Viralrecon

Sarai Varona Fernández

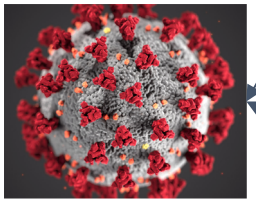
Unidad de Bioinformática
Unidades Centrales Científico Técnicas – SGAFI-ISCIII

22-26 Noviembre 2021, 1ª Edición
Programa Formación Continua, ISCIII

Outline

1. Background
2. Sequencing approaches
3. State of the art
4. Standarization
5. Nextflow
6. Nf-core
7. Viralrecon
    1. Pipeline
    2. Results

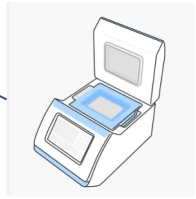
1. Background



December 2019

SARS-Cov-2

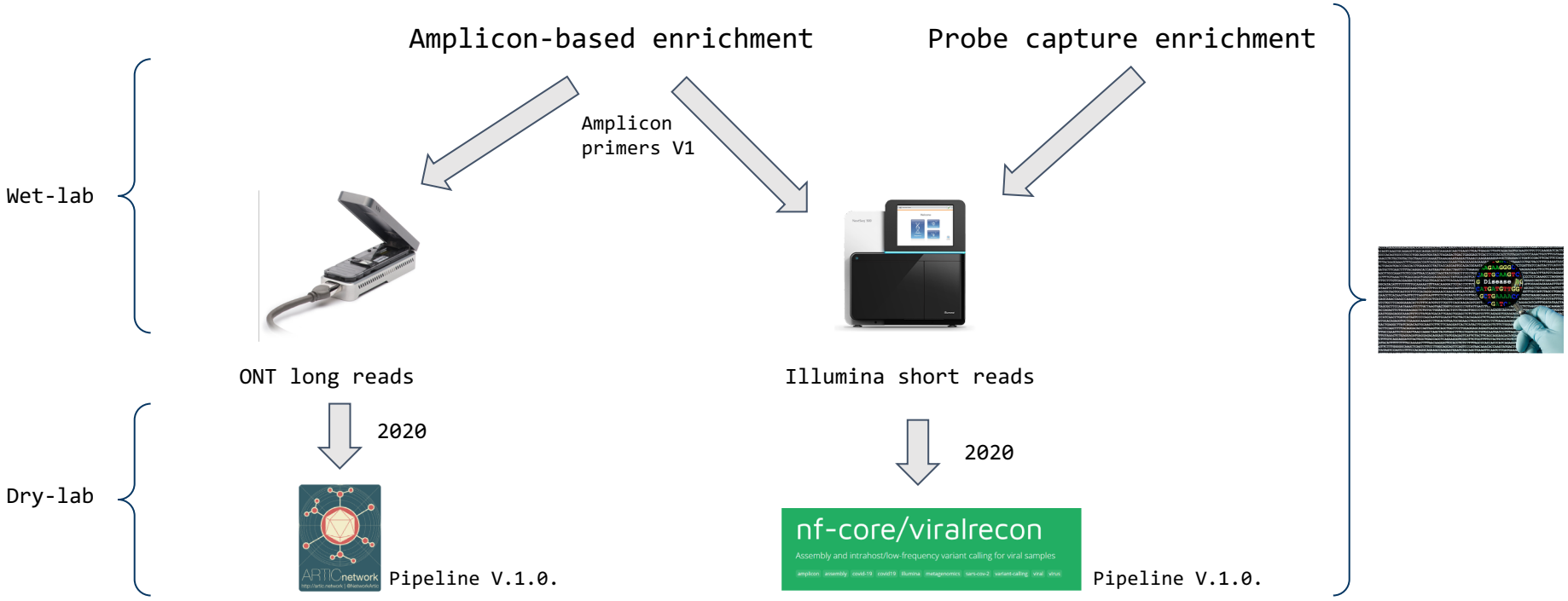
Surveillance:
Genome Sequencing



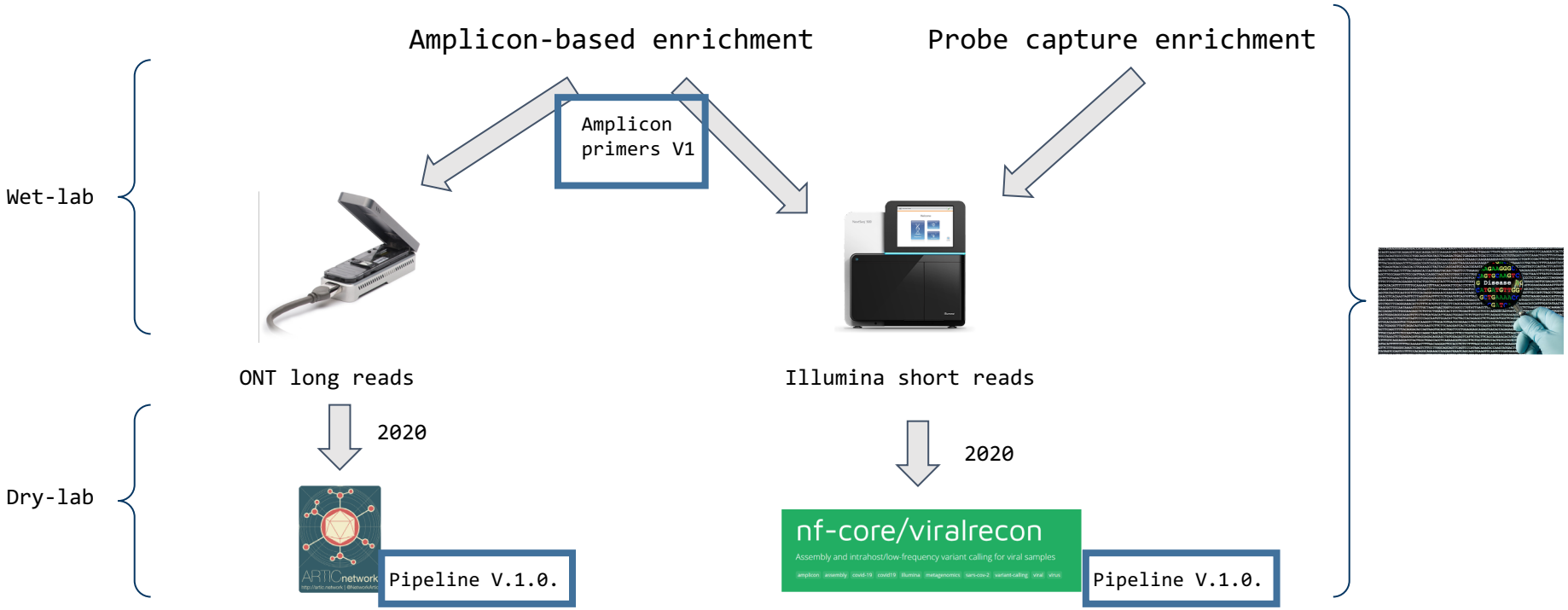
Detection:
RT-PCR

- RT-PCR Primer & Probe efficiency
- Viral variability behaviour
- Correlate viral lineage & severity
- Vaccine development and efficiency surveillance
- Future established disease control (flu)
- Coordinated & rapid response
- Understand immune response to infection

