

Organización, gestión y cómputo de datos genómicos

BU-ISCIII

Unidades Comunes Científico Técnicas - SGSAFI-ISCIII

2-10 Noviembre 2022
Programa Formación AESAN

Index

The Computing Revolution in Biosciences:

- The Century of Biology
- Computing in Biosciences
- The Omics Era
- Change of Paradigm
- HPC infrastructure
- Workflows
- The need of standardisation
- Nextflow
- Containers
- Data Management Plan
- LIMS

The Century of Biology

“If the 20th century was the century of physics, the 21st century will be the century of biology. While combustion, electricity and nuclear power defined scientific advance in the last century, the new biology of genome research—which will provide the complete genetic blueprint of a species, including the human species—will define the next.”

VENTER, C., & COHEN, D. (2004). The Century of Biology. *New Perspectives Quarterly*, 21(4), 73–77.
doi:10.1111/j.1540-5842.2004.00701.x

Healthcare IT News

GLOBAL EDITION ▼ TO

Obama's next move: Precision medicine and genomics venture capitalist?

By [Jessica Davis](#) | June 29, 2016 | 04:48 PM



Healthcare IT News

GLOBAL EDITION ▼ TO

Microsoft, Google invest in precision medicine startup DNAnexus

By [Bernie Monegain](#) | January 02, 2018 | 12:25 PM



Computing in Biosciences I

Research used to be focussed in a small number of samples and researchers analysed them with the whatever means they had and/or felt more comfortable with:

- Windows based PC using programs with visual interface
- Macs and Linux based workstations
- Remote web servers
- Web-based platforms (i.e. Galaxy) and remote HPC
- HPC local environments

Computing in Biosciences II

- Windows based PC using programs with visual interface

Pros	Cons
Data remains private	No backups or data management schemes
Software easy to install	Software version not easy to control, binaries are black boxes
Graphic interface	No control over hidden parameters
	Analysis are irreproducible

Computing in Biosciences III

- Macs and Linux based workstations

Pros	Cons
Data remains private	No backups or data management schemes
Control over software installed versions, open source programs	Software may not be easy to install, library and dependencies problems
All parameters are available for the command	Command line interface
	Analysis are irreproducible

Computing in Biosciences IV

- Remote web servers

Pros	Cons
No need to storage intermediate files	Your data is in someone else's computer No backups or data management schemes
No need to install software	Software version not easy to control, black boxes
Graphic interface	No control over hidden parameters
	Quotas Analysis are irreproducible

Computing in Biosciences V

- Web-based platforms (i.e. Galaxy) and remote HPC

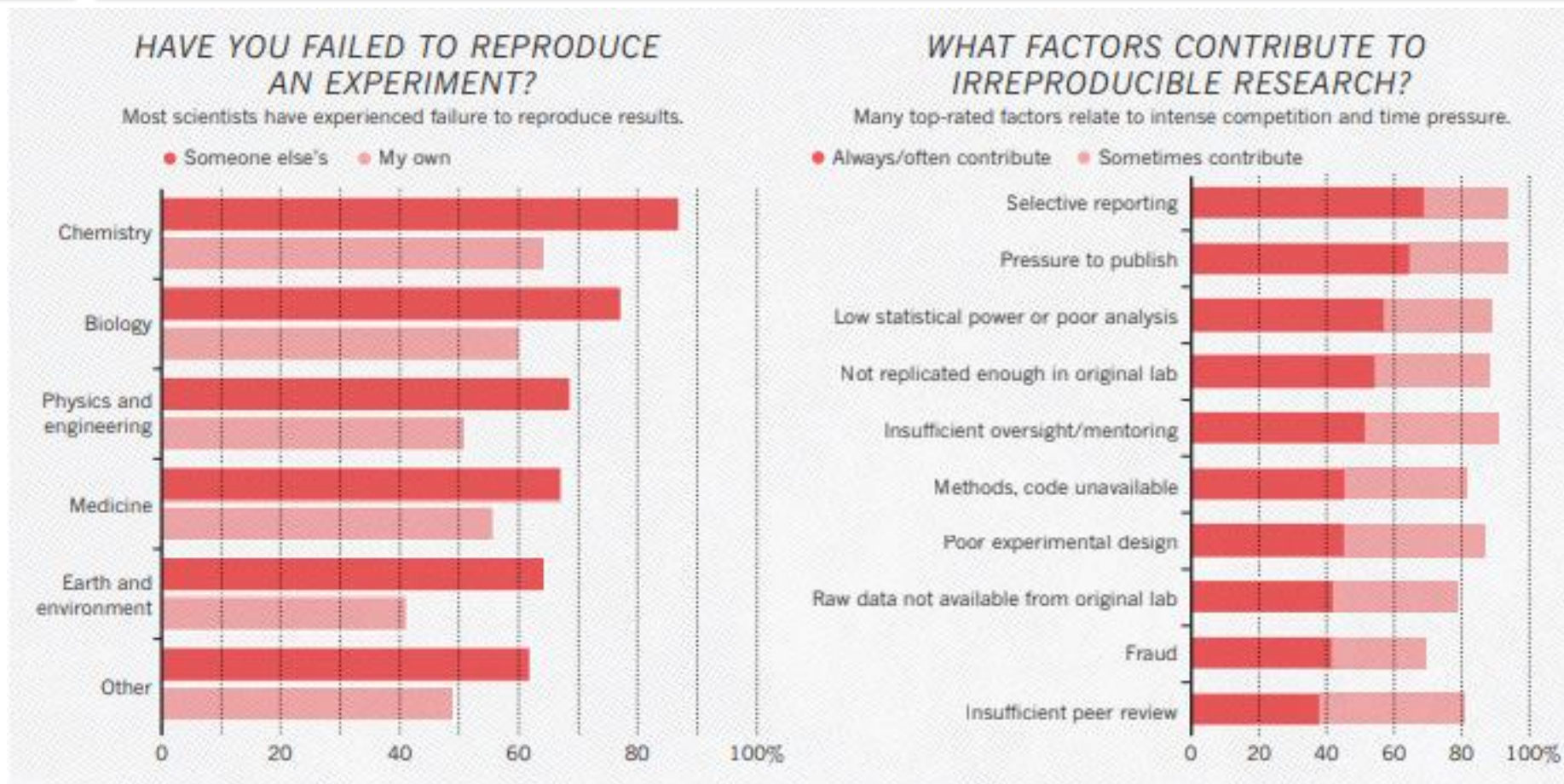
Pros	Cons
No need to storage intermediate files	Your data is in someone else's computer No backups or data management schemes
No need to install software Partial control over installed software	No control over installed software, versions and future availability
Graphic interface	No control over hidden parameters
Analysis are partially reproducible	Quotas

Computing in Biosciences VI

- HPC local environments

Pros	Cons
Data remains private Backups and data management schemes	Quotas
No need to install software Partial control over installed software	No control over installed software, versions and future availability
All parameters are available for the command	Command line interface
Possibility of suggesting new software installations	Analysis may be irreproducible

Is there a reproducibility crisis?



Source: Baker, M. "Reproducibility Crisis (Nature)," 3–5. doi:10.1038/533452A.

