Oxford Nanopore Technologies

Data analysis for all levels of expertise

Accessible, scalable, workflows and tools

Jonghwa Ahn, Ph.D
Senior Field Applications Scientist
Oxford Nanopore Technologies

Oxford Nanopore: from sample to answer

Comprehensive solutions for library preparation, sequencing and data analysis

Prepare

- Field kits
- Lab kits
- Manual & Automated

Sequence

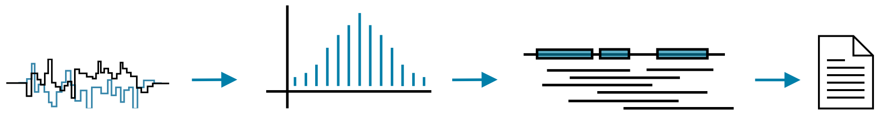
- Field devices
- Lab devices
- Low & high output platforms (low $ / test & low $ / Gb)

Analyse

- Accessible
- Scalable
- Versatile

Which sequencing data analysis is covered?

Oxford Nanopore provides a full range of solutions

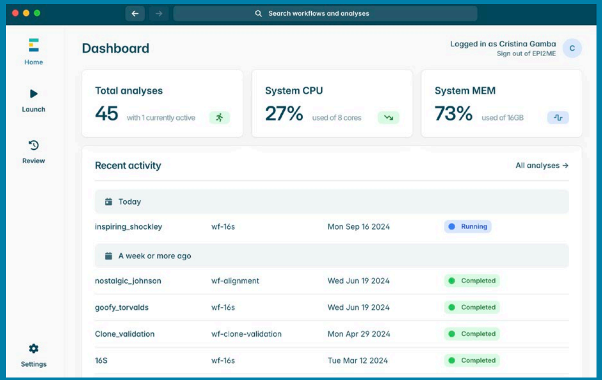


			Bioinformatics expertise	OS	Basecall	QC	Analyse	Report
Full support	MinKNOW		●○○	Windows, Mac & Linux	✓	✓	✓	✓
	EPI2ME workflows*	local	●○○	Windows, Mac & Linux	✓	✓	✓	✓
		cloud	●○○	Windows, Mac & Linux		✓	✓	✓
Limited support	Research tools		●●●	Linux	✓		✓	
	Community developed tools		●●●	Linux	✓	✓	✓	

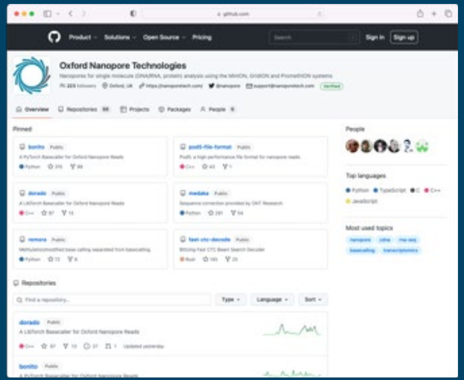
* Accessible from both an intuitive interface and the command line

More at: https://nanoporetech.com/data-analysis

Oxford nanopore data analysis options



EPI2ME



GitHub

	EPI2ME	GitHub
Location	Local, distributed or in the cloud	User defined (laptop, cluster or cloud)
Timing	Post-run or real time	Post-run
Configurability	Pre-configured	User defined
Reporting	Detailed output, shareable reports	User built
Operating systems	Windows, Mac, Linux	User defined
Expertise needed	● ○ ○	● ● ●

Download the EPI2ME app for free

Get the latest tools from GitHub

Basecalling

- MinKNOW
- Dorado

MinKNOW: overview

Our devices software: from basecalling to reporting

Basecalling

QC

Analyse

Report

Basecalling

- Real-time basecalling
  - Choose desired accuracy
- Methylation calling

Real time QC

- Real-time quality control of the run
  - Stop and resume as needed

Analyse your data

- Seamless Analysis in real time including
  - Enrichment or depletion with Adaptive sampling
  - Alignment
  - Barcode demultiplexing
  - And more

Generate automatic report

- QC reports of each run

Store only the data you need

Data formats

Oxford Nanopore format — POD5 (or FAST5)

- Contains raw signal data and read metadata, permitting re-basecalling
- Large in size
- Optional

Standard format — FASTQ

- Standard format for basecalled sequence data with per-read quality scores
- Text-based file format, small in size
- Default

Standard format — BAM

- Standard sequence alignment format
- Compressed text-based format, small in size
- Optional / default for modified bases

* Size may vary depending on information stored

Example storage requirements

GPU: V100 GPU
Memory:64 GB RAM
Storage: **4 TB SSD**

Per flow cell storage examples

Flow cell output	POD5 (optional storage)	FASTQ.GZ	Unaligned BAM with modifications
30 Gb	150 GBytes	19.5 GBytes	18 GBytes
150 Gb	1 TByte	97.5 GBytes	90 GBytes