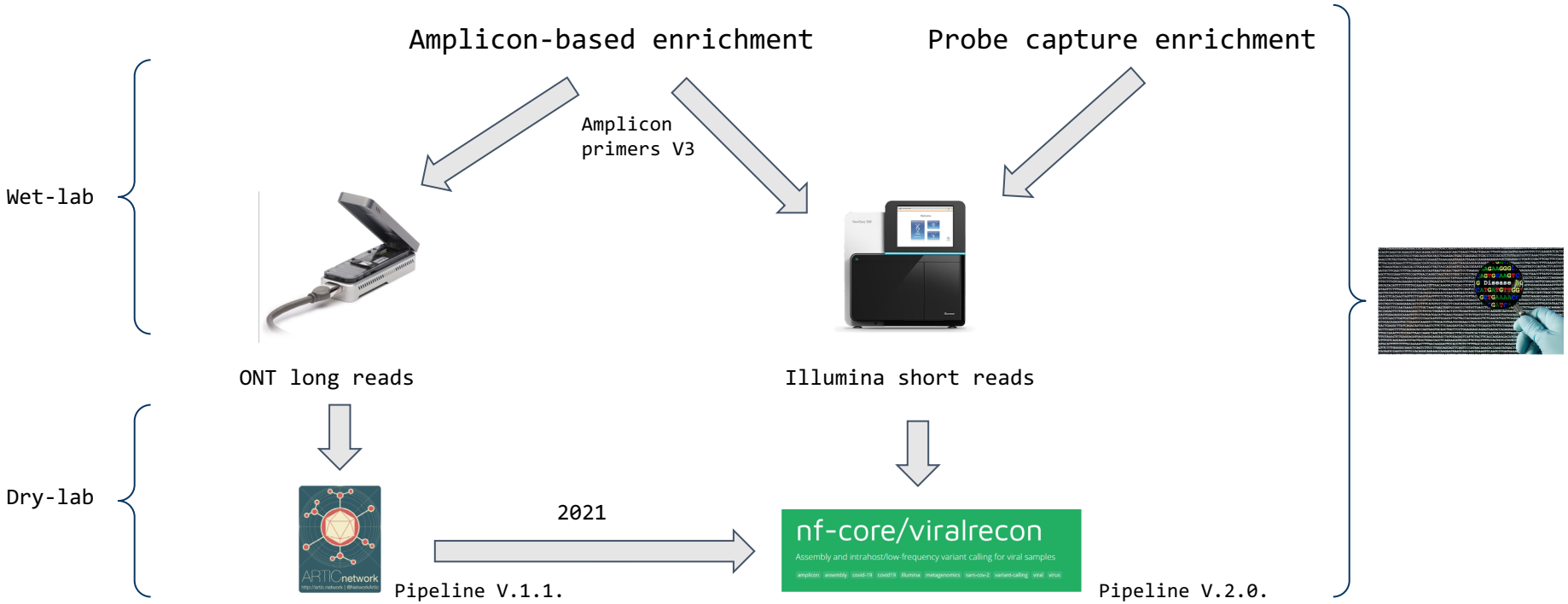
2. Sequencing approaches



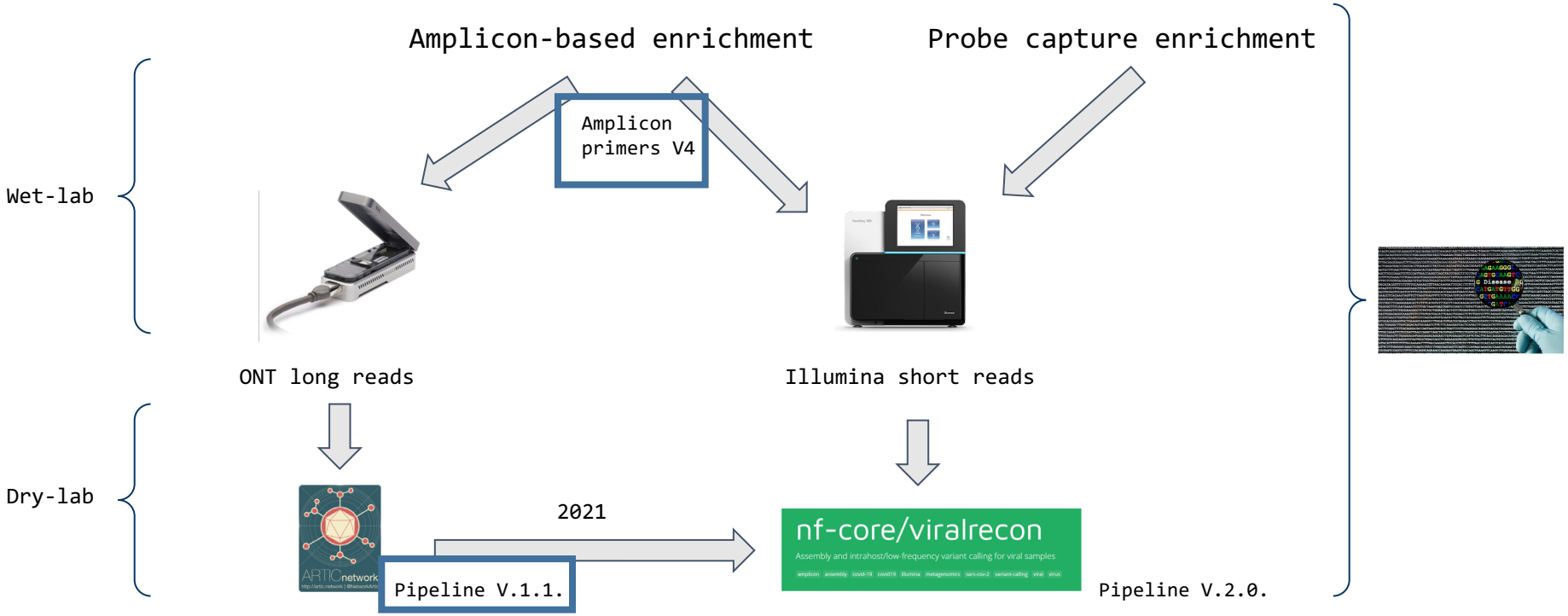
2. Sequencing approaches



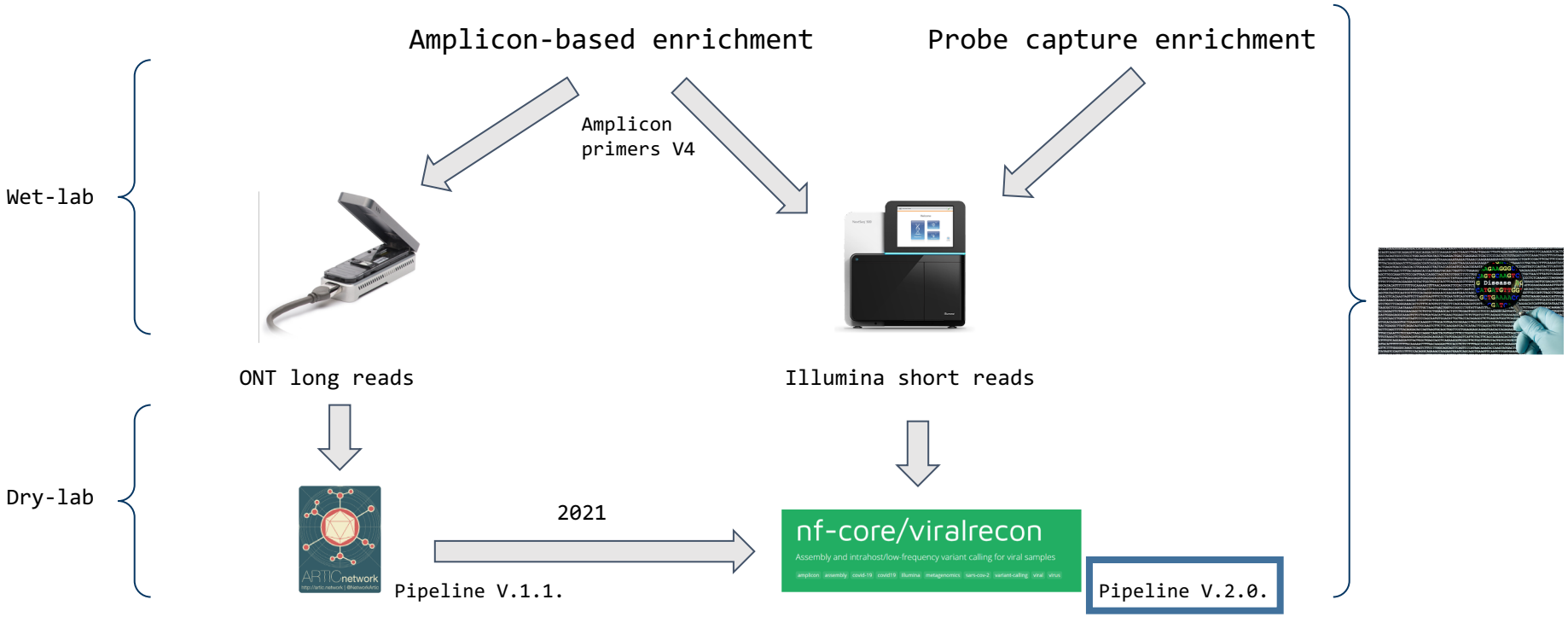
2. Sequencing approaches



2. Sequencing approaches



2. Sequencing approaches



3. History

January 22 2020 Artic Network protocols

https://github.com/BU-ISCIII/SARS_Cov2_consensus-nf

March 18 2020 Our 1st pipeline started

nf-core

March 30 2020 nf-core collaboration

April 5 2020 1st COVID19 Virtual BioHackathon

June 1 2020 viralrecon released v1.0.0

https://github.com/jaleezyy/covid-19-signal

June 29 2020 covid-19-signal v1.0.0

https://github.com/nodrogluap/nanostripper

September 3 2020 nanostripper v1.0.0

https://github.com/nf-core/viralrecon

May 13 2021 viralrecon v2.0 release

4. Standarization

nf-core/viralrecon

nextflow

https://github.com/nf-core/viralrecon

https://www.nextflow.io/

4. Standarization

Research used to 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

4. Standarization

- 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

4. Standarization

- 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

4. Standarization

- 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

4. Standarization

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

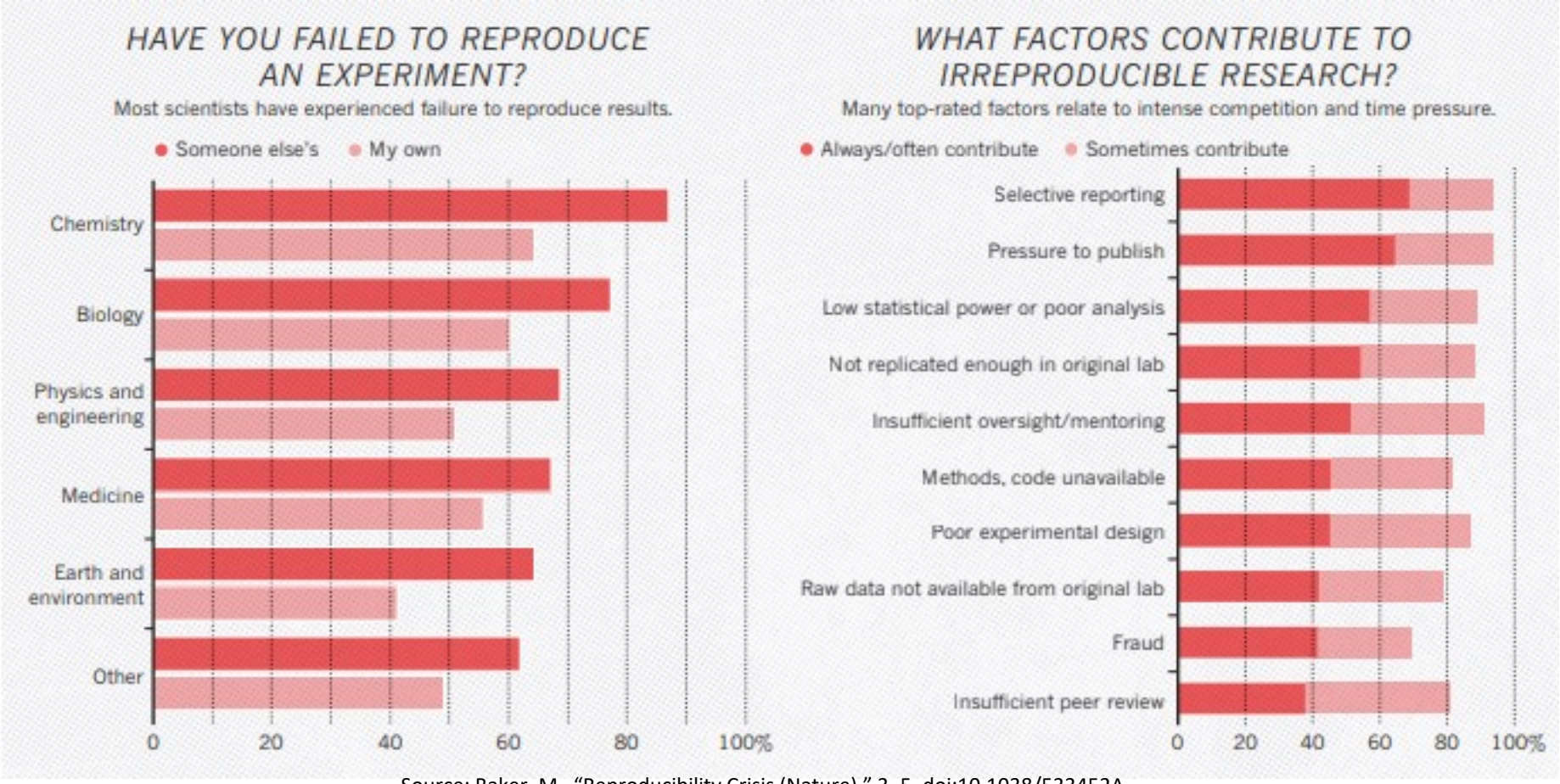
Pros	Cons
No need to storage intermediate files	Your data is in someone else's computer No backups or data management schemes
No need to install software Partial control over installed software	No control over installed software, versions and future availability
Graphic interface	No control over hidden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

4. Standarization

- 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

4. Standarization



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

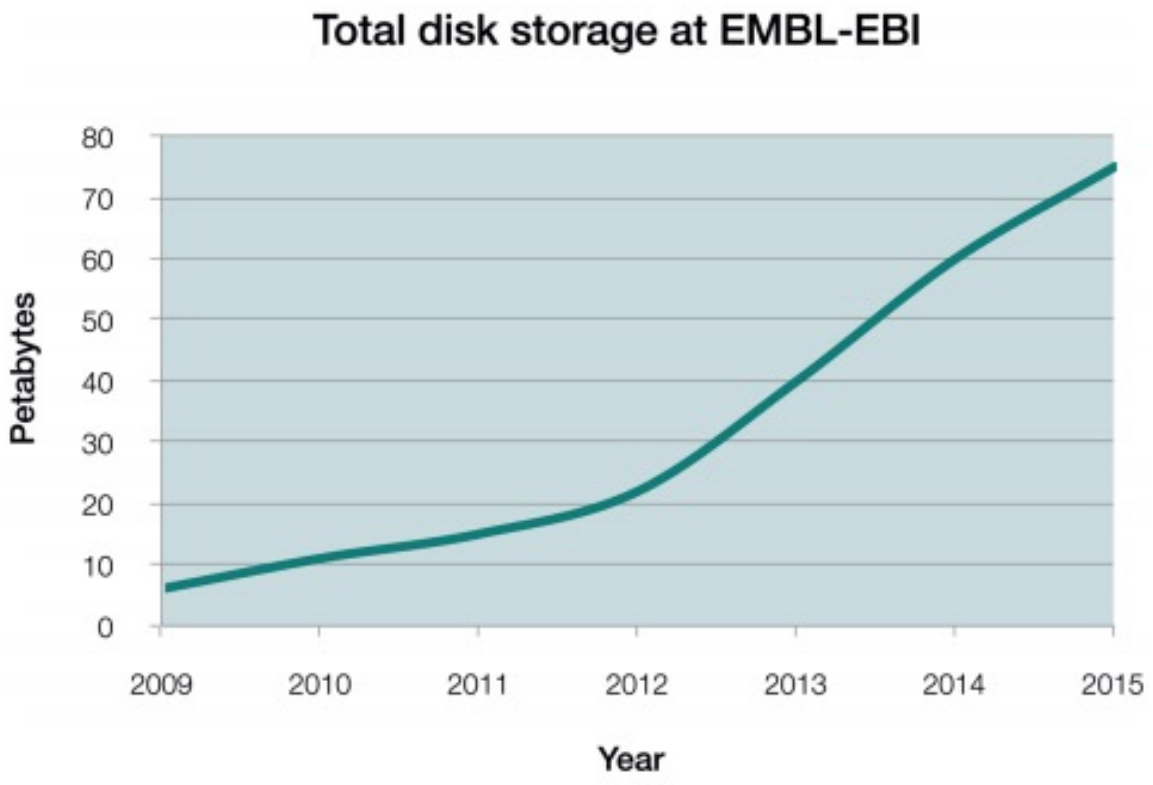
4. Standarization

	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

4. Standarization



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.

4. Standarization

1 sample	multiple samples
Research only: NGS was still a new thing, no applications 10 years ago	**Many applications:** research, clinical, industrial, forensic, military, …