The Omics Era I

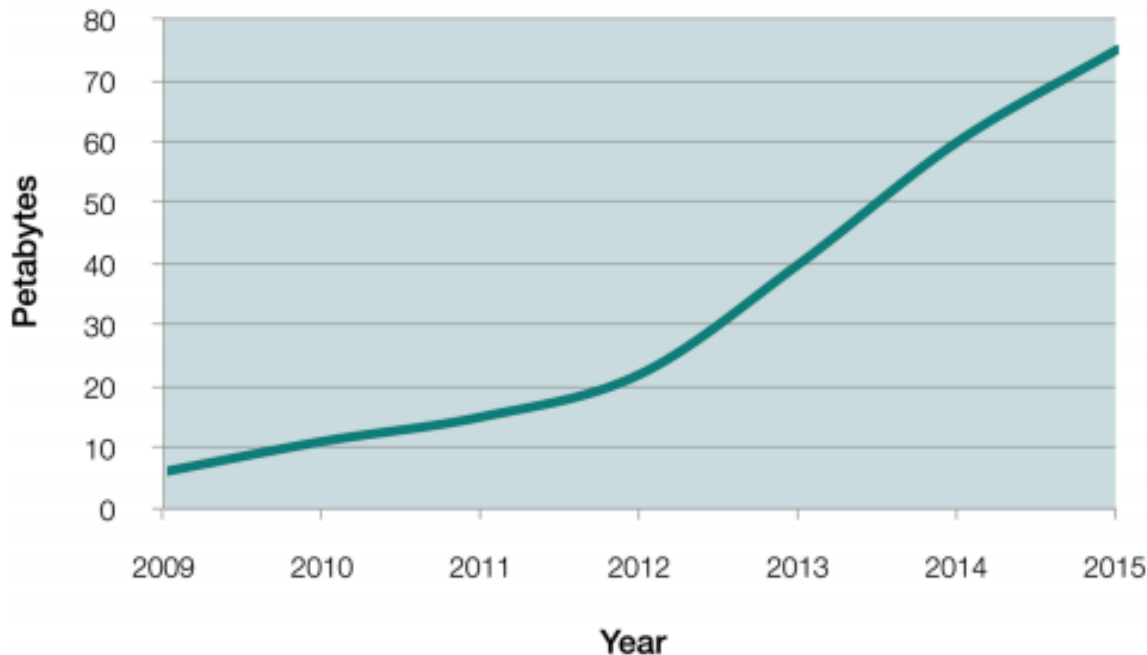
	Coverage	No. of Reads	Read Length	BAM File Size	Strand NGS Size
Whole Genome	37.7x	975,000,000	115	82 GB	104 GB
Whole Genome	38.4x	3,200,000,000	36	138 GB	193 GB
Exome	40x	110,000,000	75	5.7 GB	7.1 GB

Whole Genome Samples	Exome Samples	Space	Space including Backup
0	200	1.6 TB	3.2 TB
0	1000	8.0 TB	16 TB
100	0	15 TB	30 TB
1000	0	150 TB	300 TB
100	1000	23 TB	46 TB

Source: <https://www.strand-ngs.com/support/ngs-data-storage-requirements>

The Omics Era II

Total disk storage at EMBL-EBI



Installed (2008–2015) storage at EMBL-EBI. These figures include all installed storage, counting multiple backups for all data resources as well as unused storage to handle submissions in the immediate future

Source: Cook, Charles E et al. "The European Bioinformatics Institute in 2016: Data growth and integration" *Nucleic acids research* vol. 44,D1 (2015): D20-6.

Change of Paradigm I

1 sample

Research only: NGS was still a new thing, no applications 10 years ago

Reproducibility is not needed: Why would anyone reanalyse this?

Storage is not an issue: files of 1 sample fits everywhere in my HDD, maybe I will copy it in a CD-ROM

Computing is simple: no need to worry about resources or optimisation

multiple samples

Many applications: research, clinical, industrial, forensic, military, ...

Reproducibility, scalability , portability and standardisation are required

Storage is challenging: storage, indexation and backup required, privacy and legal standards

Computing requires optimisation and lots of resources

HPC infrastructure I

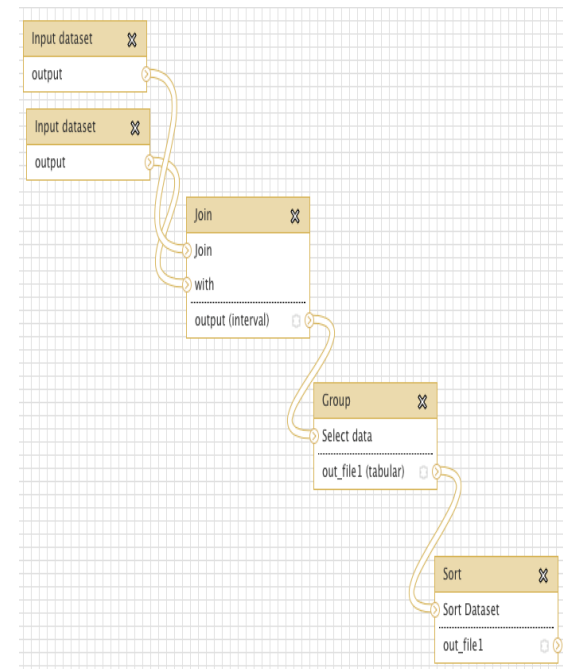
Machine	OS	Software	CPU	RAM	Storage
Workstation (x5)	Centos 6.9	/opt(*)	4 cores	32 Gb	4 TB
Bioinfo01 (1 node)		/opt(*)	16 cores	120 Gb	500 Gb
HPC (16 nodes)		/opt(*)	320 cores	8 TB	500 Gb