Full run storage examples

Assuming 30/150 Gbases per flow cell

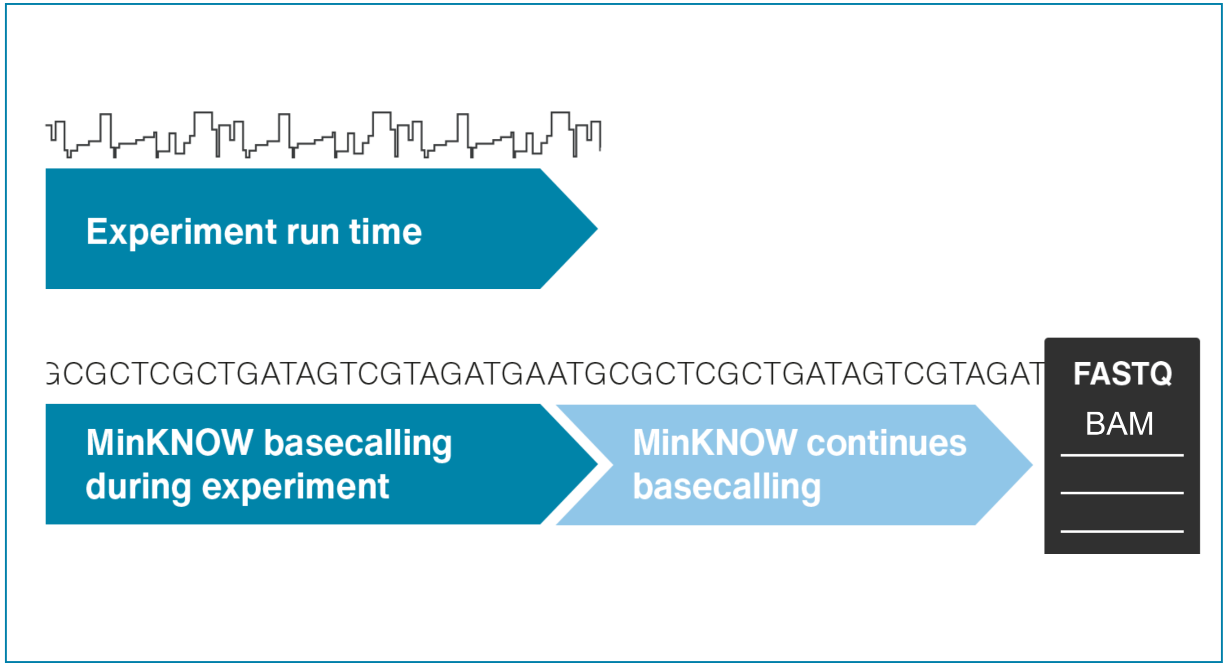
No. flow cells / run	POD5 (optional storage)	FASTQ.GZ	Unaligned BAM with modifications
5 MinION flow cells	750 GBytes	97.5 GBytes	90 GBytes
2 PromethION flow cells	2 TBytes	195 GBytes	180 Gbytes

All values are approximate

Canonical basecalling

From raw data to sequence

- Transform the “squiggle” into sequence with AI
- Basecall in real-time with MinKNOW
- Post-run basecalling with MinKNOW or Dorado
- Choose the most suitable model
  - HAC is faster, recommended for most applications
  - SUP is slower, recommended for *de novo* assembly projects or low-frequency variant analysis
- Methylation calling can be switched on with a click