Reproducibility is not needed: Why would anyone reanalyse this?	**Reproducibility, scalability , portability and standardisation are required**
Storage is not an issue: files of 1 sample fits everywhere in my HDD, maybe I will copy it in a CD-ROM	**Storage is challenging:** storage, indexation and backup required, privacy and legal standards
Computing is simple: no need to worry about resources or optimisation	**Computing requires optimisation and lots of resources**

4. Standarization

- Nowadays scientific computing paradigm

Pros	Cons
Data remains private Backups and data management schemes	High storage space Dedicated file systems Databases to index files
Control over software installed versions, open source programs	Many versions of the same software coexists
All parameters are available for the command	You have to understand all software variations
Analysis are reproducible and public	You have to publish and document your work

4. Standarization

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, …

4. Standarization

MRC | Cloud Infrastructure for Microbial Bioinformatics

HOME OVERVIEW PEOPLE GET STARTED NEWS TRAINING CONTACT

SYSTEM HIGHLIGHTS

The system will focus on features relevant to genomics researchers with features such as huge data storage capabilities, very high-memory research servers for maximum performance and integration with relevant biological databases.

7680 vCPU Cores

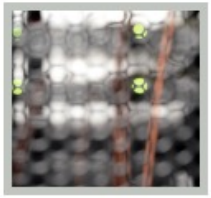
The CLIMB system is composed of over 7,500 CPU cores of processing power. This makes it the largest single system dedicated to Microbial Bioinformatics research, anywhere in the world.

500 Local Storage (TB)

To provide users with local, high performance, storage we have deployed IBM GPFS in each of the 4 sites, to provide 500TB of local storage. This storage is connected to our servers using Infiniband.

78 Total RAM (TB)

Unlike most supercomputers, the CLIMB system has been designed to provide large amounts of RAM, in order to meet the challenge of processing large, rich biological datasets. In comparison, the Spruce B supercomputer at the Atomic Weapons Research Establishment (number 68 on the Top 500 Supercomputers list, November 2014) has 35,000 cores, but only has 110 TB of RAM.

