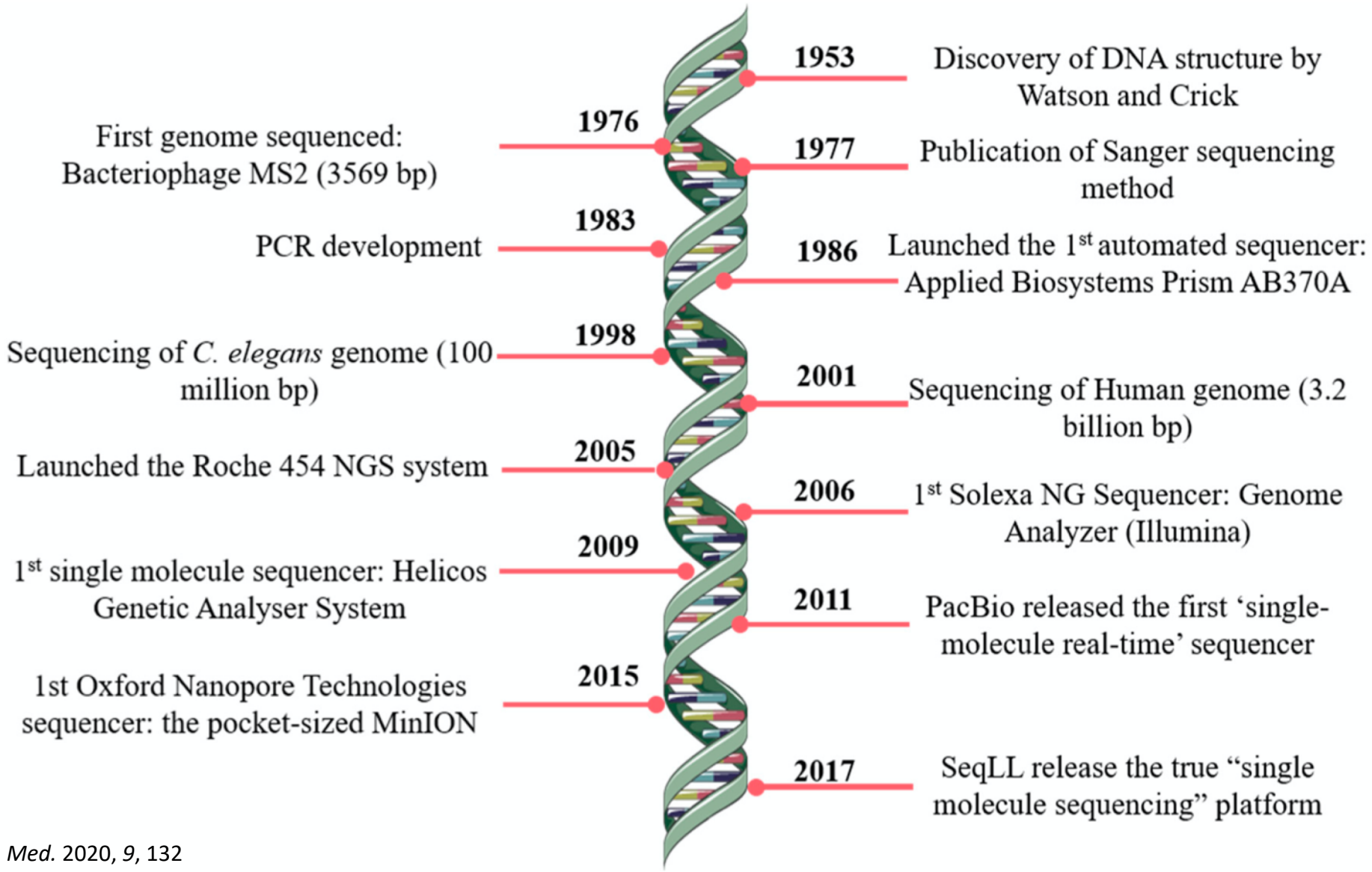
Centru Cloud si Big Data pentru participarea la Cloud-ul European pentru Stiinta Deschisa (CeCBiD-EOSC)