2 shared data storage disk boxes: 70TB + 250 TB

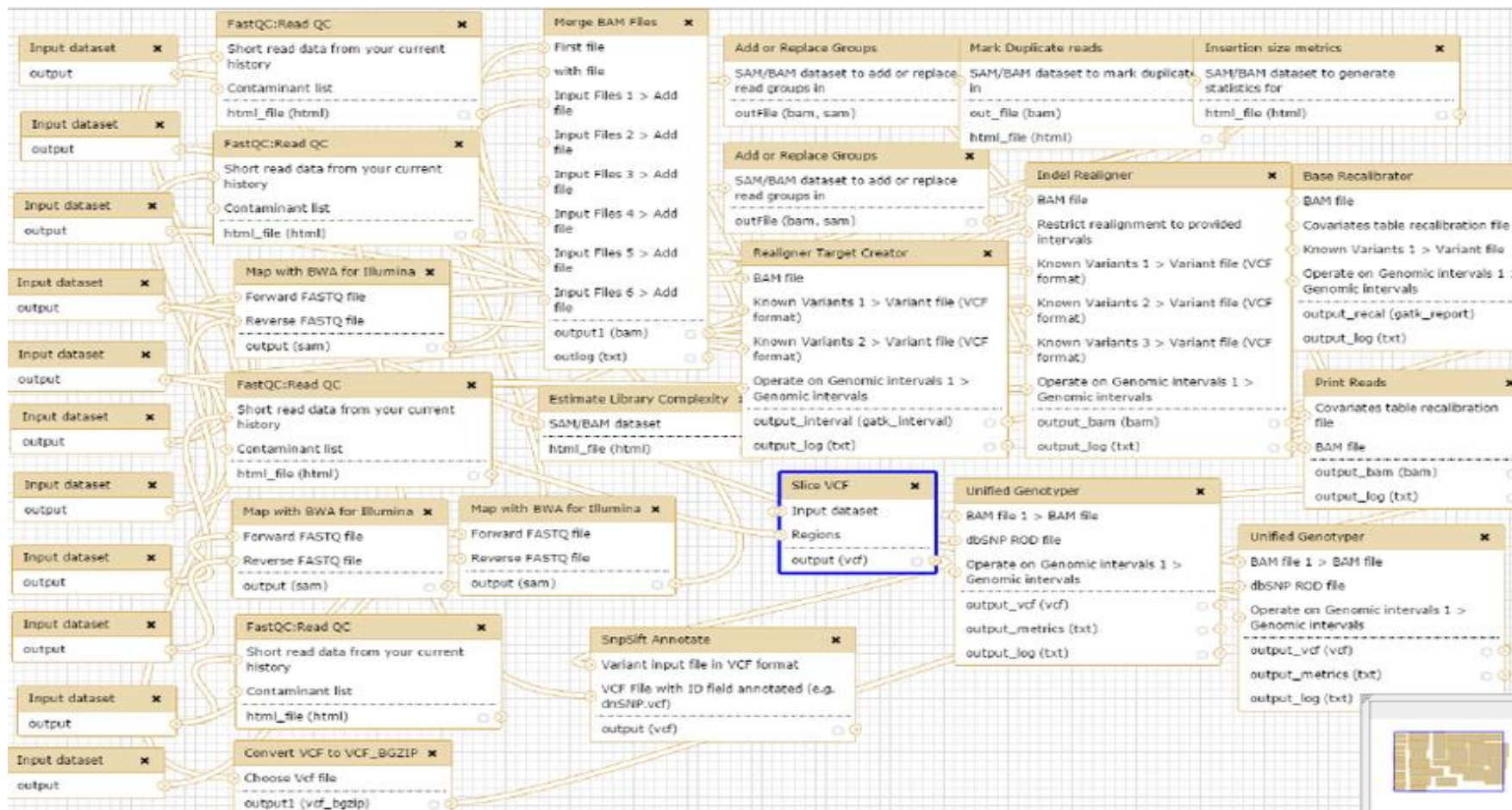
VMs, ISCIII's Windows personal terminals, personal laptops mobile platforms, cloud computing platforms, cloud storage, remote services, ...

Workflows I

- Bioinformatic analyses invariably involve shepherding files through a series of transformations, called a **pipeline** or a **workflow**.
- These transformations are done by executable **command line software** written for Unix-compatible operating systems.
- They need to be **reproducible**, **easy to maintain**, **portable** and **scalable**.



Workflows II



The need of standardisation I

Sequencing techniques are starting to be used in clinical diagnosis, and therefore workflows have to assure:

- **Reproducibility**

Results always have to be reproducible

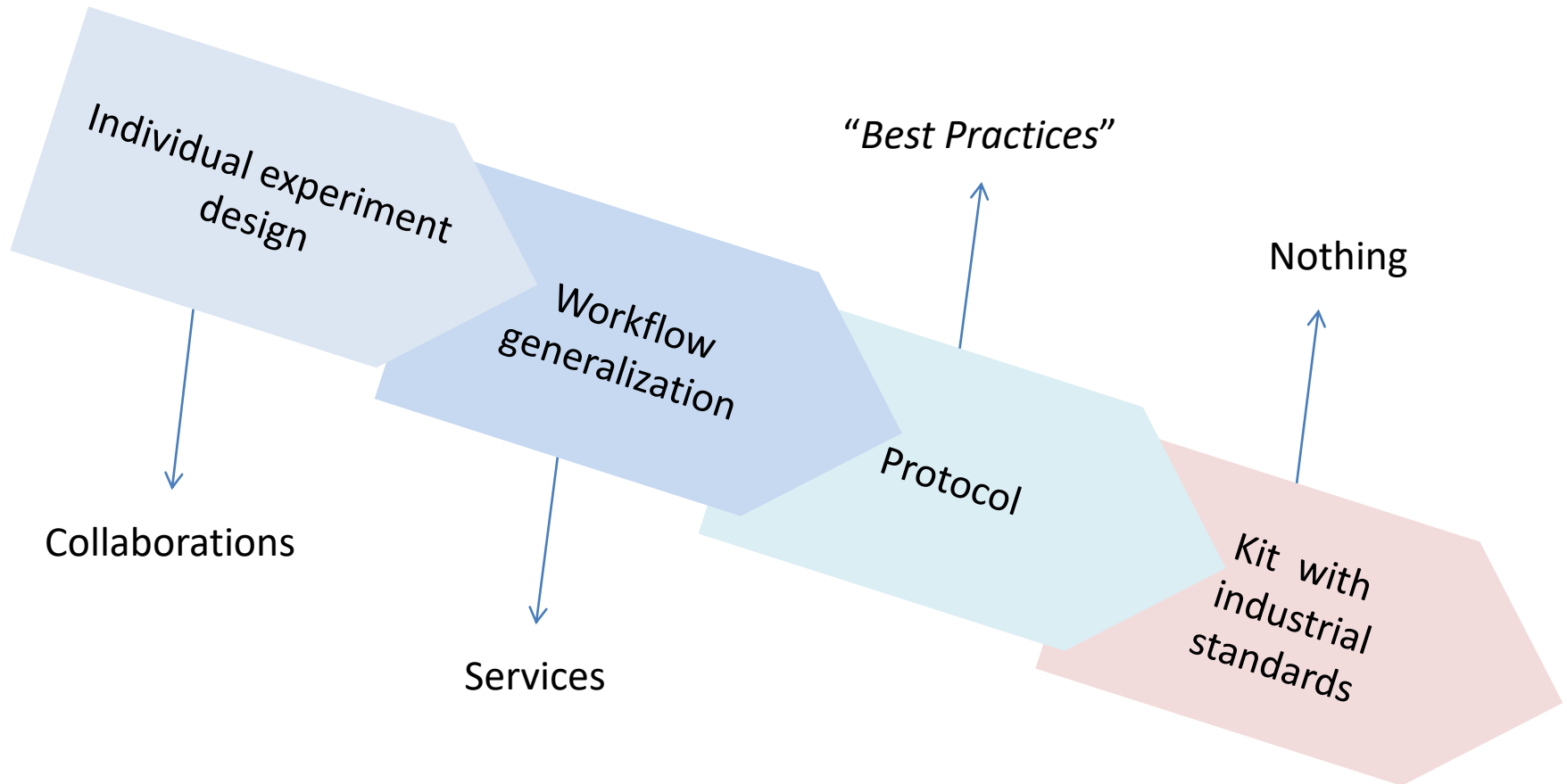
- **Portability**

The analysis workflow must be executable in different platforms

- **Scalability**

The analysis workflow must be able to work with different numbers of samples

The need of standardisation II



Nextflow

Fast prototyping

Nextflow allows you to write a computational pipeline by making it simpler to put together many different tasks.

You may reuse your existing scripts and tools and you don't need to learn a new language or API to start using it.

Portable

Nextflow provides an abstraction layer between your pipeline's logic and the execution layer, so that it can be executed on multiple platforms without it changing.

It provides out of the box executors for SGE, LSF, SLURM, PBS and HTCondor batch schedulers and for [Kubernetes](#) and [Amazon AWS](#) cloud platforms.

Continuous checkpoints

All the intermediate results produced during the pipeline execution are automatically tracked.

This allows you to resume its execution, from the last successfully executed step, no matter what the reason was for it stopping.

Reproducibility

Nextflow supports [Docker](#) and [Singularity](#) containers technology.