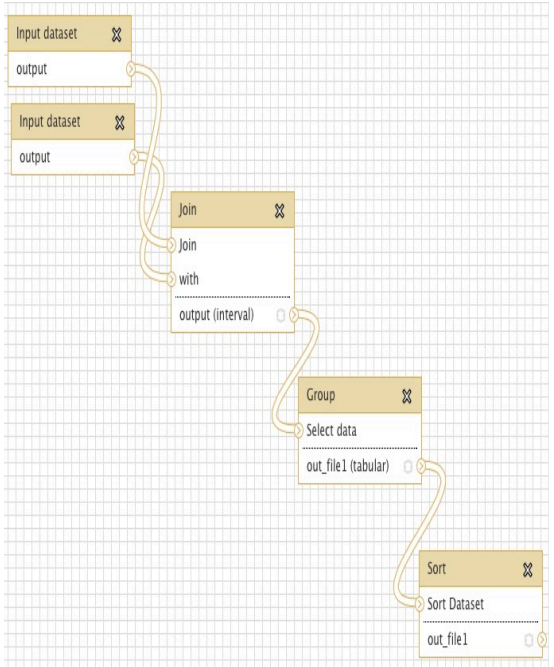


1000 Virtual Machines

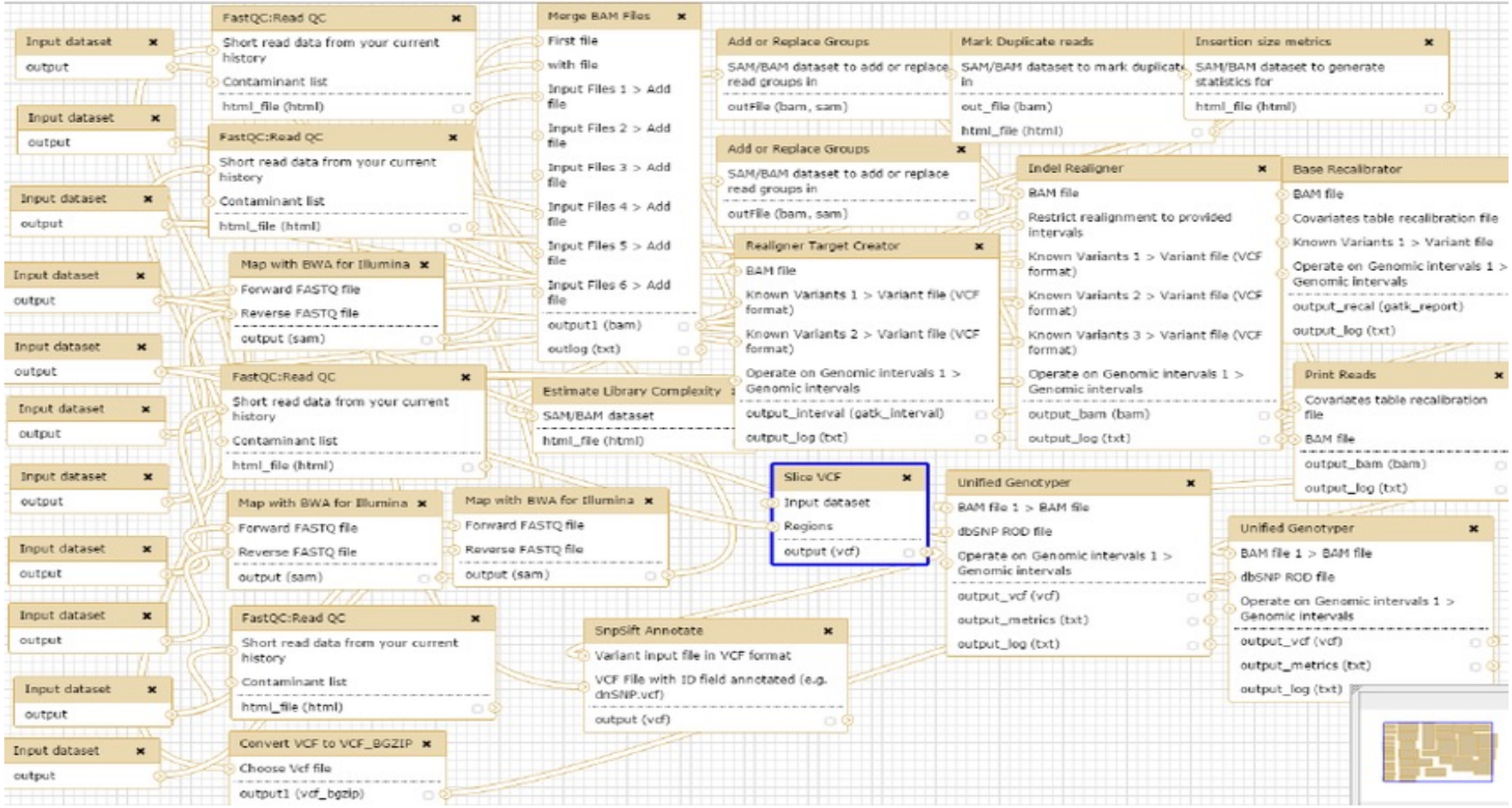
The CLIMB system is not designed to provide a single HPC system, as is often the case within academic computing; rather, the CLIMB system provides a pool of CPU cores and RAM that Medical Microbial Bioinformatics researchers can gain access to. The system has been designed to support over 1,000 VMs running simultaneously, potentially supporting most of the Microbial Bioinformatics community within the UK.

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



5. Nextflow

https://www.nextflow.io/

Sequencing used in clinical diagnosis, workflows have to assure:

- Reproducibility
- Portability: different platforms
- Scalability: different numbers of samples

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

DSL

domain-specific language

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

Portable

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

Portable

Continuous checkpoints

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

Portable

Continuous checkpoints

Reproducibility

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

Portable

Continuous checkpoints

Reproducibility

Unified parallelism

5. Nextflow

nf-core/viralrecon

nextflow

https://www.nextflow.io/

Fast prototyping

Portable

Continuous checkpoints

Reproducibility

Unified parallelism

Stream oriented

5. Nextflow

nf-core/viralrecon

https://www.nextflow.io/

nextflow

Multiple compute infrastructures

Portable

Easy to install

Reproducible

Stable code

5. nf-core

nf-core/viralrecon

nextflow

https://nf-co.re/

- Project (2018) Phil Ewels
- Standards pipelines
- Outgrow
- nf-core community
- Voluntary
- Proyects&Grats

5. nf-core

nf-core/viralrecon

nextflow

https://nf-co.re/

- Project (2018) Phil Ewels
- Standards pipelines
- Manteinance& visibility!
- Proyects&Grats

