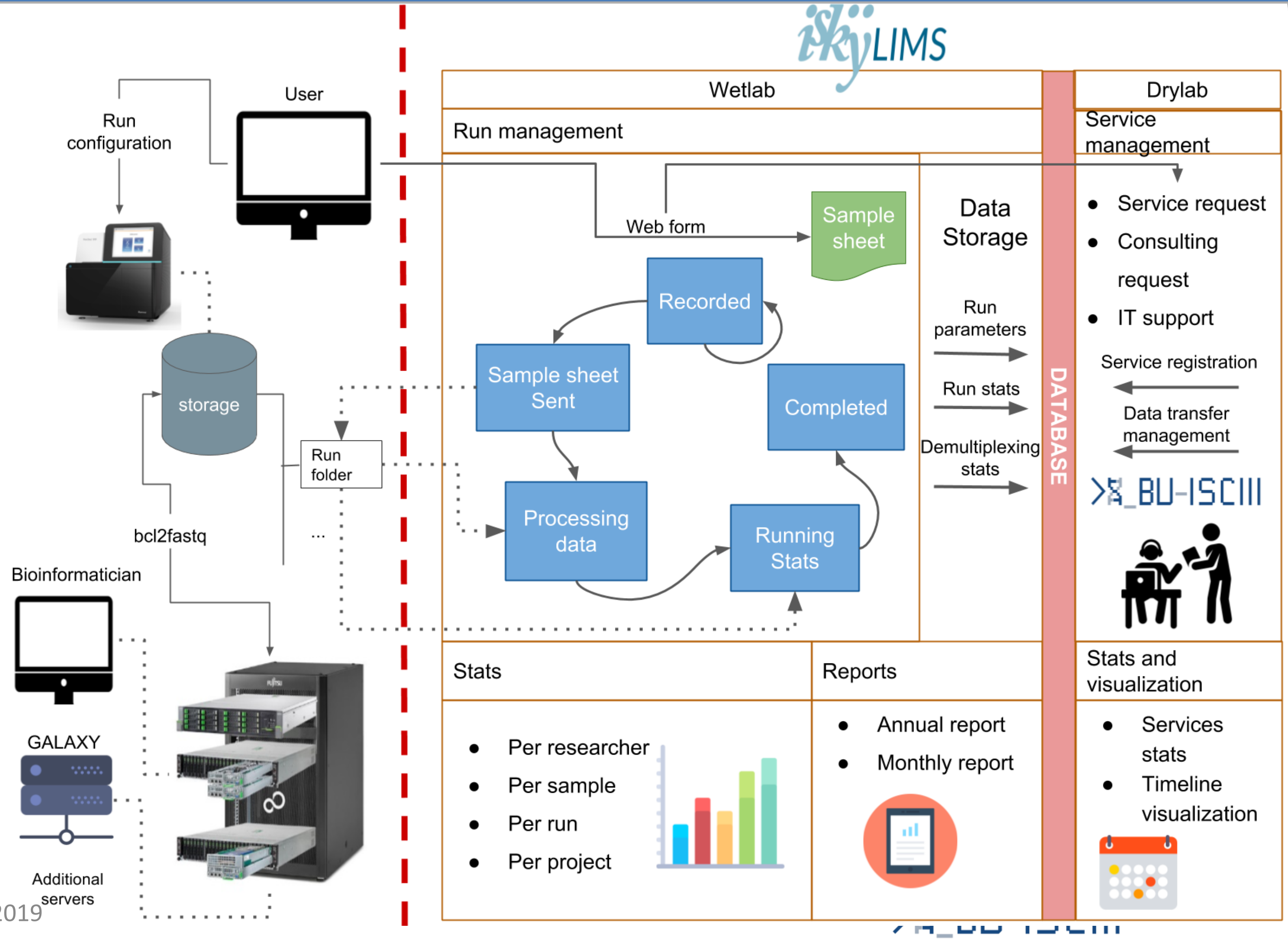
When Primary Analysis completes, kick off Secondary Analysis Pipeline

Load metrics from secondary analysis

LIMS for NGS Labs

For NGS labs (Genomics Unit), there are common sub-processes that comprise the overall workflow, these include:

- Request for Services
- Sample Receiving
- DNA/ RNA Extraction
- Library Construction
- Sequencing
- Post Sequencing Processes such as Primary and Secondary Analysis



User

Run configuration

storage

Run folder

bcl2fastq

...