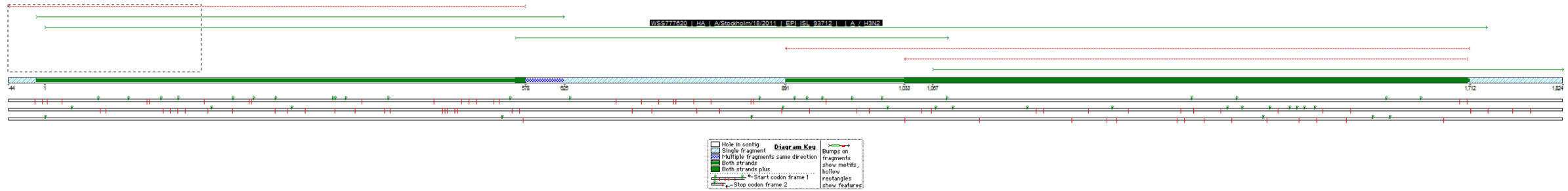
Realizarea de servicii si aplicatii informatice pentru suportul activitatii de analiza a datelor de secventiere de noua generatie

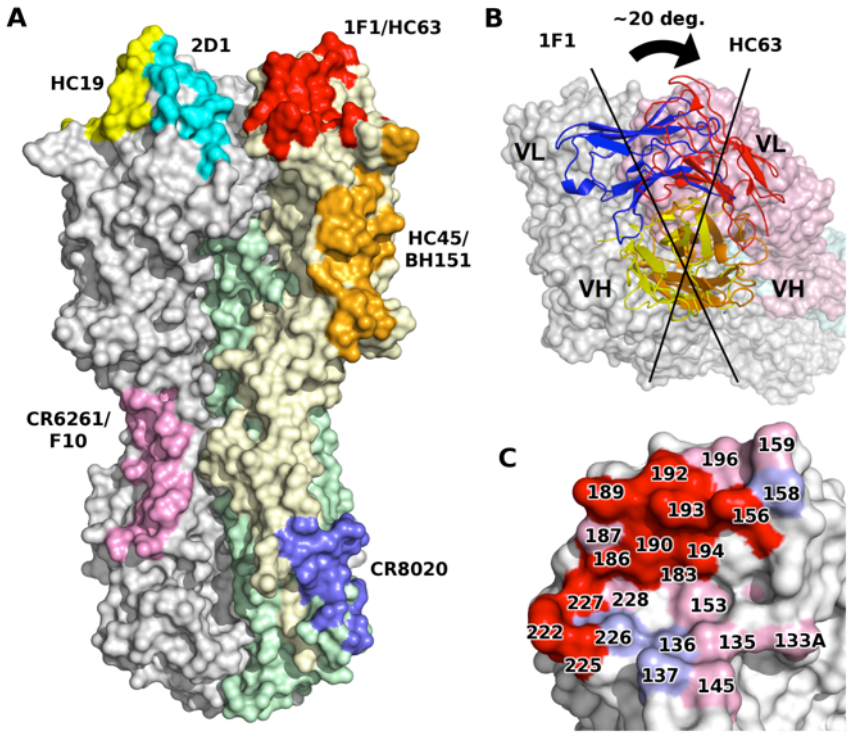
DNA sequencing timeline



R. Pereira et al. J. Clin. Med. 2020, *9*, 132

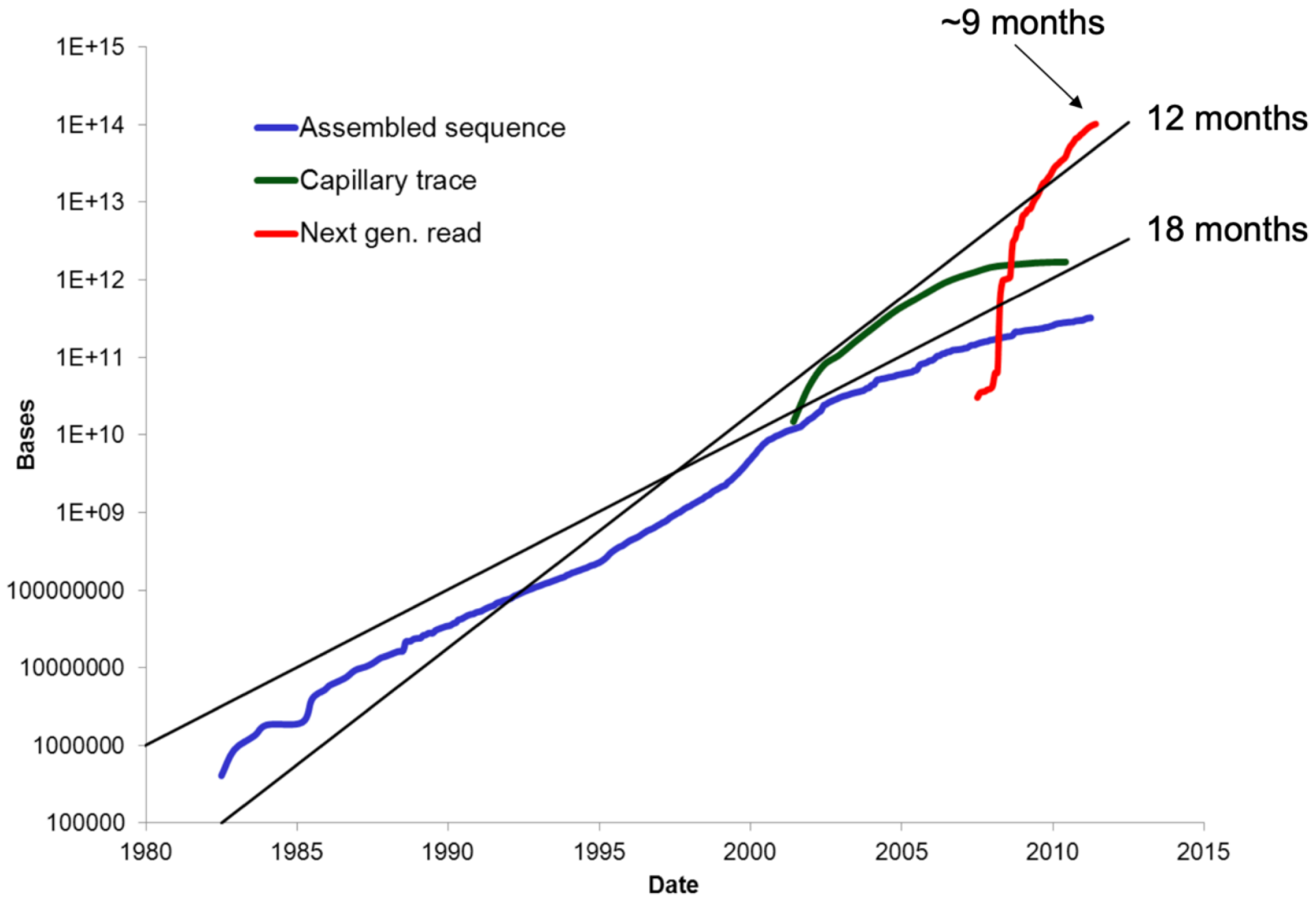
Sanger vs NGS DNA sequencing



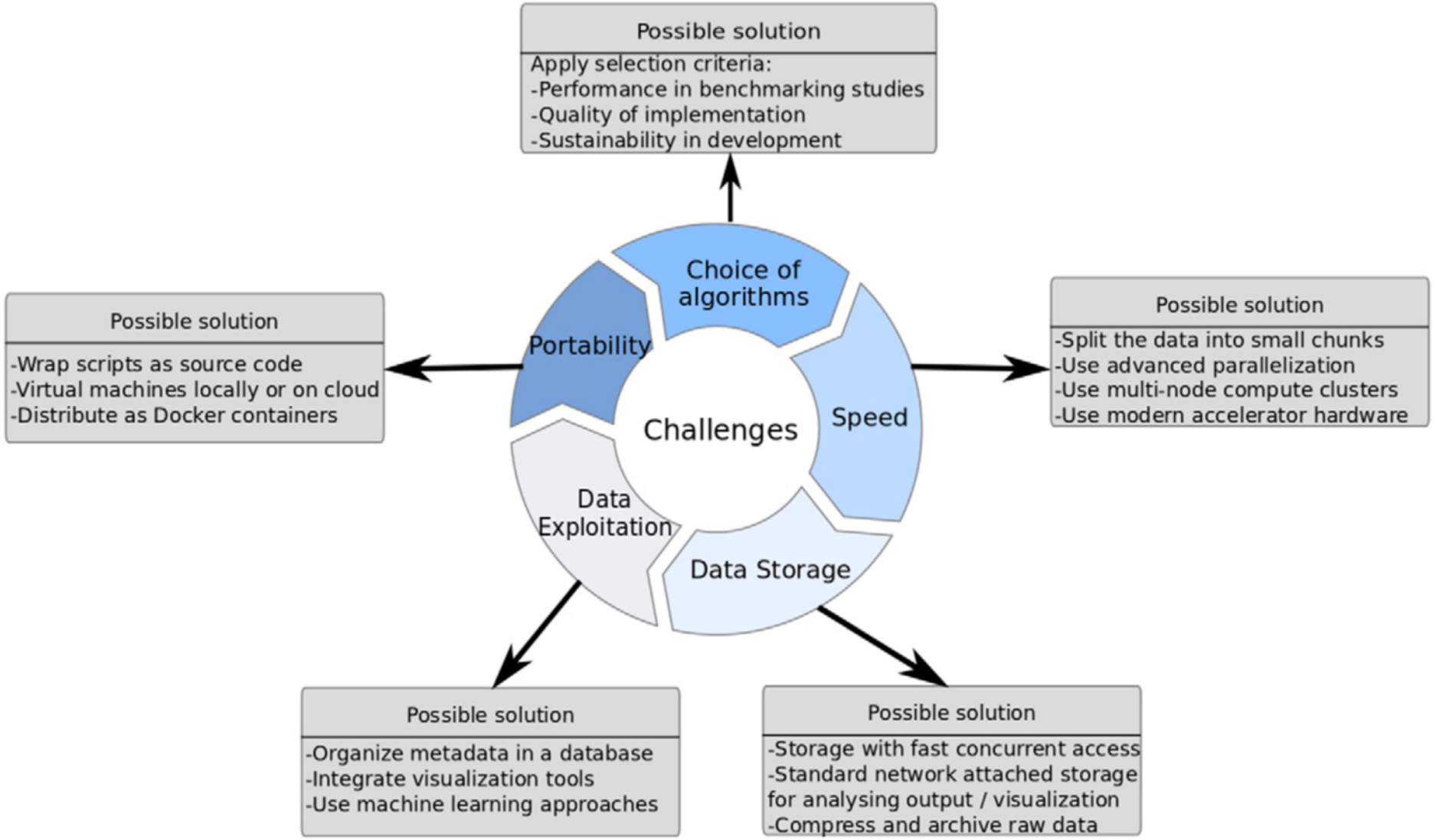


T. Tsibane et al. PLoS Pathog 8(12)

Sanger and NGS uploads to EMBL Nucleotide Sequence Database