5. nf-core

nf-core/viralrecon

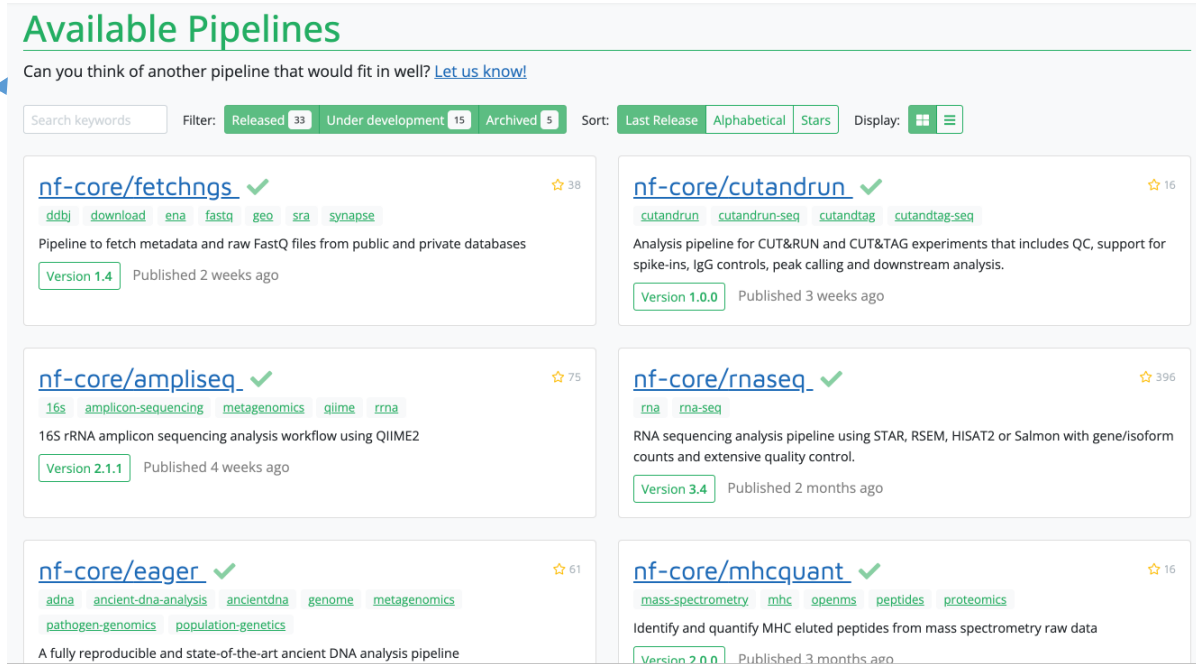
nextflow

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

VIEW PIPELINES

https://nf-co.re/



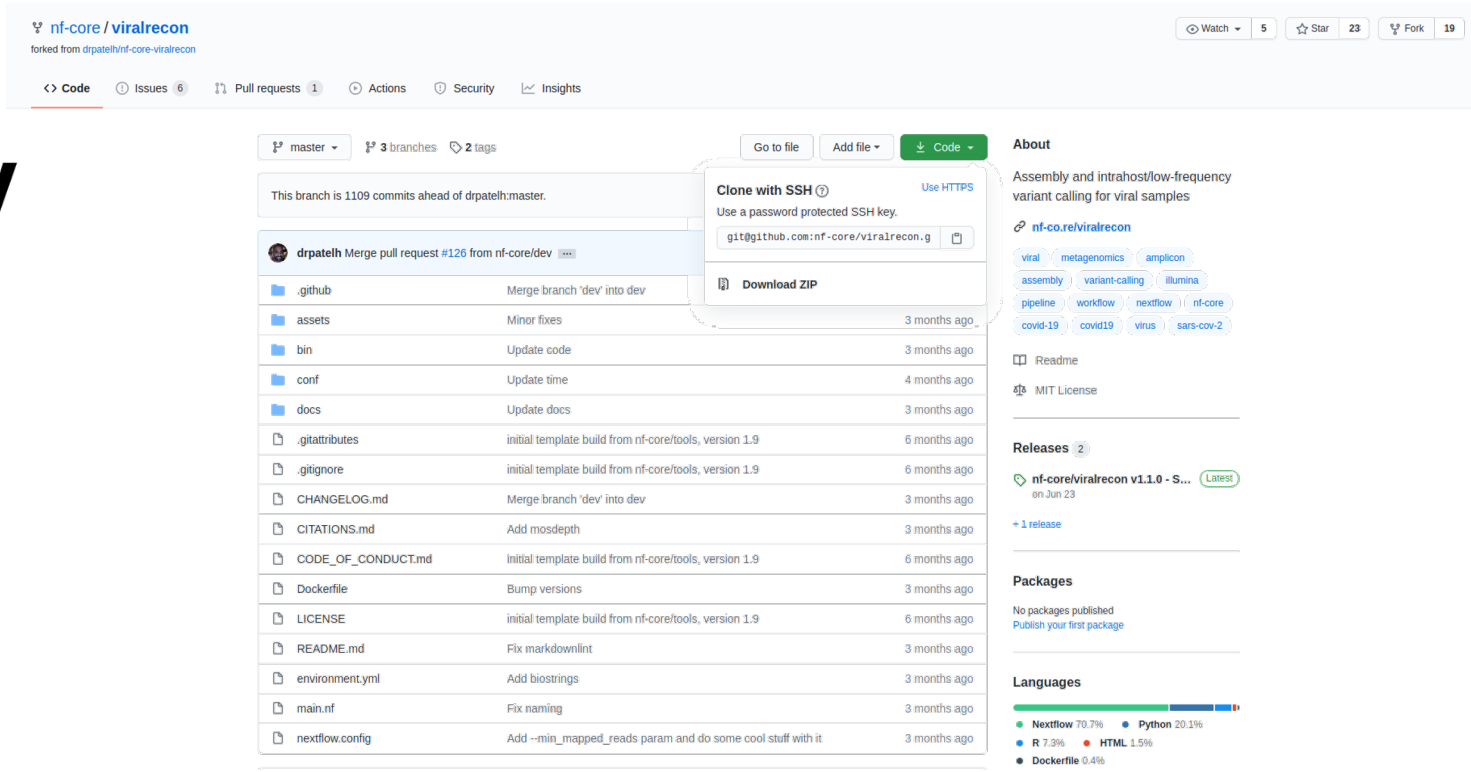
7. viralrecon

nf-core/viralrecon

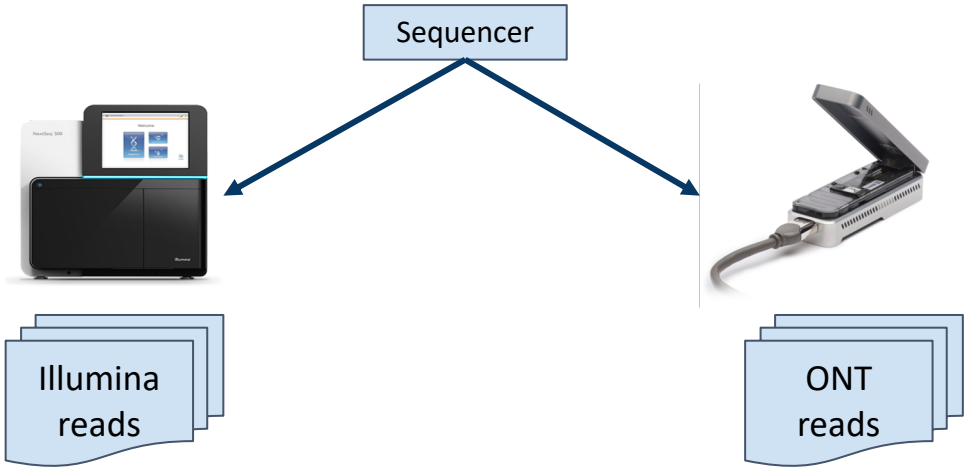
https://github.com/nf-core/viralrecon

https://github.com/BU-ISCIII/viralrecon

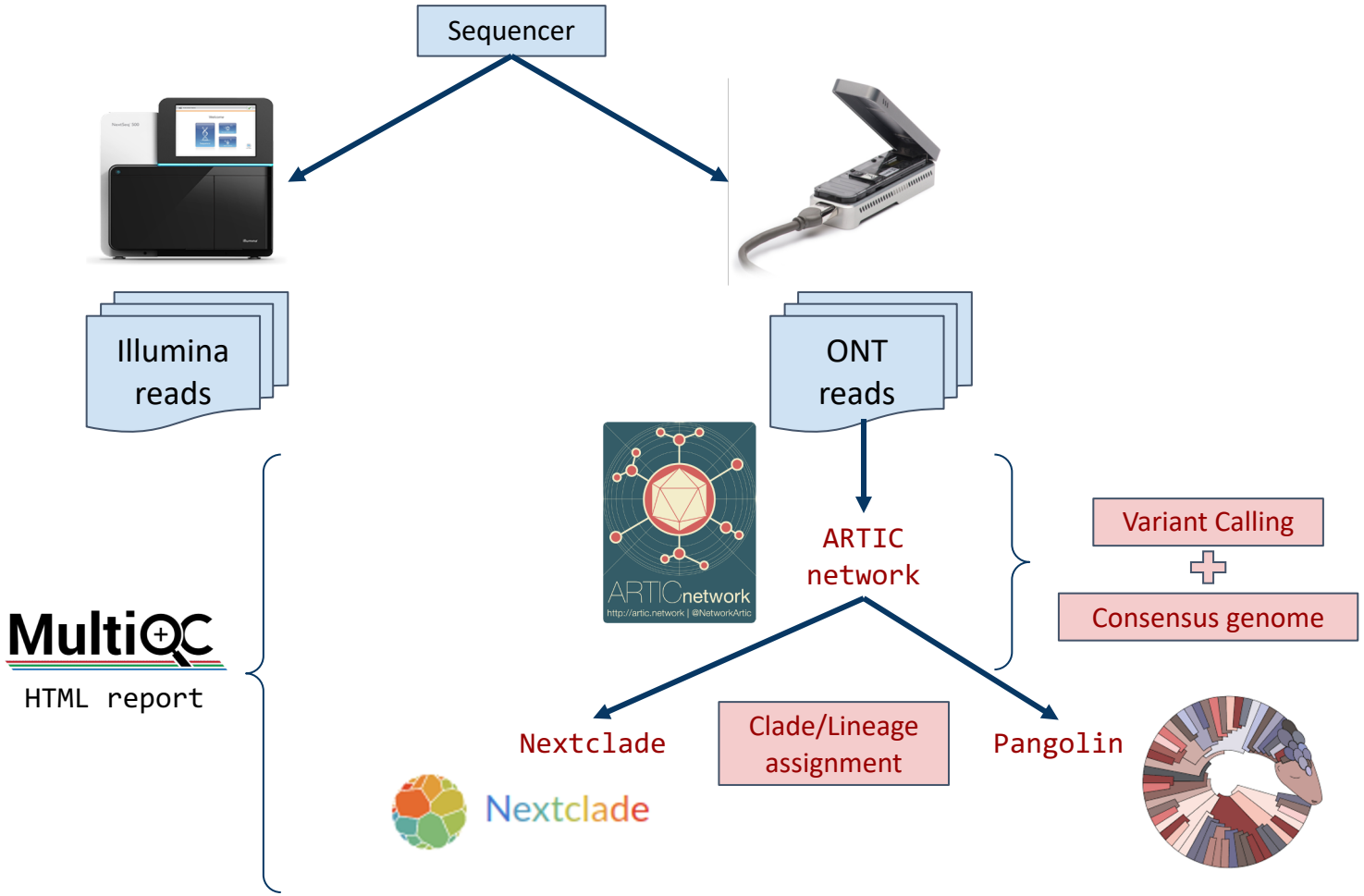
nextflow



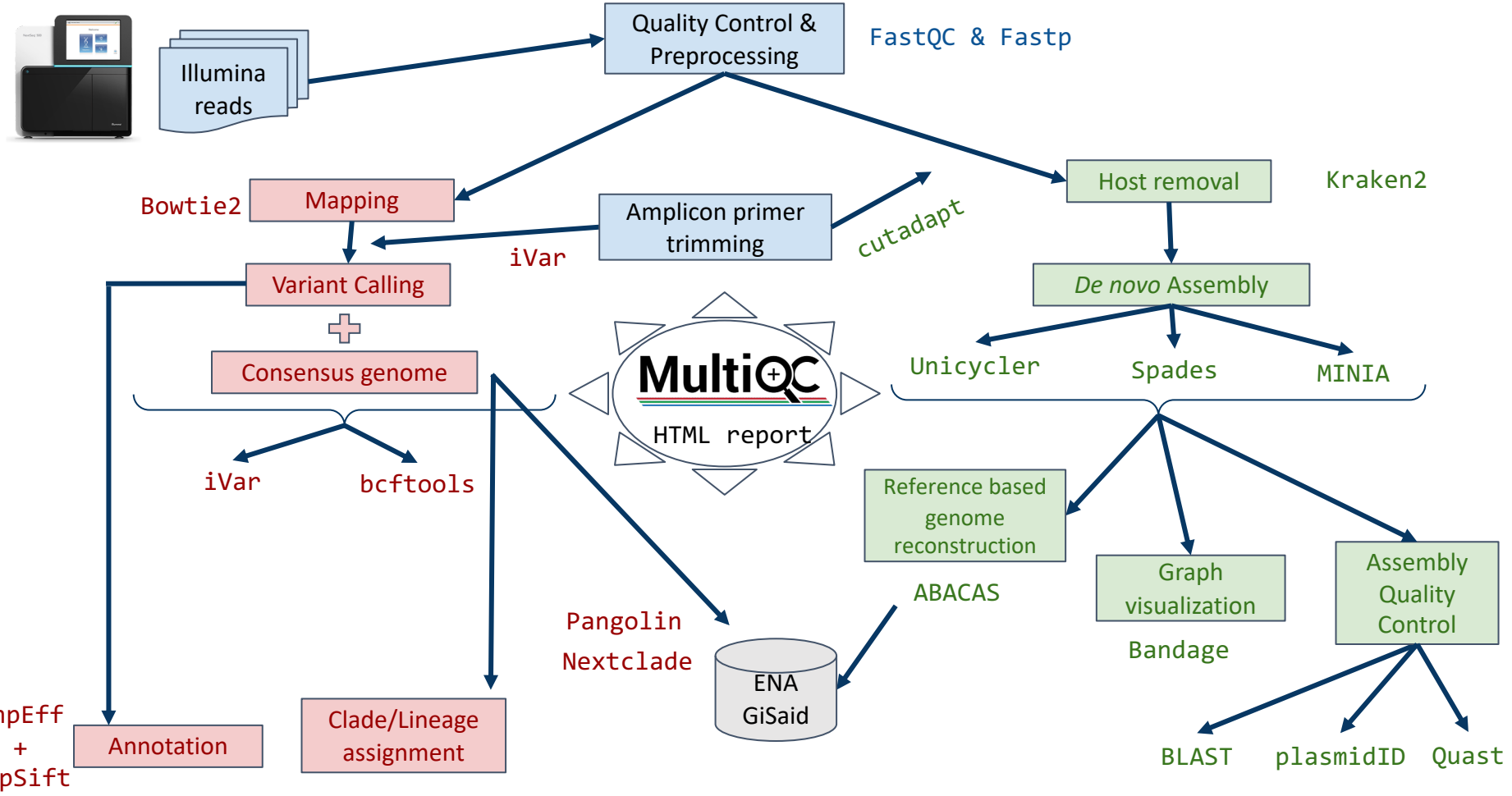
7.1. pipeline



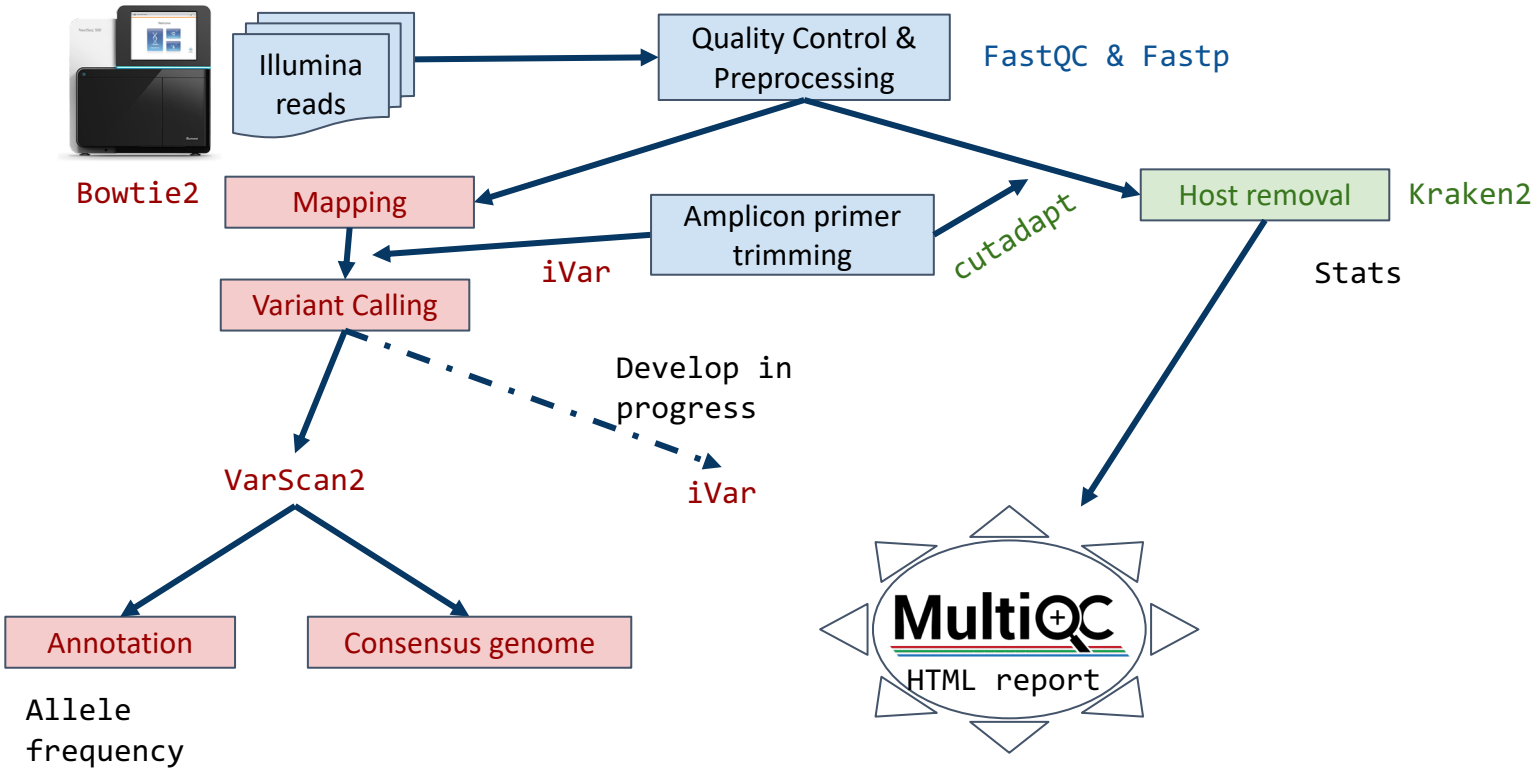
7.1. pipeline


Sequencer
Illumina reads
ONT reads
ARTIC network
Variant Calling