Bioinformatician

GALAXY

Additional servers

Wetlab

Run management

Web form

Sample sheet

Recorded

Sample sheet Sent

Completed

Processing data

Running Stats

Data Storage

Run parameters

Run stats

Demultiplexing stats

DATABASE

Drylab

Service management

- Service request
- Consulting request
- IT support

Service registration

Data transfer management

BU-ISCIII

Stats

- Per researcher
- Per sample
- Per run
- Per project

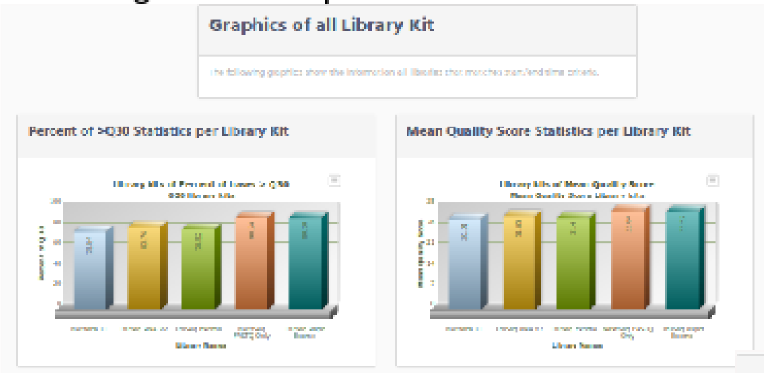
Reports

- Annual report
- Monthly report

Stats and visualization

- Services stats
- Timeline visualization

Collecting Statistics per RUN



Graphics of all Library Kit
Percent of >Q30 Statistics per Library Kit
Mean Quality Score Statistics per Library Kit

Fig. 3 Library kits, Samples with Q >30.

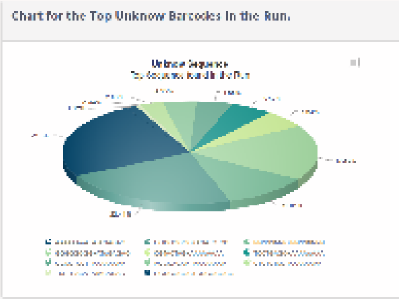


Fig. 4 Pie graphics for Unknow Barcodes.

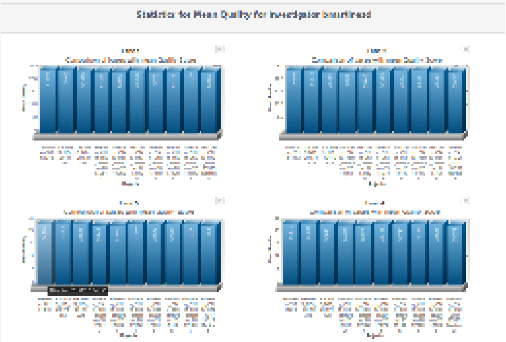
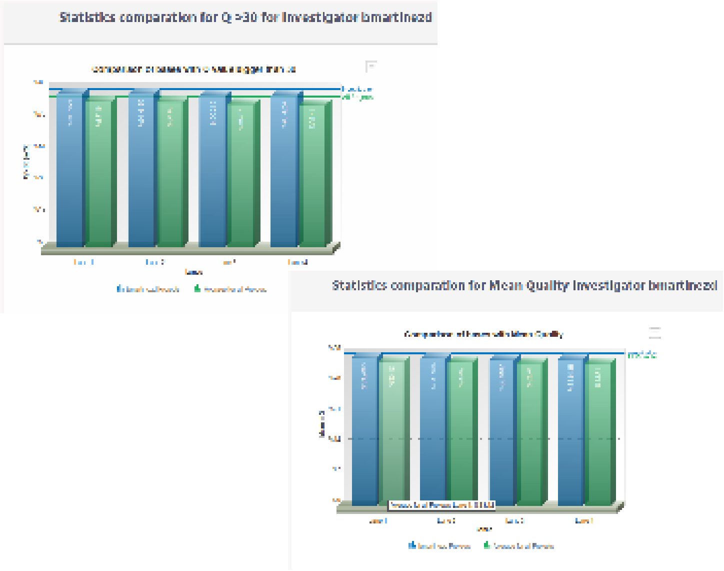
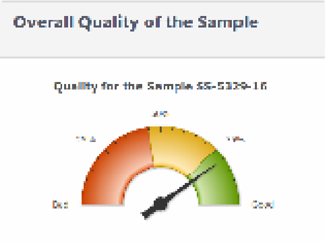


Fig. 5 Mean Quality of Investigator.