Challenges in the design and implementation of NGS analysis workflows

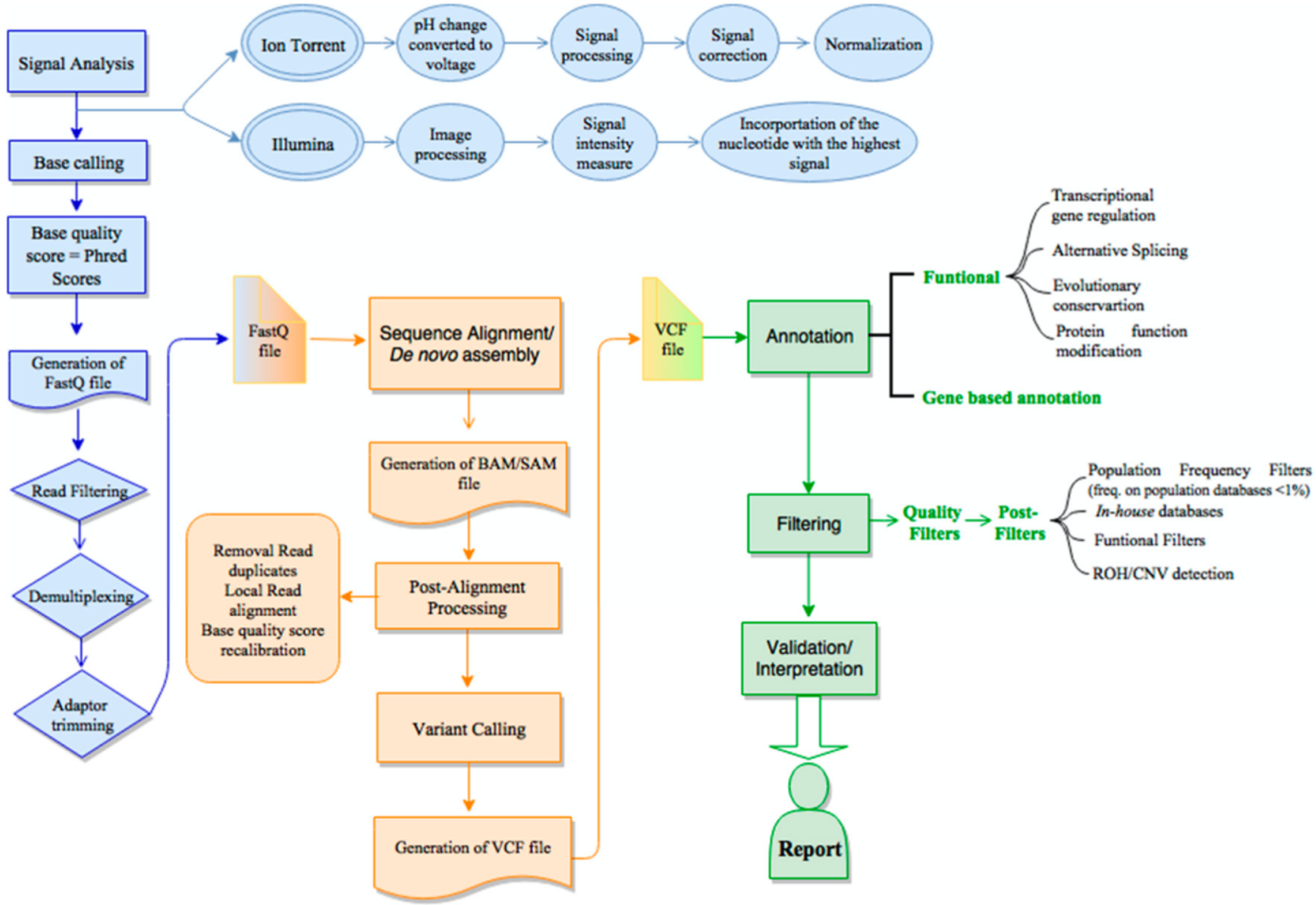


P. Kulkarni, P. Frommolt / Comput. Struct. Biotec. 15 (2017) 471-477

Tipical first steps in an NGS analysis workflows

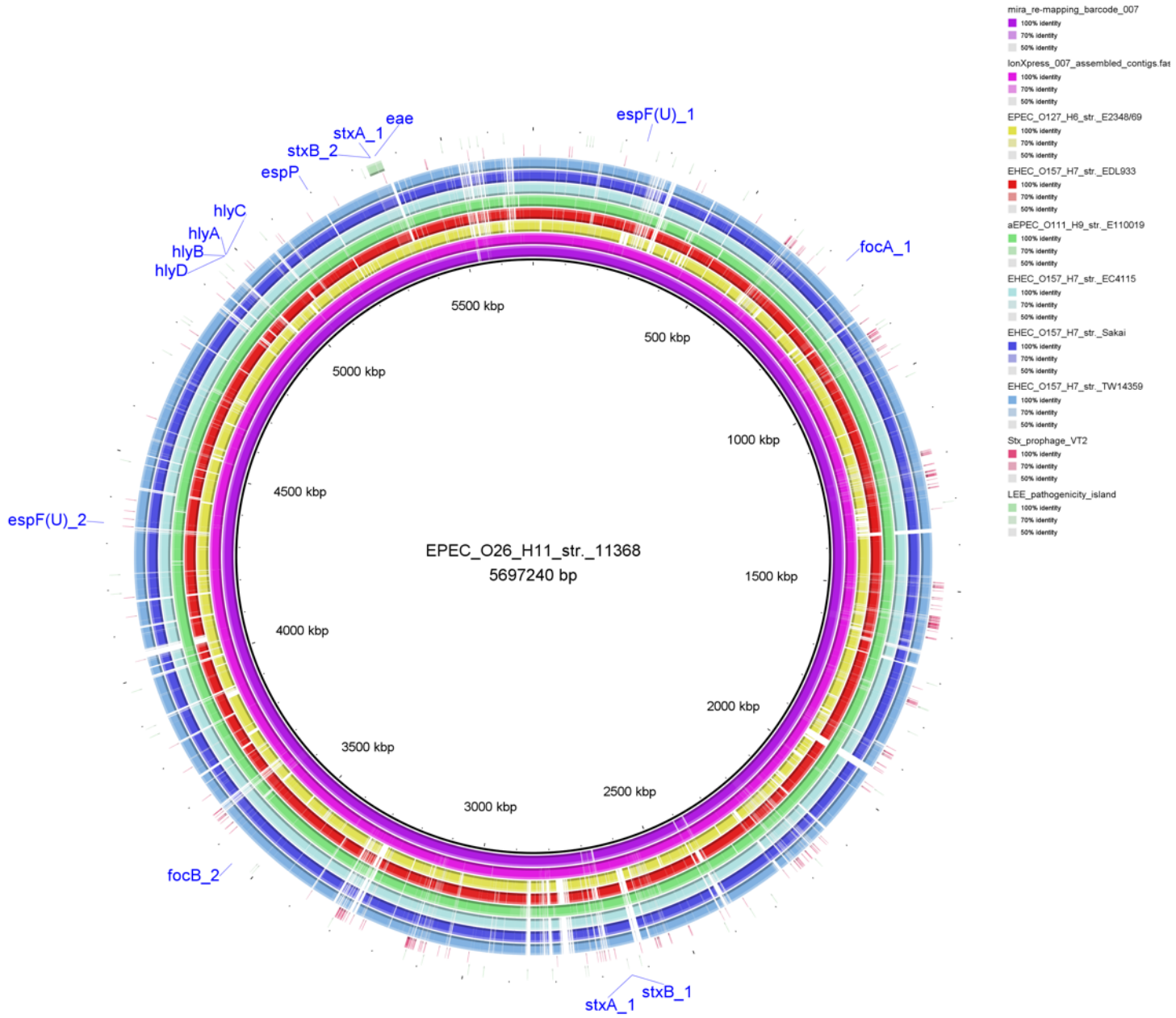
- Raw data are usually provided in FastQ, an ASCII-based format containing sequence reads at a typical length of up to 100 base pairs – Ion Torrent 150-200
- FastQ format provides an ASCII-coded quality score for every single base