Consensus genome
MultiQC
HTML report
Nextclade
Clade/Lineage assignment
Pangolin
Nextclade

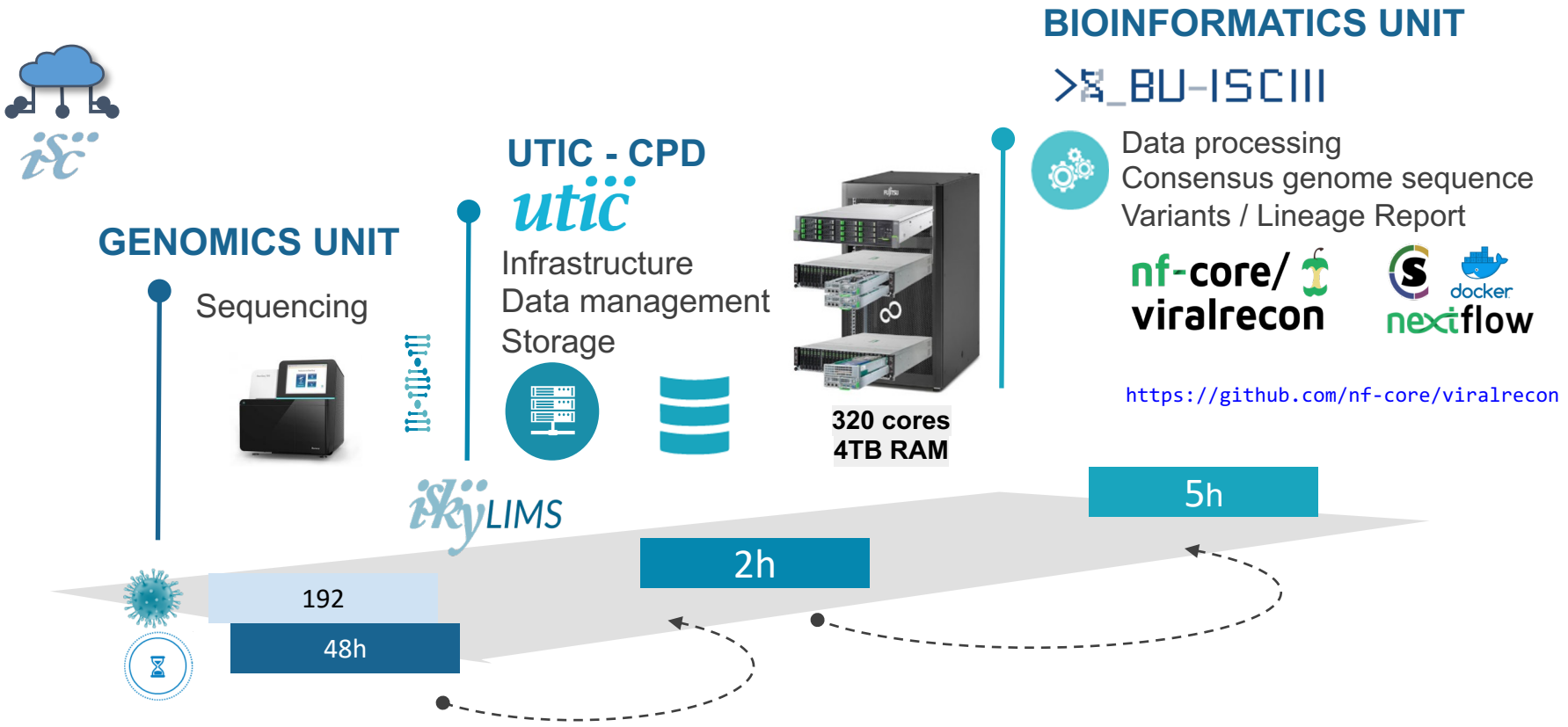
7.1. pipeline


Illumina reads
Quality Control & Preprocessing
FastQC & Fastp
Bowtie2
Mapping
Amplicon primer trimming
iVar
cutadapt
Host removal
Kraken2
Variant Calling
De novo Assembly
Consensus genome
MultiQC
HTML report
Unicycler
Spades
MINIA
iVar
bcftools
Reference based genome reconstruction
ABACAS
Graph visualization
Bandage
Assembly Quality Control
Pangolin
Nextclade
ENA
GiSaid
SnpEff
+
SnpSift
Annotation
Clade/Lineage assignment
BLAST
plasmidID
Quast