This, along with the integration of the [GitHub](#) code sharing platform, allows you to write self-contained pipelines, manage versions and to rapidly reproduce any former configuration.

Unified parallelism

Nextflow is based on the *dataflow* programming model which greatly simplifies writing complex distributed pipelines.

Parallelisation is implicitly defined by the processes input and output declarations. The resulting applications are inherently parallel and can scale-up or scale-out, transparently, without having to adapt to a specific platform architecture.

Stream oriented

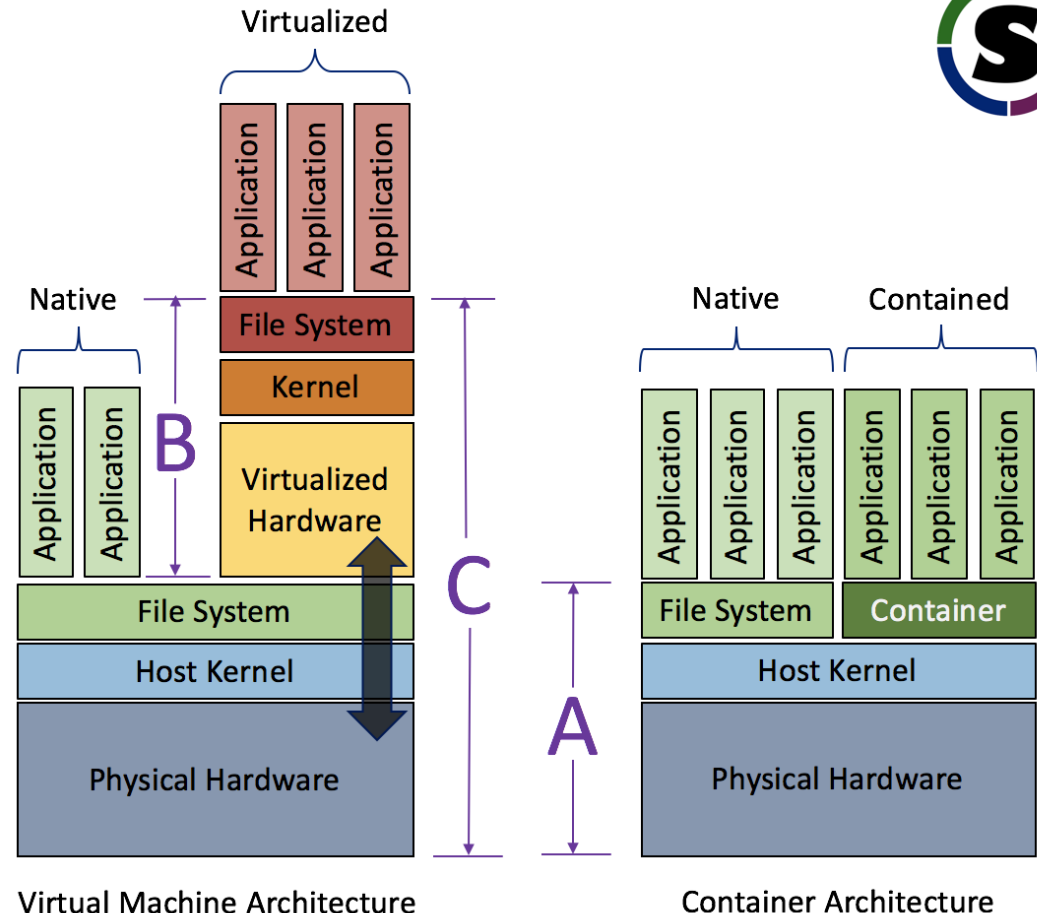
Nextflow extends the Unix pipes model with a fluent DSL, allowing you to handle complex stream interactions easily.

It promotes a programming approach, based on functional composition, that results in resilient and easily reproducible pipelines.

Containers



- Applications running within a container will always be “closer” to the physical hardware
 - Notice how close to native a container behaves
- Applications running through a virtual machine will always have multiple levels of indirection
- The container’s proximity to the physical hardware equates to less overhead, higher performance and lower latency



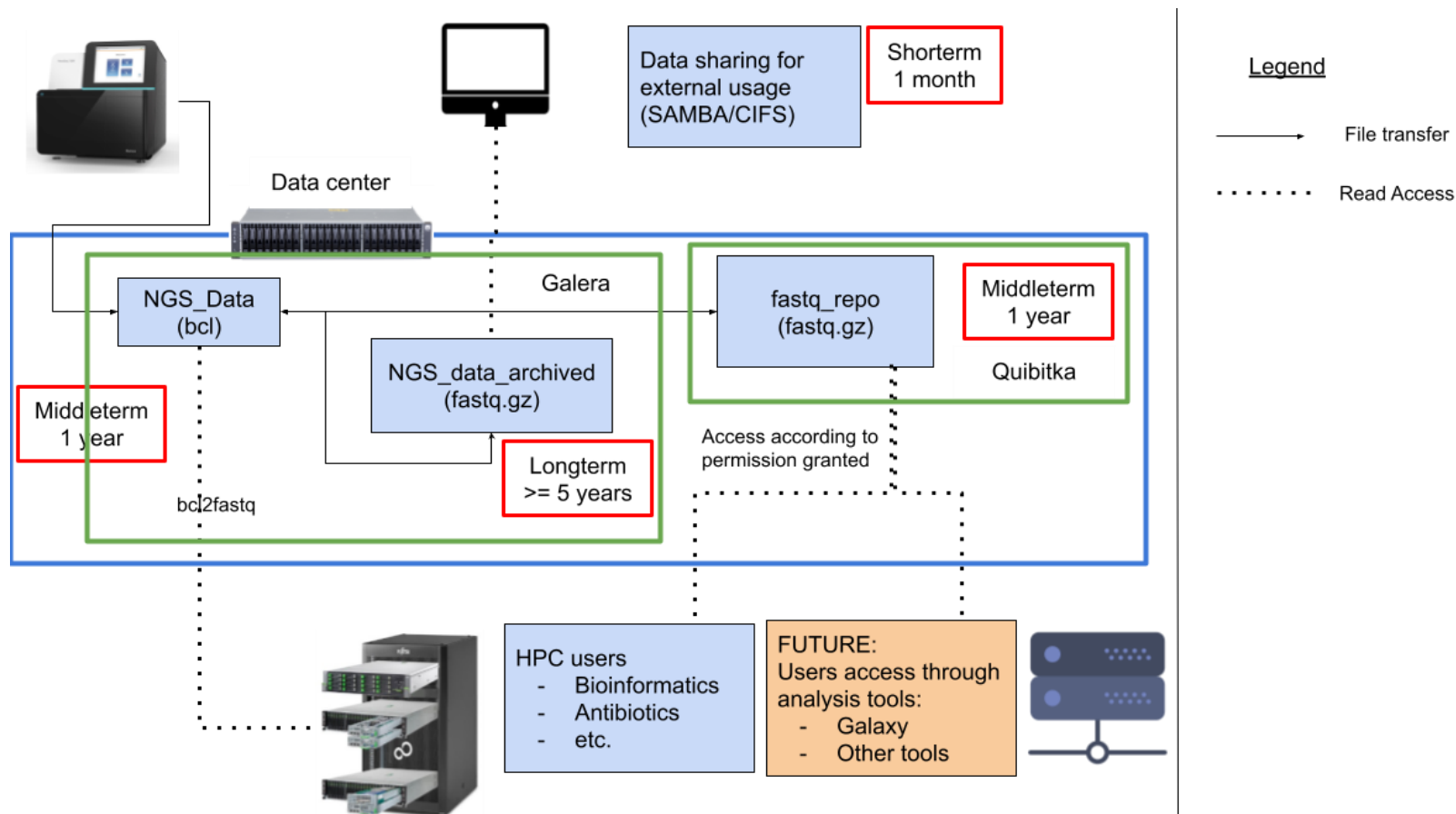
Data Management Plan

Protocolo de organización y gestión de almacenamiento de datos genómicos en el ISCIII