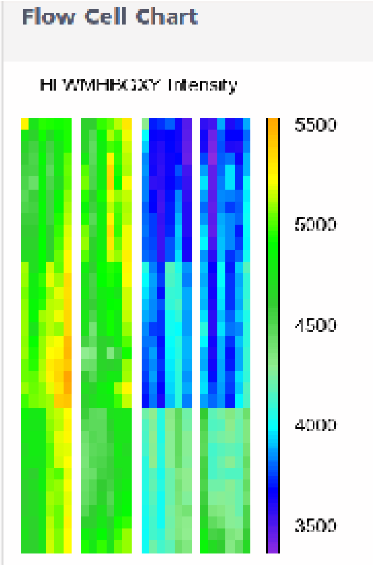
Statistics comparation for Q >30 for Investigator bmartinezd
Statistics comparation for Mean Quality Investigator bmartinezd
Comparation of Lanes with Mean Quality

Fig. 6 Quality Statistics per Investigator.



Overall Quality of the Sample
Quality for the Sample SS-S339-16
Bad
Good

Fig. 7 Overall Quality Sample.



iSkyLIMS

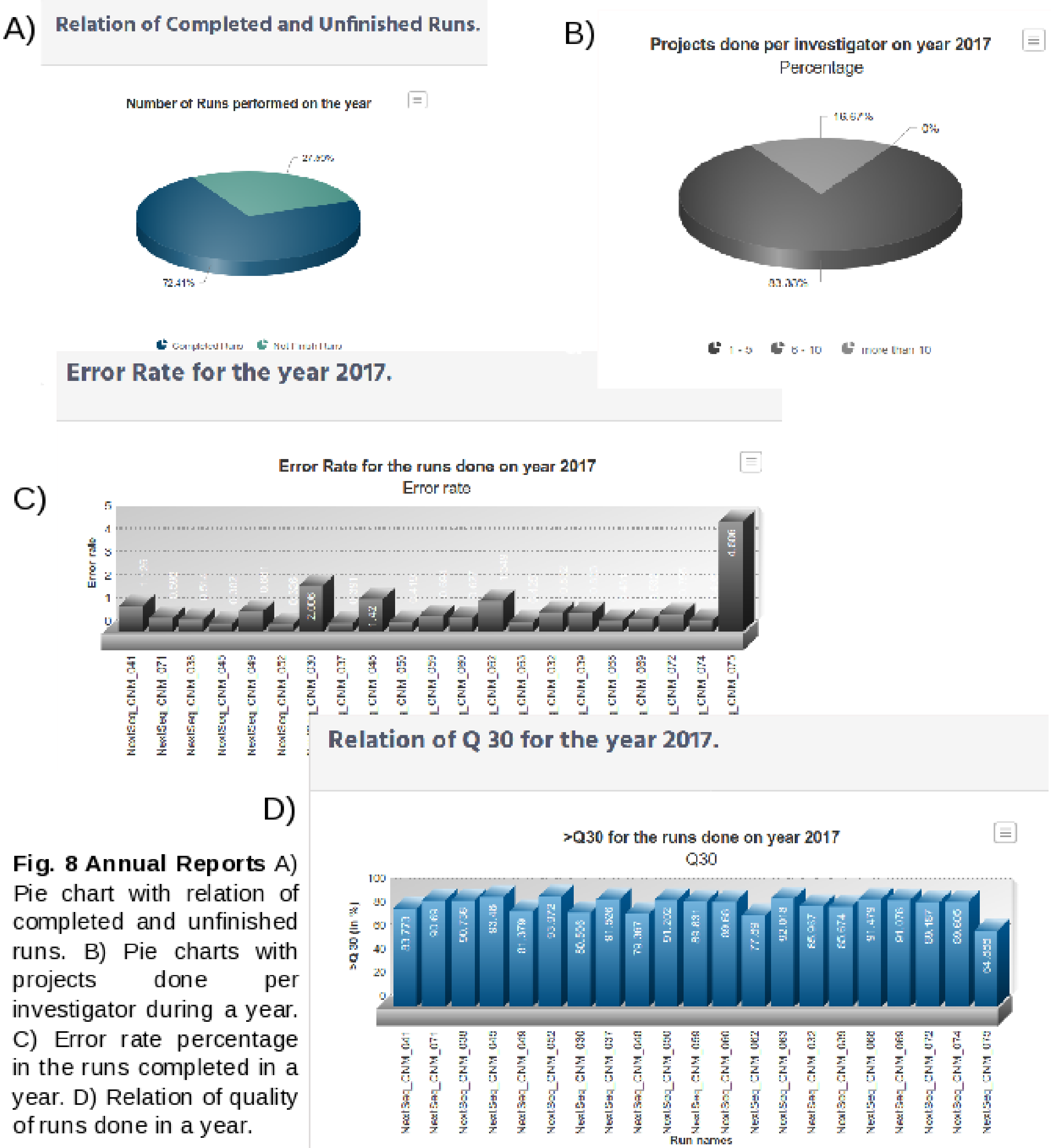
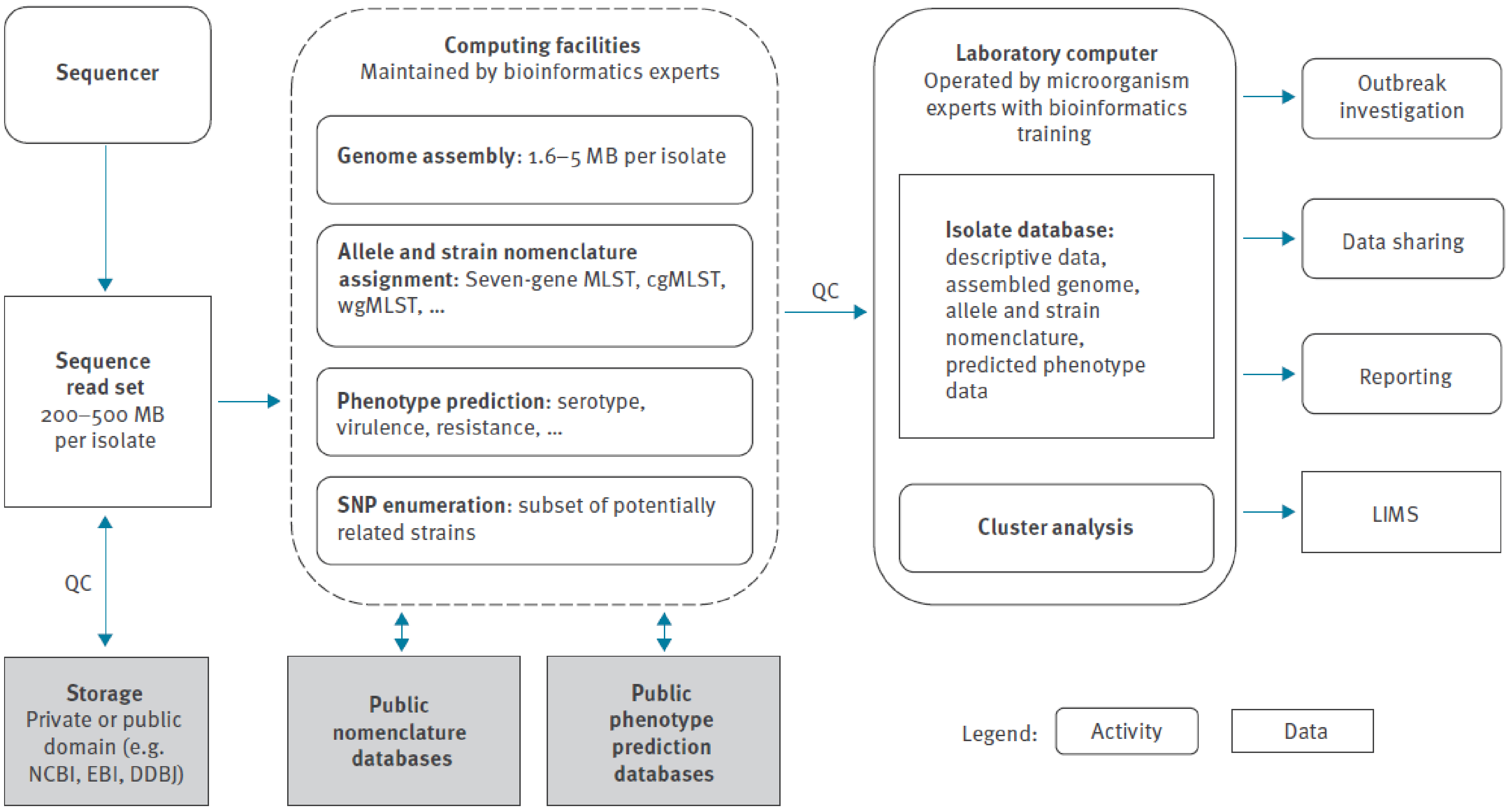