- Data quality checks and filtering procedures
- Sequence alignment to the reference genome or transcriptome of the organism of origin for resequencing
- *De novo* sequencing of previously uncharacterized genomes, a reference-free *de novo* genome assembly is required e.g. COVID-19
- These are the most data-intensive and time-consuming step of the overall procedure and many mapping algorithms have been developed with a focus on both accuracy and speed

Overview of the next generation sequencing (NGS) bioinformatics workflow



R. Pereira et al. J. Clin. Med. 2020, *9*, 132

E. coli whole genome mapping for Shiga toxin and other virulence factors



CeCBiD-EOSC bioinformatics analysis of NGS data - development stages and implementation

- Installation of Apache Taverna and Galaxy workflow management systems. Galaxy Cloud will also be implemented through CloudMan and tested on the local cluster.
- Interconnecting Apache Hadoop and Spark with Galaxy Cloud, and testing applications that use techniques for massive data infrastructures (Big Data) for bioinformatics analysis of NGS data
- Installation and testing of bioinformatics applications for NGS analysis running on graphics accelerators.
- Design and testing of bioinformatics workflows for assembling de novo and reference sequences, genotyping, SNP detection, indel detection, etc.
- Validation of bioinformatics workflows for the analysis of NGS data using human genomic data.

Thank you!