Contenidos

Introducción	2
Descripción del problema	2
Solución y organización de recursos	3
Descripción de los scripts	6
Ejecución de los scripts	7
Ubicación de los scripts	7
Logs.....	7
Descripción de los recursos.....	8
Servicios a disposición del usuario	10
Protocolo de solicitud.....	10
Mejoras futuras	11
Bibliografía.....	12
Glosario	12
Control de versiones	13

Data Management Plan



Data Management Plan

- SFTP



FileZilla

How SFTP works



SOURCE: PAUL KIRWAN, 2022.
ICONS: COMPUTER: SHUTTERSTOCK/GETTY IMAGES, INTERNET AND WEBSITE SERVER: GENESTRQ/GETTY IMAGES, ALL OTHERS: KADIRKABA/GETTY IMAGES

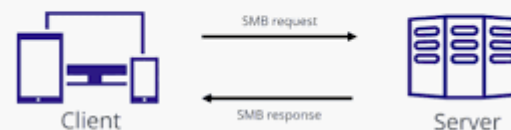
©2022 TECHTARGET. ALL RIGHTS RESERVED. TechTarget

Data Management Plan

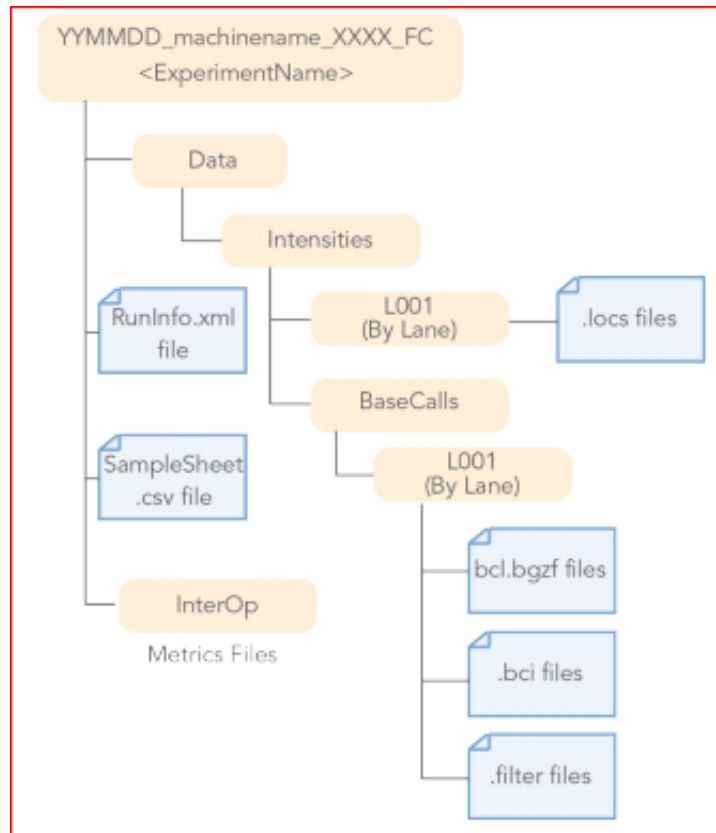
• SAMBA

	Nombre	Fecha de modificacion	lipo
> OneDrive - Personal			
✓ Este equipo			
> BI (galera.isciii.es (Isilon Server))			
> bioinfo_doc			
> Descargas			
> Documentos			
> Escritorio			
> Imágenes			
> Música			
> NovaSeq_GEN_008_20210512_BM			
> Objetos 3D			
> Vídeos			
> Windows (C:)			
> LENOVO (D:)			
> Google Drive (G:)			
> CNM_OT_Robots (\\cibeles.isciii.es)			
> Red			
	collaborations	20/01/2022 16:47	Carpeta de archiv
	congresses	25/09/2022 21:39	Carpeta de archiv
	documentation	04/01/2022 13:46	Carpeta de archiv
	internships	05/08/2020 12:40	Carpeta de archiv
	logos	04/01/2022 13:38	Carpeta de archivos
	lost+found	11/04/2019 16:41	Carpeta de archivos
	meetings	05/01/2022 17:29	Carpeta de archivos
	miscellaneous	04/01/2022 13:45	Carpeta de archivos
	nanopore	18/05/2021 12:57	Carpeta de archivos
	papers_consulta	16/10/2020 13:23	Carpeta de archivos
	prensa	30/07/2020 16:00	Carpeta de archivos
	projects	15/03/2022 17:16	Carpeta de archivos
	publications	04/01/2022 13:45	Carpeta de archivos
	research	08/06/2022 14:37	Carpeta de archivos
	seminarios	04/01/2022 13:33	Carpeta de archivos
	services	11/10/2022 13:30	Carpeta de archivos
	temporal	25/08/2022 19:17	Carpeta de archivos
	training	17/03/2022 12:20	Carpeta de archivos
	unit_organization	04/02/2022 11:07	Carpeta de archivos

Server Message Block (SMB)



Data Management Plan



Data Management Plan



Data Management Plan

DESCARGA DE DATOS DE SECUENCIACIÓN MASIVA

Los datos de secuenciación masiva secuenciados en el ISCIII estarán disponibles para su descarga desmultiplexados por muestra en formato fastq una vez completada la carrera de secuenciación durante el periodo de un mes. Una vez completada la carrera de secuenciación el usuario recibe un email (soporte.hpc@isciit.es) con instrucciones de cómo descargar los datos. Estos datos estarán disponibles para su descarga durante 1 mes.

ANÁLISIS EN SISTEMAS DE ALTA COMPUTACIÓN

Disposición de los datos de secuenciación masiva secuenciados en el ISCIII en recurso de datos activos en el entorno de alta computación del ISCIII durante 1 año para su uso y análisis. Para hacer uso de este entorno de computación se puede solicitar acceso mediante la aplicación GLPI <https://sau.isciit.es> (se requiere conocimiento de Sistema operativo Linux y entorno de alta computación) o se puede solicitar Servicio a la Unidad de Bioinformática desde (<https://iskylims.isciit.es>), más información póngase en contacto con bioinformatica@isciit.es).

ARCHIVADO

Todos los datos de secuenciación masiva secuenciados en el ISCIII son archivados durante un periodo de **10 años** con mecanismos de seguridad adecuados en las cabinas de la UTIC del ISCIII. Estos archivos pueden ser recuperados en cualquier momento por el investigador mediante solicitud en GLPI <https://sau.isciit.es>. Una vez pasados esos **10 años** los datos se borrarán, realizándose un aviso al investigador 1 mes antes de su borrado para que pueda disponer de ellos si así lo desea.