Fig. 8 Annual Reports A) Pie chart with relation of completed and unfinished runs. B) Pie charts with projects done per investigator during a year. C) Error rate percentage in the runs completed in a year. D) Relation of quality of runs done in a year.

Potential solutions for computing and storage in PulseNet International laboratories, May 2017

Nadon et al., Eurosurv. 2017



Sequencer

Sequence read set 200–500 MB per isolate

QC

Storage
Private or public domain (e.g. NCBI, EBI, DDBJ)

Computing facilities
Maintained by bioinformatics experts

Genome assembly: 1.6–5 MB per isolate

Allele and strain nomenclature assignment: Seven-gene MLST, cgMLST, wgMLST, ...

Phenotype prediction: serotype, virulence, resistance, ...

SNP enumeration: subset of potentially related strains

Public nomenclature databases

Public phenotype prediction databases

QC

Laboratory computer
Operated by microorganism experts with bioinformatics training

Isolate database: descriptive data, assembled genome, allele and strain nomenclature, predicted phenotype data

Cluster analysis

Outbreak investigation

Data sharing

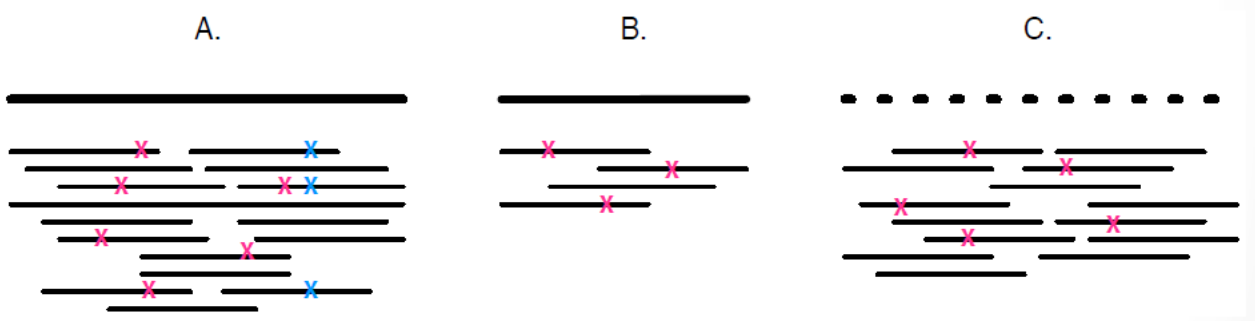
Reporting

LIMS

Legend: Activity | Data

Sequencing terms, three aims

(A) Resequencing (B) Read counting (C) Assembly



Sequencing terms

Breadth of coverage

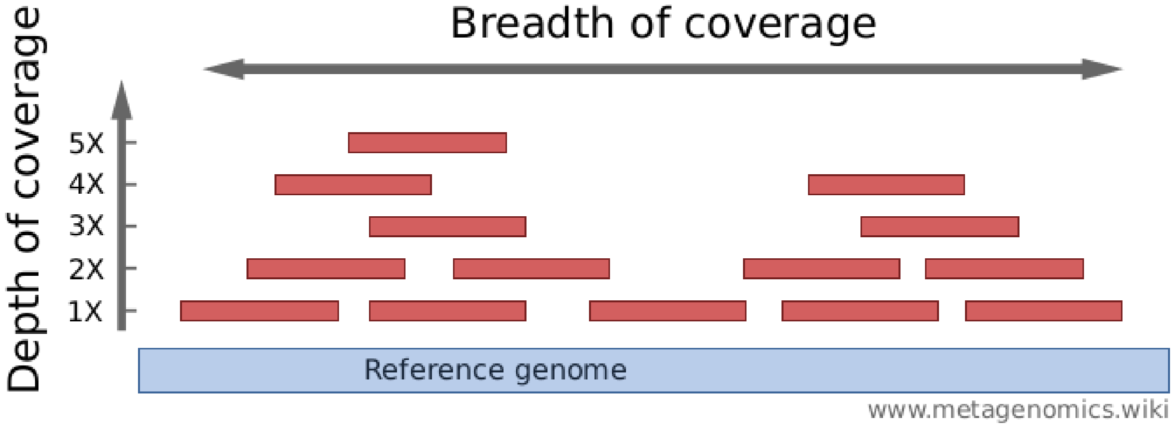
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.

Depth of coverage

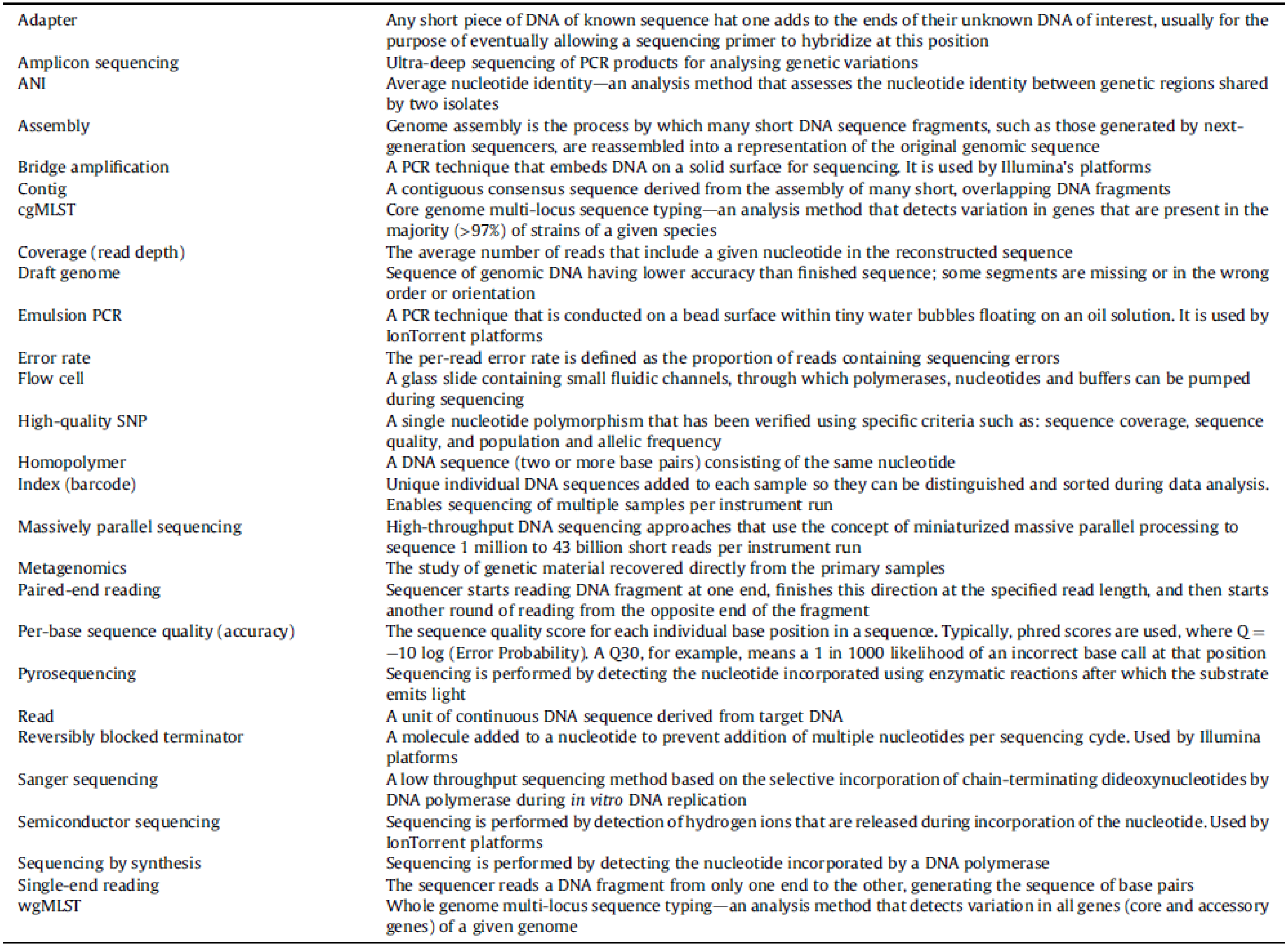
How strong is a genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).