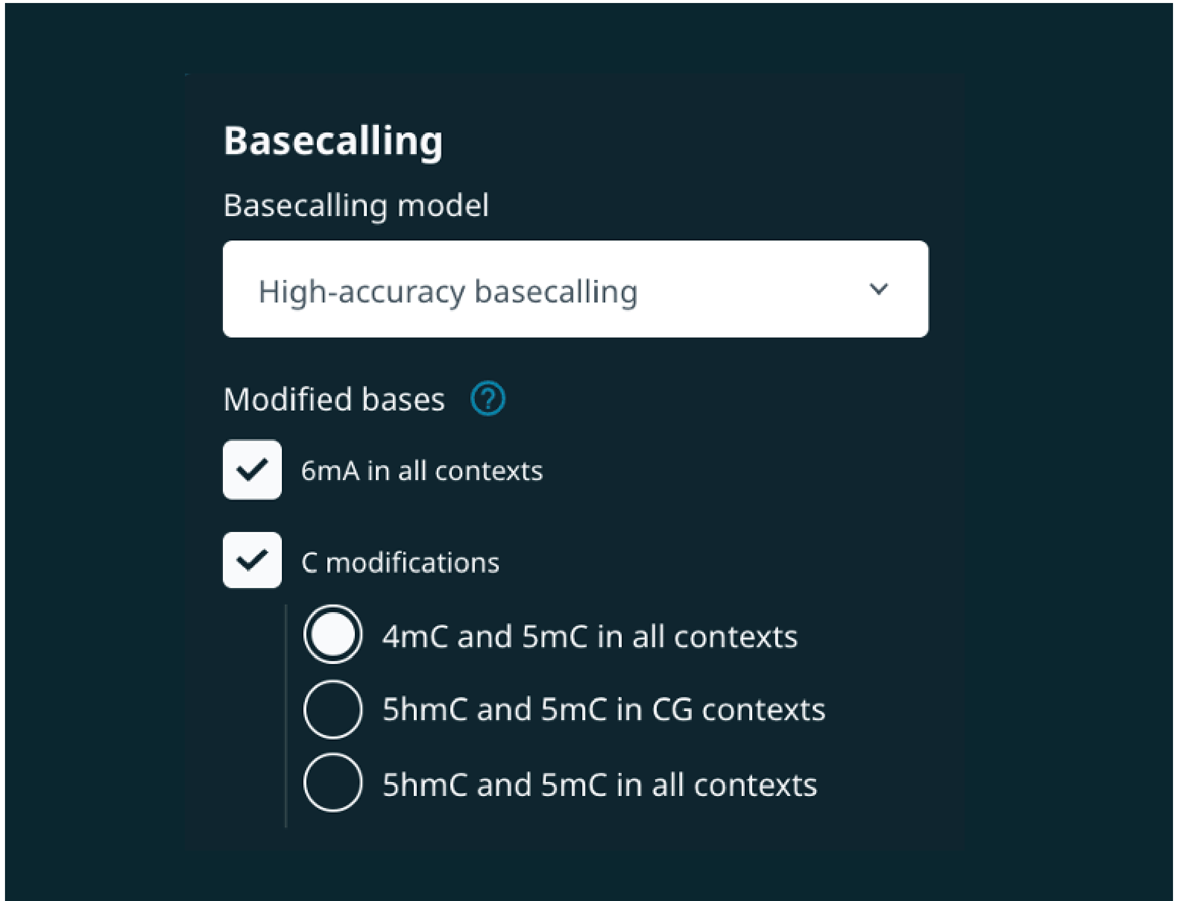


Find out more at: nanoporetech.com/how-it-works/basecalling

Inbuilt modification calling

Most accurate methylation

- Seamless base modification analysis of native DNA data in MinKNOW
- In parallel to basecalling, real-time options
- Currently supports multiple DNA and RNA modification calling
- Latest models available in Dorado standalone



Basecalling options in MinKNOW

Find out more at: https://nanoporetech.com/platform/accuracy

Canonical basecalling models and versions available

In MinKNOW and Dorado

Basecalling Models

	Model Version	Key Features	Dorado	MinKNOW
DNA	v4.3	Live basecalling with HAC Improved bacterial/low-complexity regions Mods: 5mC/5hmC, 5mC, 6mA	✓	✓
	v5.0*	Raw read accuracy ~Q26 SUP bacterial improvements Mods: 4mC/5mC Raw read accuracy ~Q21 HAC	✓	✓
	v5.2*	Raw read accuracy ~Q22 HAC	✓	in following version
RNA	v3.0.1	Raw read accuracy ~Q13 SUP Mods: m6A DRACH	✓	✓
	v5.0.0*	Raw read accuracy ~Q19 SUP Mods: m6A, pseU	✓	✓
	v5.1*	Mods: inosine, m5C	✓	✓
	v5.2*	Mods: 2'Ome (A,C, U, G)	✓	in following version

* SUP transformer architecture

Modified Bases

	Model	Context	Dorado	MinKNOW
DNA	5mC/5hmC (mammal/plant)	CpG	✓	✓
		All	✓	✓*
	4mC/5mC (bacterial)	All	✓	✓*
	6mA	All	✓	✓*
RNA	m6A	DRACH	✓	✓*
		All	✓	✓*
	PseU	All	✓	✓*
	m5C	All	✓	✓*
	Inosine	All	✓	✓*
	2'Ome-A	All	✓*	in following version
	2'Ome-C	All	✓*	in following version
	2'Ome-U	All	✓*	in following version
	2'Ome-G	All	✓*	in following version

*Currently only available in SUP

Accuracies for v5.0 modified base models

In MinKNOW 25.05.14

Modified base model	Context	HAC accuracy	SUP accuracy
4mC and 5mC	All	97.1%	98.0%
5mC and 5hmC	CpG	99.3%	99.3%
	All	98.6%	98.7%
6mA	All	97.5%	98.3%

Benchmarking Modified Base Detection with Modkit in Oct. 2024

Modified Base	Context	HAC Accuracy	SUP Accuracy
5mC+5hmC	All	97.30	97.80
5mC+5hmC	CpG	98.21	98.15
5mC only	All	99.20	99.48
5mC only	CpG	99.76	99.81
6mA	All	96.21	97.60

Blackwell Architecture GPU Support

In MinKNOW 25.05.14

		HAC basecalling only	HAC basecalling with alignment	SUP basecalling only	SUP basecalling with alignment
Standalone MinKNOW on NVIDIA GeForce RTX 5090 - MinION Mk1D	DNA (30 kb human)	1	1	1	1
	RNA	1	1	1	1
Standalone MinKNOW on NVIDIA GeForce RTX 5090 - PromethION 2 Solo	DNA (30 kb human)	2	2	0.8	0.8
	RNA	2	2	0.8	0.8

- 30 Gb output over 72 hours from a MinION flow cell running 30 kb human reads
- 100 Gb output over 72 hours from a PromethION flow cell running 30 kb human reads
- 7 Gb output over 72 hours from a MinION flow cell running RNA
- 35 Gb output over 72 hours from a PromethION flow cell running RNA