7.1. pipeline


Illumina reads
Quality Control & Preprocessing
FastQC & Fastp
Bowtie2
Mapping
Amplicon primer trimming
cutadapt
Host removal
Kraken2
iVar
Stats
Variant Calling
Develop in progress
VarScan2
iVar
Annotation
Consensus genome
Allele frequency
MultiQC
HTML report

7.1. pipeline


BIOINFORMATICS UNIT
>_BU-ISCIII
Data processing
Consensus genome sequence
Variants / Lineage Report
nf-core/viralrecon
docker
nextflow
https://github.com/nf-core/viralrecon
UTIC - CPD
utic
Infrastructure
Data management
Storage
320 cores
4TB RAM
GENOMICS UNIT
Sequencing
iSkyLIMS
192
48h
2h
5h

7.2. results

Mapping + Variant Calling + Consensus results:

1. Mapping .bam files → IGV
2. mapping_illumina.tab

host	Virussequence	sample	totalreads	readshost	%readshost	readsvirus	%readsvirus	unmapped reads	%unmaped reads	meanDP coverage virus	Coverage >10x(%)	Variants in consensusx10	Missense Variants	%Ns10x	Lineages
human	NC_045512.2	1910-21-8	825524	77610	9,4	4953	0,6	742961	90	9,34	12,61	20	14	82,91	None
human	NC_045512.2	218984	866586	270022	31,16	579919	66,92	16645	1,92	1127,06	86,58	32	19	9,29	AY.9
human	NC_045512.2	218985	951302	5968	0,63	934274	98,21	11060	1,16	1910,63	95,59	41	28	3,93	B.1.617.2
human	NC_045512.2	218986	927772	3728	0,4	919051	99,06	4993	0,54	1899,15	95,01	41	29	3,56	B.1.617.2
human	NC_045512.2	218987	779434	11082	1,42	765248	98,18	3104	0,4	1772,53	96,82	50	34	2,5	B.1.617.2
human	NC_045512.2	218991	882992	86020	9,74	790454	89,52	6518	0,74	1508,68	93,06	42	29	5,94	B.1.617.2
human	NC_045512.2	218993	724660	6436	0,89	715747	98,77	2477	0,34	1707,01	97,46	43	30	2,3	AY.5
human	NC_045512.2	219002	822458	12600	1,53	805022	97,88	4836	0,59	1976,08	98,13	46	32	0,92	AY.33
human	NC_045512.2	219003	829206	194660	23,48	546447	65,9	88099	10,62	626,65	66,72	29	19	27,19	AY.4
human	NC_045512.2	219004	1034136	765762	74,05	233198	22,55	35176	3,4	362,46	51,81	25	16	39,39	None
human	NC_045512.2	219005	832886	337032	40,47	472413	56,72	23441	2,81	553,85	62,94	31	20	32,15	None
human	NC_045512.2	219006	751474	242642	32,29	343649	45,73	165183	21,98	437,04	45,06	25	16	50,29	None
human	NC_045512.2	219007	935756	155234	16,59	771250	82,42	9272	0,99	1089,15	87,32	41	29	10,65	AY.4
human	NC_045512.2	219008	773260	8762	1,13	762434	98,6	2064	0,27	1639,52	93,07	40	29	4,68	AY.33
human	NC_045512.2	219009	843482	43918	5,21	797765	94,58	1799	0,21	1828,52	96,34	45	30	2,66	AY.33
human	NC_045512.2	219010	842022	107520	12,77	715887	85,02	18615	2,21	594,23	64,03	29	19	30,1	AY.7.1
human	NC_045512.2	219011	688048	438016	63,66	122541	17,81	127491	18,53	226,08	31,88	17	9	63,19	None
human	NC_045512.2	219012	1164144	1860	0,16	1071012	92	91272	7,84	1118,53	85,14	38	25	11,85	AY.4
human	NC_045512.2	219014	995438	43194	4,34	949150	95,35	3094	0,31	1994,32	95,01	45	30	3,16	B.1.617.2
human	NC_045512.2	219016	630838	383018	60,72	148625	23,56	99195	15,72	233,26	29,5	16	10	67,07	None
human	NC_045512.2	219017	934944	270604	28,94	629404	67,32	34936	3,74	559,31	63,02	33	21	32,04	None
human	NC_045512.2	219018	681136	513224	75,35	153460	22,53	14452	2,12	315,75	52,13	29	17	39,76	None
human	NC_045512.2	219019	785808	321366	40,9	452154	57,54	12288	1,56	690,17	72,55	32	20	22,83	AY.7.1
human	NC_045512.2	219020	633106	598280	94,5	19183	3,03	15643	2,47	44,19	15,02	8	4	82,98	None
human	NC_045512.2	219021	819454	285798	34,88	443489	54,12	90167	11	469,52	46,95	29	21	46,42	None
human	NC_045512.2	219022	1242186	391956	31,55	826054	66,5	24176	1,95	1025,52	85,73	37	23	10,48	AY.4
human	NC_045512.2	219023	986322	258072	26,17	639531	64,84	88719	8,99	605,84	66,88	32	22	24,28	AY.4
human	NC_045512.2	219024	1022824	111886	10,94	901415	88,13	9523	0,93	1217,56	87,54	38	23	8,87	AY.4
human	NC_045512.2	219025	807626	731652	90,59	33759	4,18	42215	5,23	76,27	28,15	8	3	66,87	None
human	NC_045512.2	219029	923796	175040	18,95	648320	70,18	100436	10,87	626,72	64,54	29	19	30,48	AY.20

7.2. results

Mapping + Variant Calling + Consensus results:

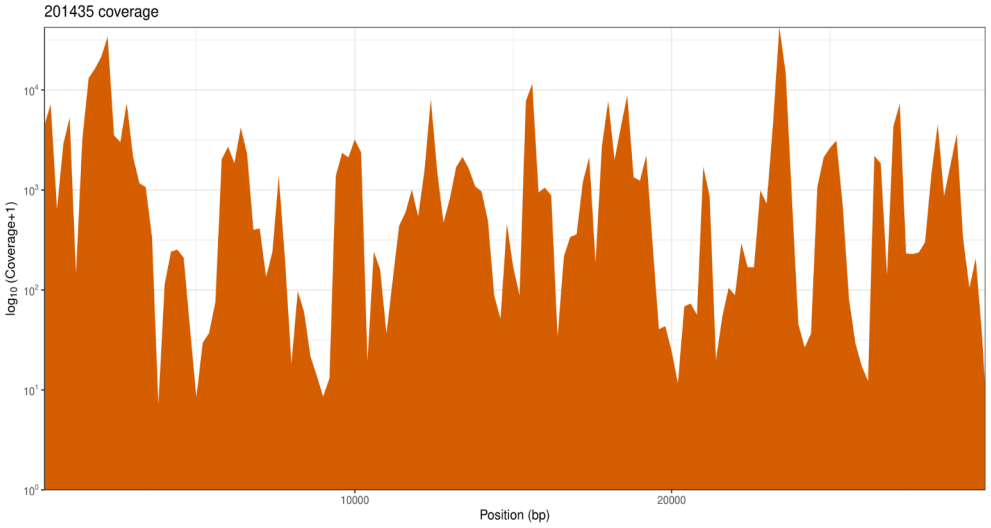
1. Mapping .bam files
2. mapping_illumina.tab
3. wide_variants_table.tab