Laboratory Information Management System: iSkyLIMS

- **LIMS. (Laboratory Information Management System)**
 - Conjunto de herramientas basadas en sistemas informáticos que permite la adquisición y gestión de toda la información generada en el laboratorio.
- **¿Por qué se necesita un LIMS?**
 - Poder manejar la enorme cantidad de información que genera un laboratorio.
 - Identificación de muestras
 - Tiempos y procedimientos de recogida, procesamiento, transporte, eliminación, etc...
 - Personal encargado de cada procedimiento
 - Gestión de reactivos (lotes, fechas de caducidad)
 - ...
 - Ayuda a la Gestión de la Calidad (Normas de certificación y acreditación ISO – ENAC. P.e ISO 17025, ISO 15189)
 - En resumen: intentar por todos los medios REDUCIR/ELIMINAR errores.
- **Problemática:**
 - Distintos laboratorios tienen distintos requerimientos dependiendo de sus procedimientos y de su infraestructura.

iSkyLIMS

- **¿Por qué un LIMS orientado al proceso de secuenciación masiva?:**
 - Los procedimientos de secuenciación masiva generan una gran cantidad de datos ya que permite la secuenciación de muchas muestras al mismo tiempo.
 - Ayuda en la gestión de muestras/carreras: reducir errores en la preparación y configuración de la carrera.
 - Control de la calidad de la secuenciación
 - Mejorar el control sobre el proceso y la capacidad de predicción
 - Necesidad de una respuesta más rápida en la relación muestra/tiempo de los resultados
- **¿Por qué desarrollar nuestro propio LIMS?:**
 - Actualmente no existen otros LIMS que cubran todas las necesidades de la Unidad de Genómica y de Bioinformática y adecuado a la infraestructura de la que disponemos.

iSkyLIMS



iskylims.isciii.es

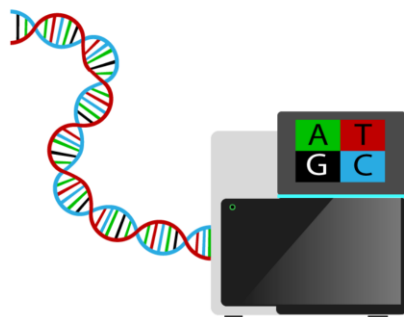
HOME ABOUT US TUTORIALS FAQs REGISTER CONTACT

bioinfoadm

Logout

My account

Módulo para la gestión de datos de secuenciación masiva



Módulo para solicitar servicio a la Unidad de Bioinformática

Drylab



IMPORTANTE:

Para acceder se necesita usuario y contraseña

Para solicitar usuario escribir un correo a:

bioinformatica@isciii.es

Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB

iSkyLIMS

HOME SERVICES REQUEST COUNSELING REQUEST INFRASTRUCTURE REQUEST


bmartinezd Logout My account

BU-ISCIII

iSkyLIMS: DryLab

Welcome

This section will allow you to check BU-ISCIII service activity. Available processes are: requesting new services, collaborations, counseling and infrastructure. You will be able to check the status of your ongoing services.



Services ongoing and queued

Service Number	Status	Expected delivery date
SRVIER080	in_progress	15 September, 2018
SRVCNM090	in_progress	10 May, 2019
SRVCNM116	archived	09 August, 2019
SRVIER177	in_progress	25 September, 2020
SRVIER207	in_progress	30 June, 2020

Timeline of services

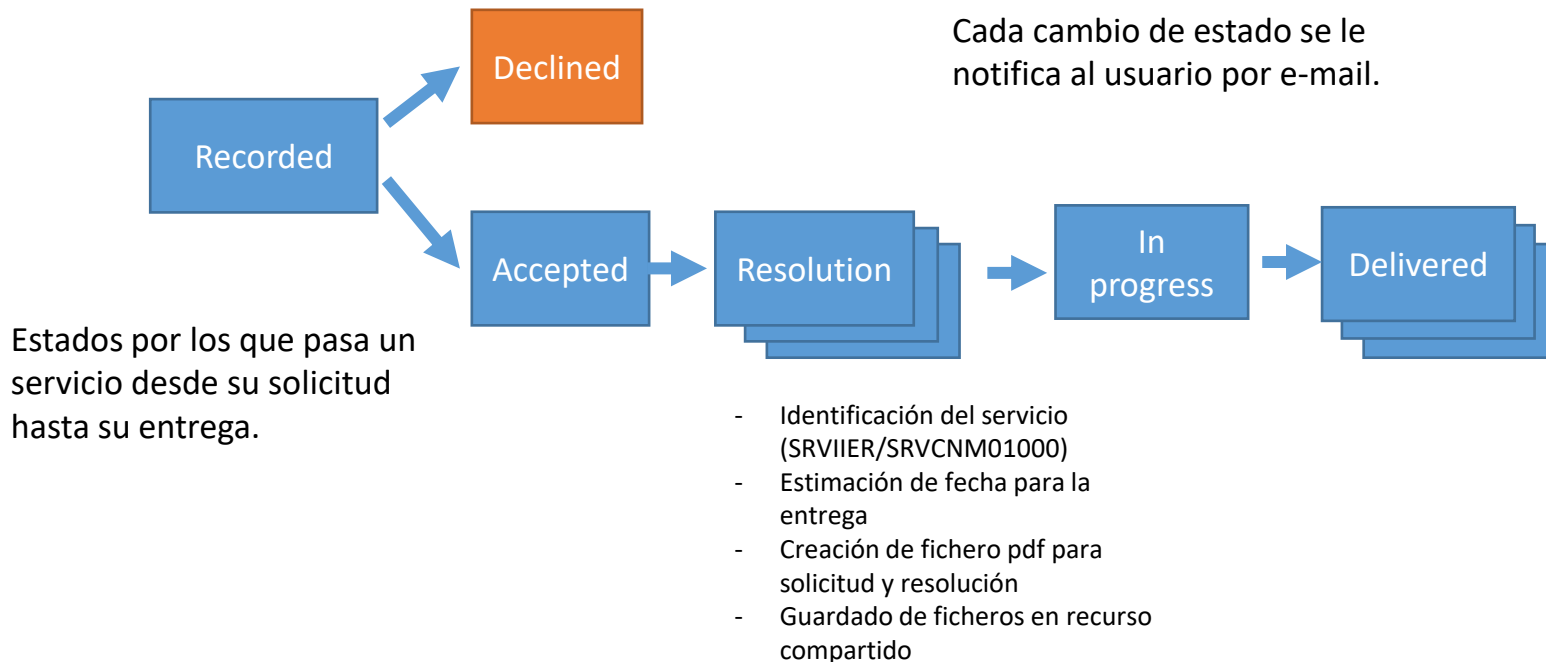
Under construction. Kind of diagram with services dates.

Solicitud de servicios a la Unidad de Bioinformática incluidos en su cartera

Seguimiento del servicio solicitado

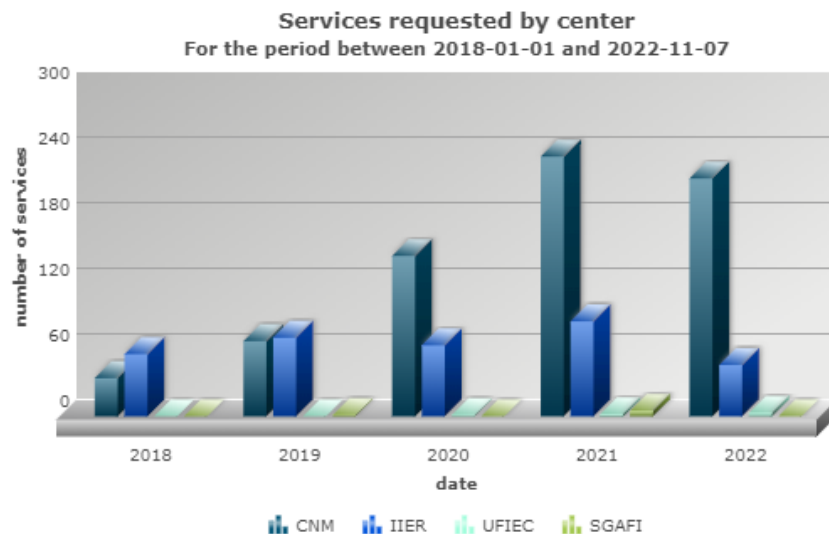
Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB



Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB



- iSkyLIMS nos permite gestionar, cuantificar y obtener estadísticas de los servicios realizados.