Sequence data types

File formats: FASTA

The reference file

Notes

- Header always begins with >
- Unique Identifier followed by optional information separated by a space
- Followed by lines of sequence
- Required input for many applications

Often used for read alignments:

- Reference bias
- Population representation
- Species diversity
- Strain/breed/sub-species relatedness
- Not representative of any one member

Unique identifier

Optional metadata

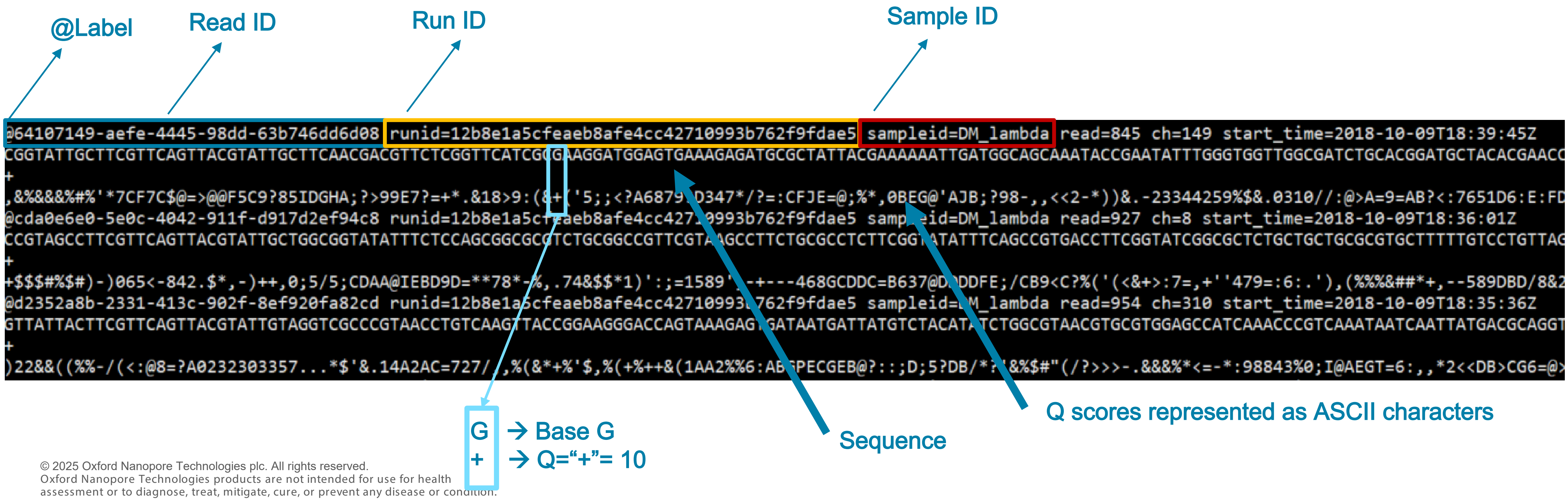
```
>XI dna:chromosome chromosome:R64-1-1:XI:1:666816:1 REF
CACCACACCCACACACCACACCCACACACACACCACACCCACACACCACACCCACACACC
ACACCCACTACTCTAACCCTATTCTAATCCAACCCTGATCAACCTGTCTCCAAACCTACC
CTCACATTACCCTACCTCTCCACTCGTTACCCTGCCCCACTCAACCATACCACTCCCACC
CACCATCCATCTCTCTACTGTCACCAGCCCACCGTCCACCATAACCGTTACCCTCCCATT
ACTCATATTTAACCCCACTACCACTTACCCTGCCATTACCCTACCATCCACCATGTCCTA
CTCACTGTACTGTTGTTCTACCCTCCATATTGAAACGTTAACAAATAATCGTAAATAATA
CACATATACTTACCCTACCACTCTAATCCCACCACACATCACATGCCATACTCACCTTCA
CTTGTATACTGATATGGTATACGCACACGGATGCTACGTATATACCACTCTCAACTTACC
CTACTCTCACATTCCACTCCATGGCCCAGTCTCACTAAATCAGTACGATGCACTCACATC
ATTATTCACGGCACTTGCCTCAGCGGTTTATACCCTGTGCAATTTACCCATAAAACCCAC
GATTATCCACATTTTAATATCTATATCTCATTCAGCGGCTCCAAATATTGTATAACTGCT
CTTAATGCATACATTATACCACTTTTACTCCATATACTAACCATTCAATTTATACACACT
TATTTCAATATACCCACAAAATCACCACTAAAATTACCTAAACATAAAAATATTCTACTC
TTCAACAATAATACATAAACACACTCAATTGCGTATCTATACCACCATGACGTCATTAAC
```

Reference sequences provide coordinate systems to anchor our prediction against

This facilitates the comparison of predictions derived from different samples and studies

File formats: FASTQ

- *De facto* format for most sequence data types
- Standard text file (huge in size)
- Blocks of 4 lines correspond to data for 1 read
  - Header, sequence, header (denoted by "+"), per-base quality