Glossary of some commonly used sequencing terms

Besser et al., Clin Micr Infect, 2018



Adapter	Any short piece of DNA of known sequence hat one adds to the ends of their unknown DNA of interest, usually for the purpose of eventually allowing a sequencing primer to hybridize at this position
Amplicon sequencing	Ultra-deep sequencing of PCR products for analysing genetic variations
ANI	Average nucleotide identity—an analysis method that assesses the nucleotide identity between genetic regions shared by two isolates
Assembly	Genome assembly is the process by which many short DNA sequence fragments, such as those generated by next-generation sequencers, are reassembled into a representation of the original genomic sequence
Bridge amplification	A PCR technique that embeds DNA on a solid surface for sequencing. It is used by Illumina's platforms
Contig	A contiguous consensus sequence derived from the assembly of many short, overlapping DNA fragments
cgMLST	Core genome multi-locus sequence typing—an analysis method that detects variation in genes that are present in the majority (>97%) of strains of a given species
Coverage (read depth)	The average number of reads that include a given nucleotide in the reconstructed sequence
Draft genome	Sequence of genomic DNA having lower accuracy than finished sequence; some segments are missing or in the wrong order or orientation
Emulsion PCR	A PCR technique that is conducted on a bead surface within tiny water bubbles floating on an oil solution. It is used by IonTorrent platforms
Error rate	The per-read error rate is defined as the proportion of reads containing sequencing errors
Flow cell	A glass slide containing small fluidic channels, through which polymerases, nucleotides and buffers can be pumped during sequencing
High-quality SNP	A single nucleotide polymorphism that has been verified using specific criteria such as: sequence coverage, sequence quality, and population and allelic frequency
Homopolymer	A DNA sequence (two or more base pairs) consisting of the same nucleotide
Index (barcode)	Unique individual DNA sequences added to each sample so they can be distinguished and sorted during data analysis. Enables sequencing of multiple samples per instrument run
Massively parallel sequencing	High-throughput DNA sequencing approaches that use the concept of miniaturized massive parallel processing to sequence 1 million to 43 billion short reads per instrument run
Metagenomics	The study of genetic material recovered directly from the primary samples
Paired-end reading	Sequencer starts reading DNA fragment at one end, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment
Per-base sequence quality (accuracy)	The sequence quality score for each individual base position in a sequence. Typically, phred scores are used, where Q = −10 log (Error Probability). A Q30, for example, means a 1 in 1000 likelihood of an incorrect base call at that position
Pyrosequencing	Sequencing is performed by detecting the nucleotide incorporated using enzymatic reactions after which the substrate emits light
Read	A unit of continuous DNA sequence derived from target DNA
Reversibly blocked terminator	A molecule added to a nucleotide to prevent addition of multiple nucleotides per sequencing cycle. Used by Illumina platforms
Sanger sequencing	A low throughput sequencing method based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication
Semiconductor sequencing	Sequencing is performed by detection of hydrogen ions that are released during incorporation of the nucleotide. Used by IonTorrent platforms
Sequencing by synthesis	Sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase
Single-end reading	The sequencer reads a DNA fragment from only one end to the other, generating the sequence of base pairs
wgMLST	Whole genome multi-locus sequence typing—an analysis method that detects variation in all genes (core and accessory genes) of a given genome

Thanks for your attention!

Questions???