Además nos permite saber los tiempos medios de resolución de los servicios.

Trabajamos con dos tiempos:

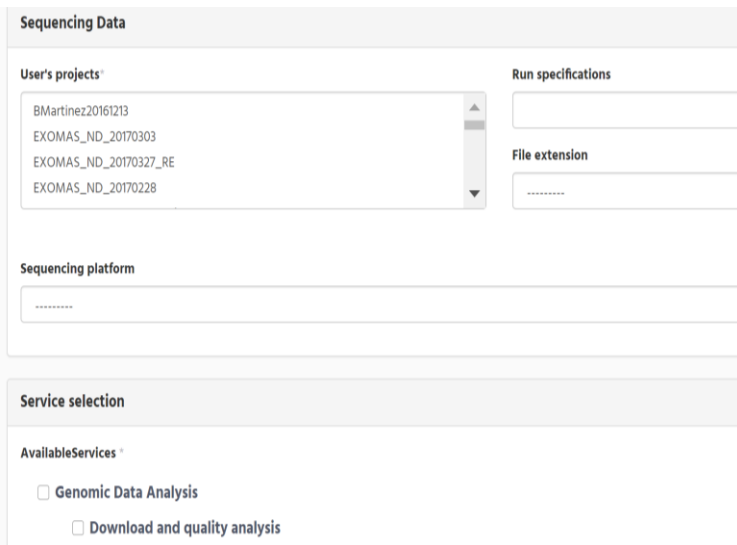
- **Tiempo total:** incluye tiempo en cola que pueda estar el servicio y que depende del personal disponible en la Unidad y la carga de trabajo en el momento de la solicitud. Además de la priorización de servicios relacionados con labores de diagnóstico y/o salud pública.
- **Tiempo de resolución:** es el tiempo de ejecución del servicio real, desde que el personal inicia el servicio (que coincide cuando el investigador recibe el mail de que el servicio está en progreso) hasta la entrega de resultados.

Ejemplos:

- Servicio de ensamblado (~100 muestras):
 - Tiempo total: 3-5 días de media
 - Tiempo de resolución: 1-2 días de media.
- Servicio de RNASeq (~20 muestras):
 - Tiempo total: 15 días de media
 - Tiempo de resolución: 4-5 días de media
- Servicio Análisis de variantes en exoma/RNAseq: (~10 muestras)
 - Tiempo total: 15 días de media.
 - Tiempo de resolución: 5-7 días de media.

Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB



The screenshot shows the 'Sequencing Data' section of the DRYLAB interface. It includes a 'User's projects' list with four entries: BMartinez20161213, EXOMAS_ND_20170303, EXOMAS_ND_20170327_RE, and EXOMAS_ND_20170228. To the right, there are input fields for 'Run specifications' and 'File extension'. Below this is a 'Sequencing platform' field. The 'Service selection' section at the bottom shows two available services: 'Genomic Data Analysis' and 'Download and quality analysis', both with unchecked checkboxes.

- **Services request** -> servicios relacionados con análisis genómicos. Se dan dos opciones:
 - **Internal sequencing:** Secuenciaciones realizadas en la Unidad de Genómica. Seleccionar el proyecto de secuenciación para el que se solicita el servicio.
 - **External sequencing:** Secuenciaciones realizadas de manera externa.
- **Counseling request** -> servicios relacionados con consultoría, estancias, formación, etc.
- **Infrastructure request** -> servicios relacionados con solicitud de máquinas virtuales, scripts, instalación de software, robots opentrons, etc.
- **IMPORTANTE:**
 - se puede adjuntar un fichero a la solicitud (**con nombre corto y conciso, sin tildes**).
 - La descripción del servicio debe no superar las 5-6 líneas.
 - Si se quiere adjuntar más información se puede complementar enviando un mail a bioinformatica@isci.es

Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB

- Services request

- ☐ Genomic Data Analysis
 - ☐ Download and quality analysis
 - ☐ Data download
 - ☐ Sequence quality analysis
 - ☐ Sequence pre-processing (quality filtering)
 - ☐ Next Generation Sequencing data analysis
 - ☐ DNAseq: Exome sequencing (WES) / Genome sequencing (WGS) / Target sequencing
 - ☐ Trio/family variant calling pipeline
 - ☐ Variant calling and annotation pipeline
 - ☐ Microbial: Whole genome outbreak analysis pipeline
 - ☐ Microbial: wgMLST
 - ☐ Microbial: MLST + virulence + AMR + plasmid analysis
 - ☐ Microbial: Assembly + automatic annotation
 - ☐ Microbial: plasmidID pipeline - strain plasmid characterization
 - ☐ RNAseq: Transcriptome sequencing
 - ☐ miRNA-Seq pipeline
 - ☐ mRNA-Seq pipeline
 - ☐ Amplicon sequencing (Deep sequencing)
 - ☐ Low frequency variant detection
 - ☐ Viral: assembly and minor variants detection
 - ☐ Metagenomics
 - ☐ 16S taxonomic profiling
 - ☐ Shotgun metagenomics profiling
 - ☐ Shotgun metagenomics - Virus genome reconstruction

Análisis de exomas
en familias



Ensamblado de genoma
bacteriano



Genoma viral
consenso



Módulo para gestión de servicios de la Unidad de Bioinformática

DRYLAB

- Counseling request

AvailableServices *

- ☐ **Bioinformatics consulting and training**
 - ☐ Bioinformatics analysis consulting
 - ☐ In-house and outer course organization
 - ☐ Student training in collaboration: Master thesis, research visit,...