File Formats: SAM & BAM

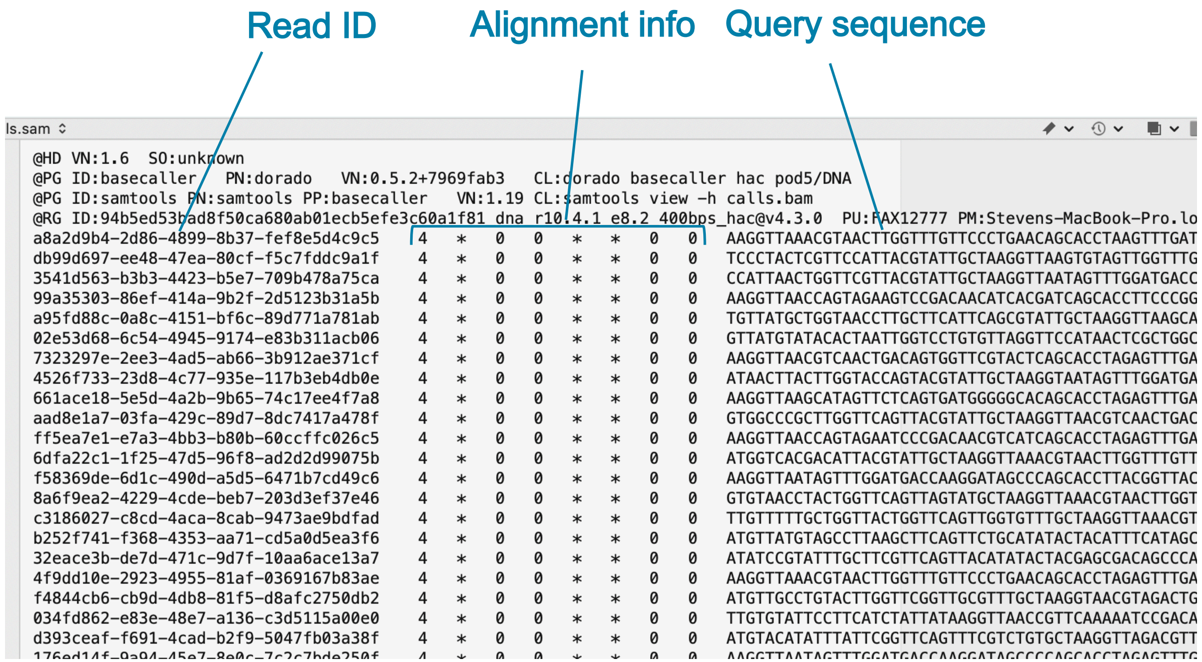
SAM files

- Sequence alignment map (SAM) first published by Li *et al.*, 2009
- File specification: https://www.htslib.org/doc/sam.html
- SAM contains one line for each read and has 12+ columns containing:
  - Coordinates, mapping quality, and metrics describing each read alignment

- <u>**Reads can also be stored unaligned to a reference genome**</u>

BAM files

- BAM file is the binary (computer readable) version of a SAM
- Same information
- Much smaller storage space required
- Can index BAMs for rapid, random access
- If you have SAM, you should convert to BAM