POS	REF	ALT	GENE	EFFECT	HGVS_C	HGVS_P	1910.21.8 None	218984 AY.9	218985 B.1.617.2	218986 B.1.617.2	218987 B.1.617.2	218991 B.1.617.2	218993 AY.5	219002 AY.33	219003 AY.4
21575	C	T	S	missense_variant	c.13C>T	p.Leu5Phe	.	.	.	.	.	.	p.Leu5Phe	.	.
21595	C	T	S	synonymous_variant	c.33C>T	p.Val11Val	.	.	.	.	.	.	.	.	.
21608	G	T	S	missense_variant	c.46G>T	p.Val16Phe	.	.	.	.	.	.	.	.	.
21618	C	G	S	missense_variant	c.56C>G	p.Thr19Arg	.	.	p.Thr19Arg	p.Thr19Arg	p.Thr19Arg	p.Thr19Arg	p.Thr19Arg	p.Thr19Arg	.
21624	G	C	S	missense_variant	c.62G>C	p.Arg21Thr	.	.	.	.	.	.	.	.	.
21627	C	T	S	missense_variant	c.65C>T	p.Thr22Ile	.	.	.	.	p.Thr22Ile	.	.	.	.
21641	G	T	S	missense_variant	c.79G>T	p.Ala27Ser	.	.	.	.	.	.	.	.	.
21647	A	G	S	missense_variant	c.85A>G	p.Thr29Ala	.	.	.	.	.	.	.	p.Thr29Ala	.
21685	A	G	S	synonymous_variant	c.123A>G	p.Lys41Lys	.	.	.	.	.	.	.	.	.
21722	T	C	S	synonymous_variant	c.160T>C	p.Leu54Leu	.	.	.	.	p.Leu54Leu	.	.	.	.
21727	C	T	S	synonymous_variant	c.165C>T	p.Phe55Phe	.	.	.	.	.	.	.	.	.
21770	G	T	S	missense_variant	c.208G>T	p.Val70Phe	.	.	.	.	.	.	.	.	.
21802	T	C	S	synonymous_variant	c.240T>C	p.Asp80Asp	.	.	.	.	.	.	.	.	.
21830	G	T	S	missense_variant	c.268G>T	p.Val90Phe	.	.	.	.	.	.	.	.	.
21846	C	T	S	missense_variant	c.284C>T	p.Thr95Ile	.	.	.	.	p.Thr95Ile	.	.	.	.
21855	C	T	S	missense_variant	c.293C>T	p.Ser98Phe	.	.	.	.	.	.	.	.	.
21859	C	T	S	synonymous_variant	c.297C>T	p.Asn99Asn	.	.	.	.	.	.	.	.	.
21895	T	C	S	synonymous_variant	c.333T>C	p.Asp111Asp	.	.	.	.	.	.	.	.	.
21974	G	C	S	missense_variant	c.412G>C	p.Asp138His	.	.	.	.	.	.	.	.	.
21987	G	A	S	missense_variant	c.425G>A	p.Gly142Asp	.	.	p.Gly142Asp	p.Gly142Asp	p.Gly142Asp	p.Gly142Asp	p.Gly142Asp	p.Gly142Asp	.
22034	A	G	S	missense_variant	c.472A>G	p.Arg158Gly	.	.	.	.	.	.	.	.	.
28086	G	T	ORF8	missense_variant	c.193G>T	p.Ala65Ser	.	.	.	.	.	.	.	.	.
28111	A	G	ORF8	missense_variant	c.218A>G	p.Tyr73Cys	p.Tyr73Cys	.	.	.	.	.	.	.	.
28198	G	T	ORF8	missense_variant	c.305G>T	p.Cys102Phe	.	.	.	.	.	.	.	.	.
28253	CA	C	ORF8	frameshift_variant	c.361delA	p.Ile121fs	p.Ile121fs	.	.	.	.	.	.	.	.
28254	A	C	ORF8	missense_variant	c.361A>C	p.Ile121Leu	p.Ile121Leu	.	.	.	.	.	.	.	.

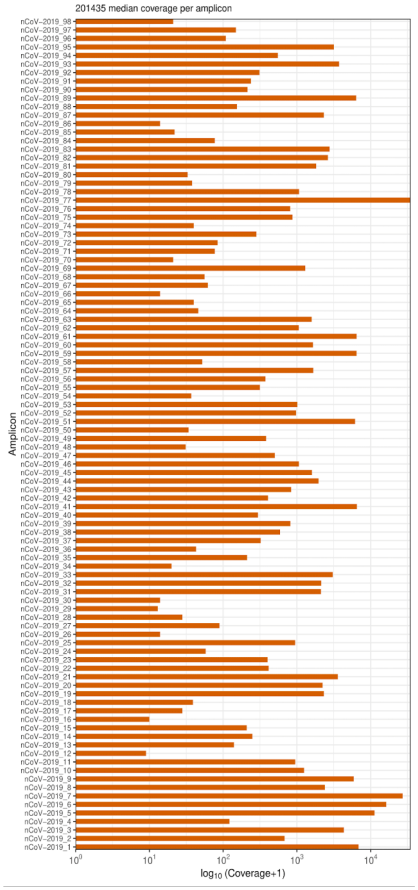
7.2. results

Mapping + Variant Calling + Consensus results:

1. Mapping .bam files
2. mapping_illumina.tab
3. wide_variants_table.tab
4. {sample_name}.trim.genome.regions.coverage.pdf
5. {sample_name}.trim.amplicon.regions.coverage.pdf



Amplicon coverage



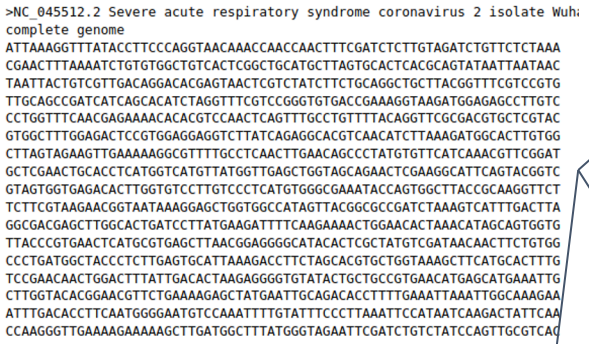
7.2. results

Assembly results:

1. summary_assembly_metrics_mqc.tsv

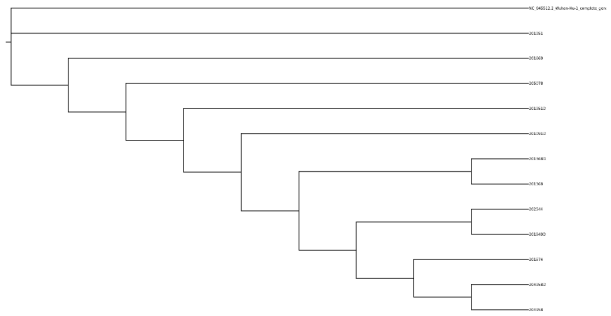
Sample	# Input reads	# Trimmed reads (fastp)	% Non-host reads (Kraken 2)	# Contigs (Assembly)	Largest contig (Assembly)	% Genome fraction (Assembly)	N50 (Assembly)
218069	1031628	803716	29.77	705	2983	60.984	1445.0
218079	1115296	866236	20.2	364	1746	34.592	916.0
218081	1075736	840544	98.49	4469	3109	87.573	1642.0
218092	1066116	834718	1.58	435	1001	1.836	661.0
218096	888172	688258	2.14	686	2680	-	625.0
218101	1303196	1013254	93.29	1841	1621	47.229	988.0
218103	1326720	1015592	86.24	1250	4578	58.666	1063.0
218104	1207650	939600	98.95	141	8233	89.677	7153.0
218105	1229108	939214	65.83	739	3005	71.732	1078.0
218106	1210156	911818	85.87	1020	3829	68.812	1159.0
218109	1071088	835558	24.17	1319	3708	60.422	792.0
218112	1220966	957970	1.5	409	1878	-	676.0
218113	1280156	988744	5.11	393	2362	18.042	1158.0
218721	1492478	1061422	99.61	4494	2431	86.483	1342.0
218722	1153738	808502	5.71	1234	1643	-	717.0
218723	1152466	817562	98.95	4270	4734	85.971	1792.0
218724	1301510	1010334	0.84	152	1376	4.013	698.0
218726	1462832	1076924	1.74	383	1755	3.842	1149.0
218727	622536	241854	0.8	200	961	3.214	662.0
218729	1442978	1087590	63.21	962	7214	83.851	1188.0
218730	1410868	868454	10.39	5347	5333	3.224	614.0
218731	1353858	993700	6.24	527	617	-	562.0
218732	1104916	855390	55.75	332	4565	73.274	1698.0
218734	1453500	1112340	5.64	944	2882	3.401	846.0
218735	1370332	1053964	6.27	839	1507	3.224	964.0

7.2. results



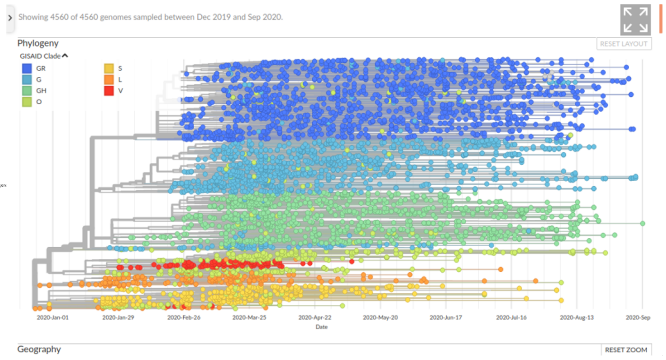
.fasta consensus genome

Outbreak detection



Other analyses

Clade classification



7.2. results

¿Questions?

Thanks to all my colleges in BU-ISCII

Thanks to Genomics Unit and the Reference Laboratory in Respiratory Viruses

Sara Monzón
Luis Chapado
Erika Kvalem
Alberto Lema
Isabel Cuesta

(Find us in https://github.com/BU-ISCIII)

And special thanks to the nf-core community (https://nf-co.re/)

Harshil Patel
(https://github.com/drpatelh)
Bioinformatics & Biostatistics Group at The Francis Crick Institute, London, now in Seqera Labs