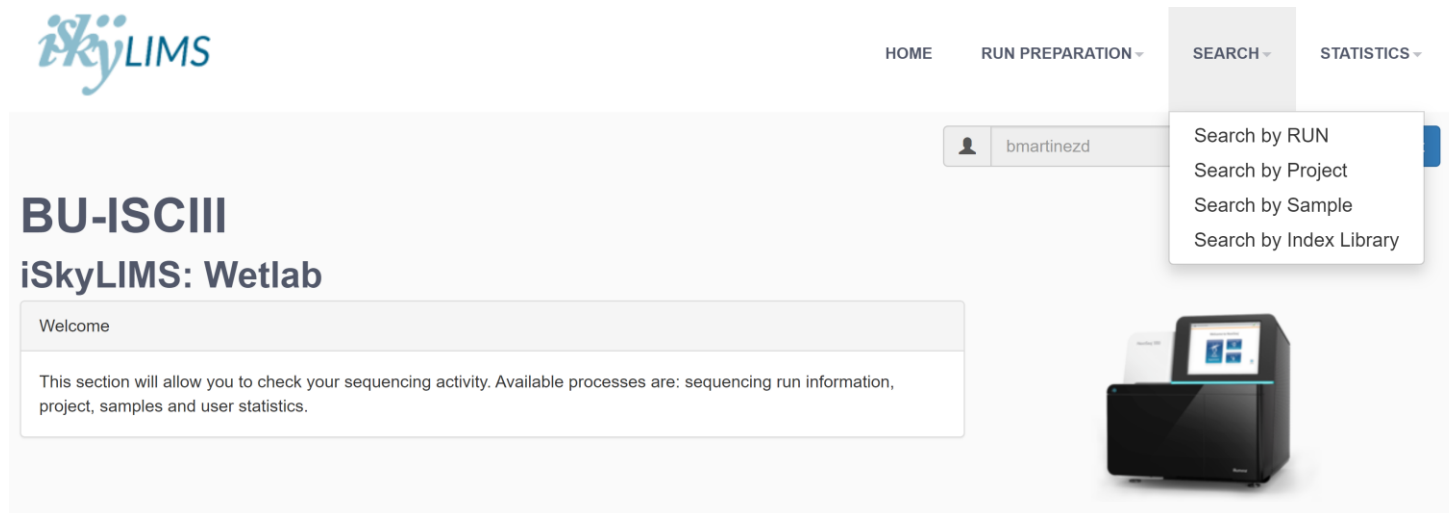
- Infrastructure request

- ☐ **User support**
 - ☐ Installation and support of bioinformatic software on Linux OS
 - ☐ Installation and access to Virtual machines in the Unit server containing bioinformatic software
 - ☐ Code snippets development
 - ☐ OT-2 robots

Módulo para la gestión de datos de secuenciación masiva

WETLAB

- Usado por la Unidad de Genómica para gestionar y poner en marcha los secuenciadores.
- Almacenamiento de datos de calidad y parámetros de secuenciación de todas las carreras que se han secuenciado en el centro.



Módulo para la gestión de datos de secuenciación masiva

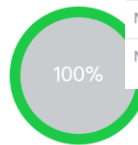
WETLAB

- Búsqueda por Carrera, por proyecto, por muestra... (Por ejemplo: localizar en qué carrera se ha secuenciado una muestra)
- Visualización de parámetros y datos de calidad.

Information for the run : NextSeq_CNM_068

State of the Run is	Completed
Run was requested by	Centro Nacional de Microbiología
Run was recorded on date	03:40PM on December 10, 2017
Run date	October 05, 2017
Run Completion Date	03:50PM on December 10, 2017
Bcl2Fastq finish Date	09:23AM on October 06, 2017
Run Finish Date	06:50AM on October 06, 2017
Disk space used for Images(in MB)	39
Disk space used for Fasta Files(in MB)	72,574
Disk space used for other Files(in MB)	336
Sample Sheet File	Sample Sheet File Download

State run is : Complete



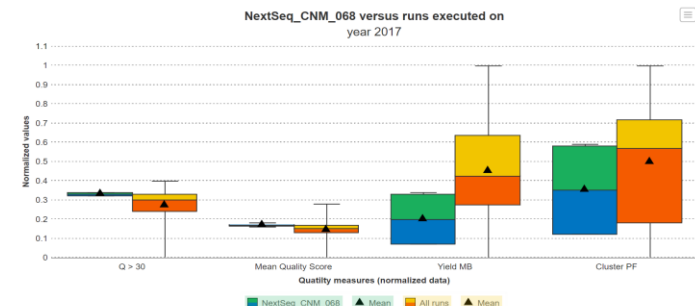
Run Quality vs all Runs

Run Quality Graphic

Sample	Barcode	PF Clusters	Percent of Project	Yield (Mbases)	% >= Q30 bases	Mean Quality Score
ND0231	TAAGGCGA+ATAGAGAG	64,529,500	20.05	9,983	90.802	33.719
ND0232	CGTACTAG+ATAGAGAG	65,723,706	20.42	10,195	91.173	33.799
ND0233	AGGCAGAA+ATAGAGAG	63,677,341	19.79	9,875	90.869	33.734
ND0241	TCCTGAGC+ATAGAGAG	63,145,540	19.62	9,778	91.117	33.787
ND0242	GGACTCCT+ATAGAGAG	64,769,275	20.12	10,070	91.279	33.821

Export Table To Excel [Download](#)

Graphic showing the NextSeq_CNM_068 versus all runs on the same year

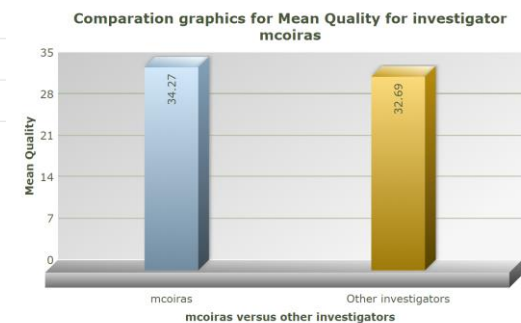
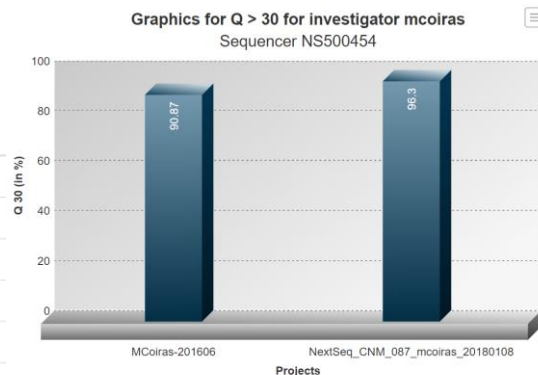


Módulo para la gestión de datos de secuenciación masiva

WETLAB

- Visión conjunta de todas o parte de las carreras realizadas. Estadísticas.

Project name	Date	Libraty Kit	Samples	Cluster PF	Yield Mb	% Q> 30	Mean
NextSeq_CNM_072_20171031BMartinez	No Date	TruSeq Rapid Exome	6	411,171,355	63,753	88.72	33.29
NextSeq_CNM_148_201900206BMartinez	No Date	Nextera DNA Exome Enrichment	5	353,823,405	54,365	95.20	34.65
NextSeq_CNM_151_20190226_Bmartinez	No Date	Nextera DNA Exome Enrichment	7	508,624,531	81,245	85.84	32.63
NextSeq_CNM_154_20190307_Bmartinez	No Date	ScriptSeqAB	12	164,546,019	50,908	73.85	30.56
NextSeq_CNM_173_Bmartinez20190712	No Date	Nextera DNA CD Indexes (96 Indexes plated)	9	45	0	77.87	30.93
NextSeq_CNM_197_20191119_BMartinez	No Date	IDT-ILMN Nextera DNA UD Indexes (96 Indexes) Set A	8	521,557,322	78,172	96.35	34.88
NextSeq_CNM_055_BMartinezd	No Date	Nextera XT	6	409,339,105	66,985	91.34	33.85
NextSeq_GEN_215_20200221_BMartinez	No Date	Nextera Flex for Enrichment	3	157,793,405	23,423	94.09	34.39
NextSeq_GEN_234_20200724_BMartinez	No Date	IDT-ILMN Nextera DNA UD Index Set A for Nextera FI	3	174,151,358	26,052	91.96	33.95



Thanks for your attention!

And this is only the tip of the iceberg...
Check this if you wanna know what's really going under the hood:



<https://github.com/BU-ISCI III>