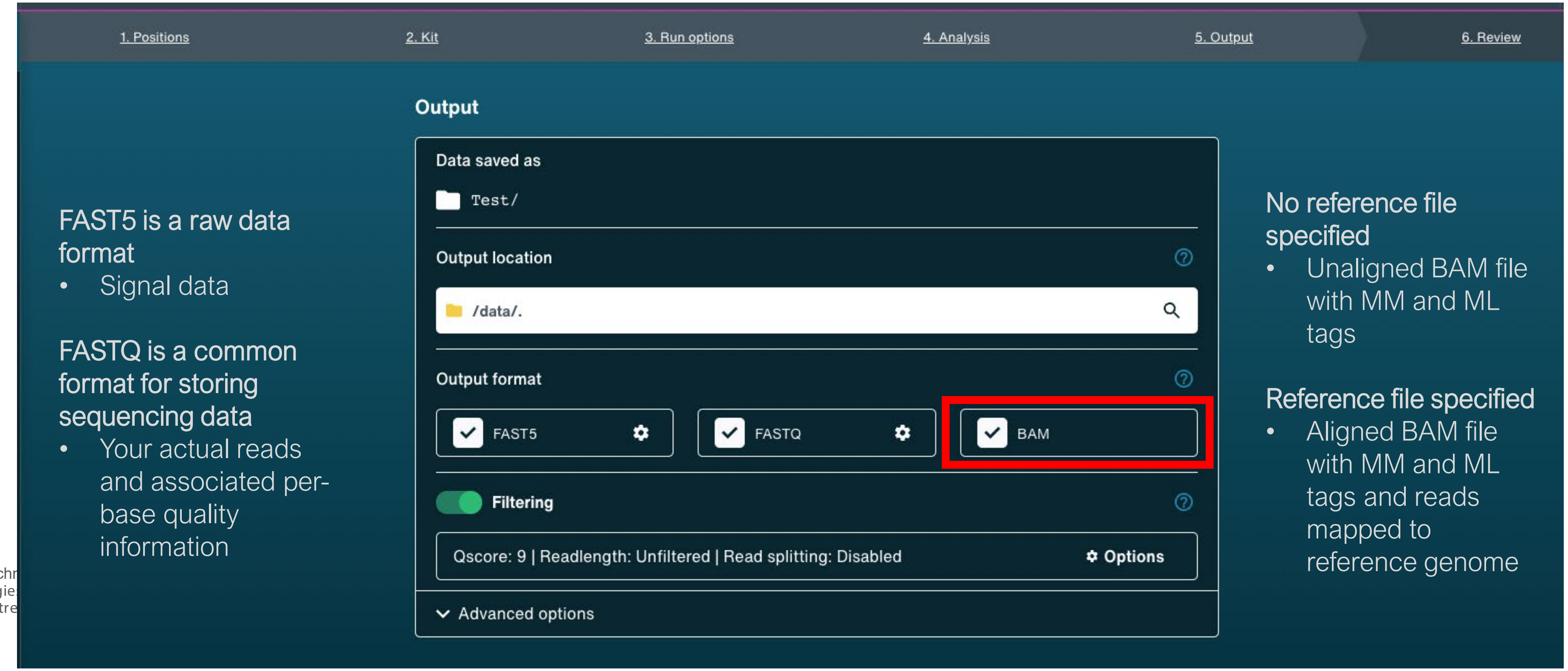


Methylation detection in MinKNOW

Ensure BAM output is enabled to store methylation information

BAM can be used to store aligned or unaligned reads (referred to as a uBAM)

- BAM or uBAM required by some EPI2ME Labs workflows (e.g. wf-human-variation)
- Methylation information described in the MM and ML tag of the BAM file



Generating an alignment

Required first step for most analyses

- **For a sequencing read, to determine the location of origin within the reference sequence**
- Reference can be complete genome, transcriptome or de novo assembly
- Input is typically FASTQ and reference genome in FASTA format
- Output is sequence alignment map (SAM) format file or binary equivalent (BAM)

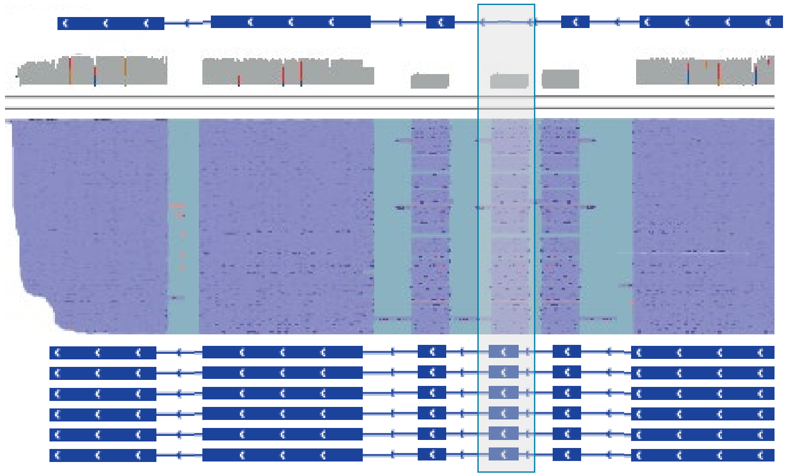
Genomics

SNPs/indels

Structural variation

Repeat content

…and more



Transcriptomics

Expression

Isoforms

Variation

…and more

Where to find reference sequences?

Public databases

Ensembl

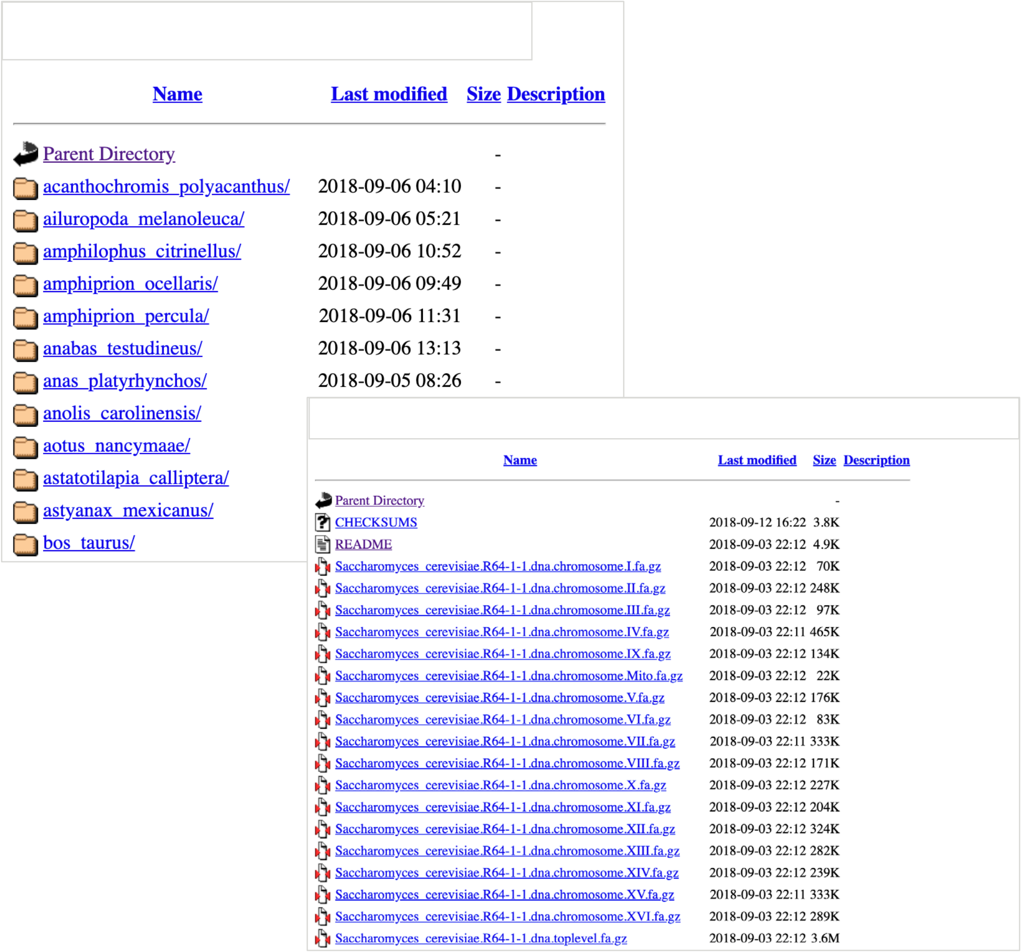
- Database of sequences and corresponding annotations
- ensemblgenomes.org

UCSC

- Database of sequences and corresponding annotations
- Genome Browser
- ucscbrowser.genenetwork.org

RefSeq

- Database of sequences
- ncbi.nlm.nih.gov/refseq/



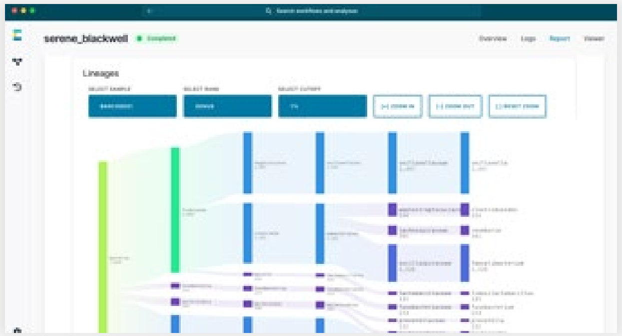
Name	Last modified	Size	Description
Parent Directory		-	
acanthochromis_polyacanthus/	2018-09-06 04:10	-	
ailuropoda_melanoleuca/	2018-09-06 05:21	-	
amphilophus_citrinellus/	2018-09-06 10:52	-	
amphiprion_ocellaris/	2018-09-06 09:49	-	
amphiprion_percula/	2018-09-06 11:31	-	
anabas_testudineus/	2018-09-06 13:13	-	
anas_platyrhynchos/	2018-09-05 08:26	-	
anolis_carolinensis/			
aotus_nancymaae/			
astatotilapia_calliptera/			
astyanax_mexicanus/			
bos_taurus/			

Name	Last modified	Size	Description
Parent Directory		-	
CHECKSUMS	2018-09-12 16:22	3.8K	
README	2018-09-03 22:12	4.9K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.I.fa.gz	2018-09-03 22:12	70K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.II.fa.gz	2018-09-03 22:12	248K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.III.fa.gz	2018-09-03 22:12	97K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.IV.fa.gz	2018-09-03 22:11	465K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.IX.fa.gz	2018-09-03 22:12	134K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.Mito.fa.gz	2018-09-03 22:12	22K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.V.fa.gz	2018-09-03 22:12	176K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VI.fa.gz	2018-09-03 22:12	83K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VII.fa.gz	2018-09-03 22:11	333K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VIII.fa.gz	2018-09-03 22:12	171K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.X.fa.gz	2018-09-03 22:12	227K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XI.fa.gz	2018-09-03 22:12	204K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XII.fa.gz	2018-09-03 22:12	324K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XIII.fa.gz	2018-09-03 22:12	282K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XIV.fa.gz	2018-09-03 22:12	239K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XV.fa.gz	2018-09-03 22:11	333K	
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XVI.fa.gz	2018-09-03 22:12	289K	
Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz	2018-09-03 22:12	3.6M	

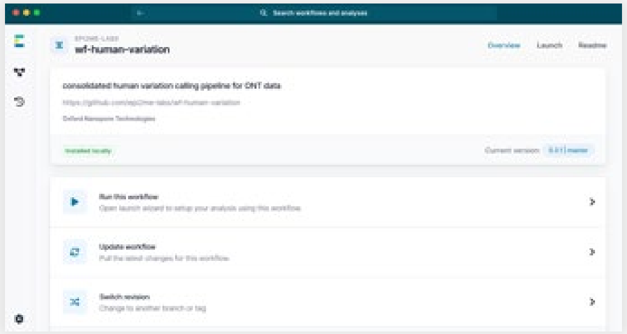
EPI2ME

Fully supported intuitive solutions

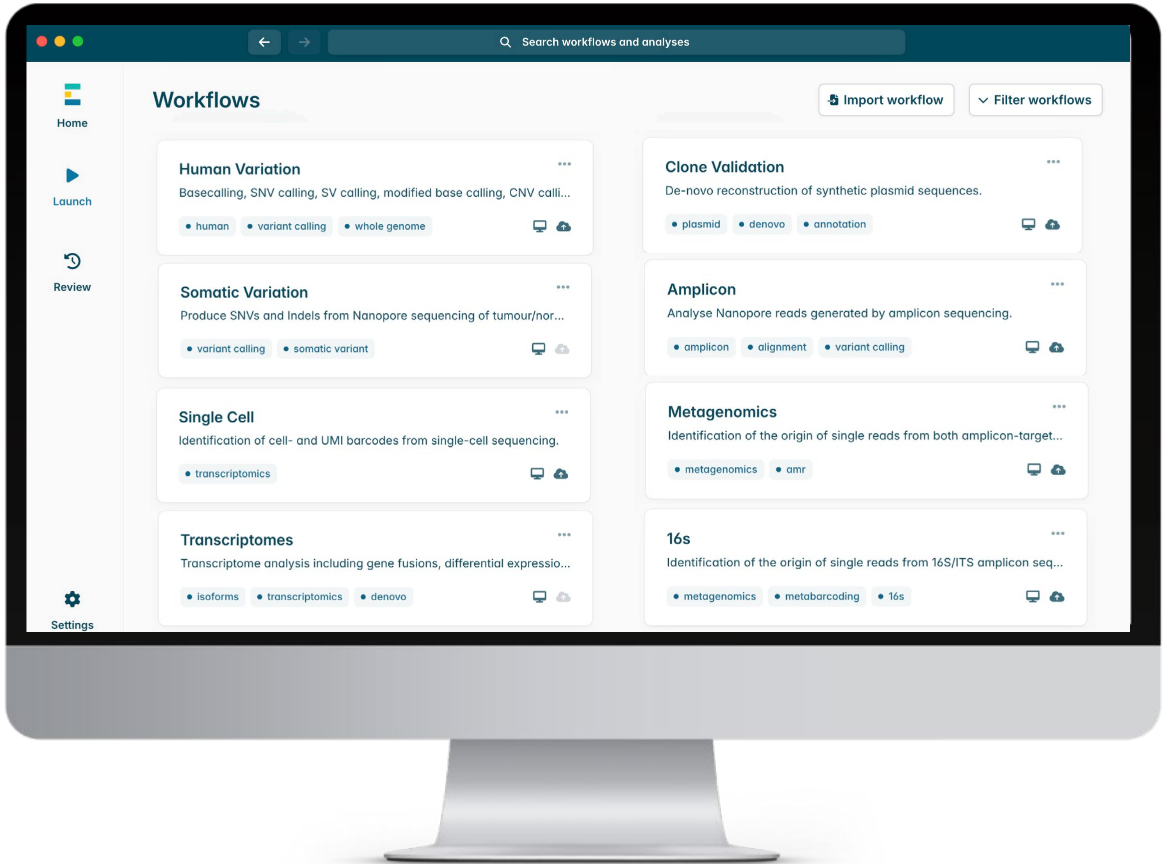
Analyse your data with EPI2ME



- Intuitive interface
- Pre-configured workflows
- Intuitive, interactive reports
- Standard output files

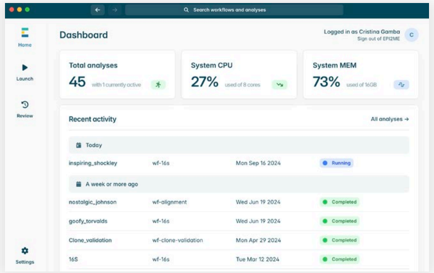


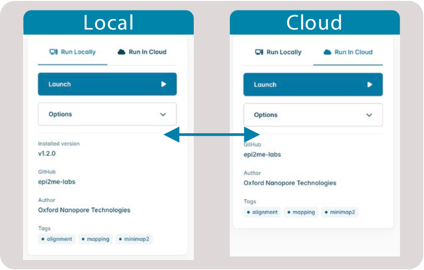
- Run locally or in the cloud
- Command-line access
- Easy integration

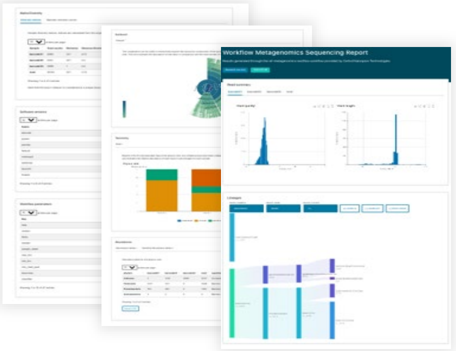


EPI2ME: how does it work?

Intuitive data analysis







DNA / RNA is sequenced and data is written to files locally

Analyse with EPI2ME desktop application or run EPI2ME workflows from the command-line

Choose between running the analysis locally or in the cloud

Analysis runs locally

Report and output files are generated

Data is uploaded and analysis run in the cloud

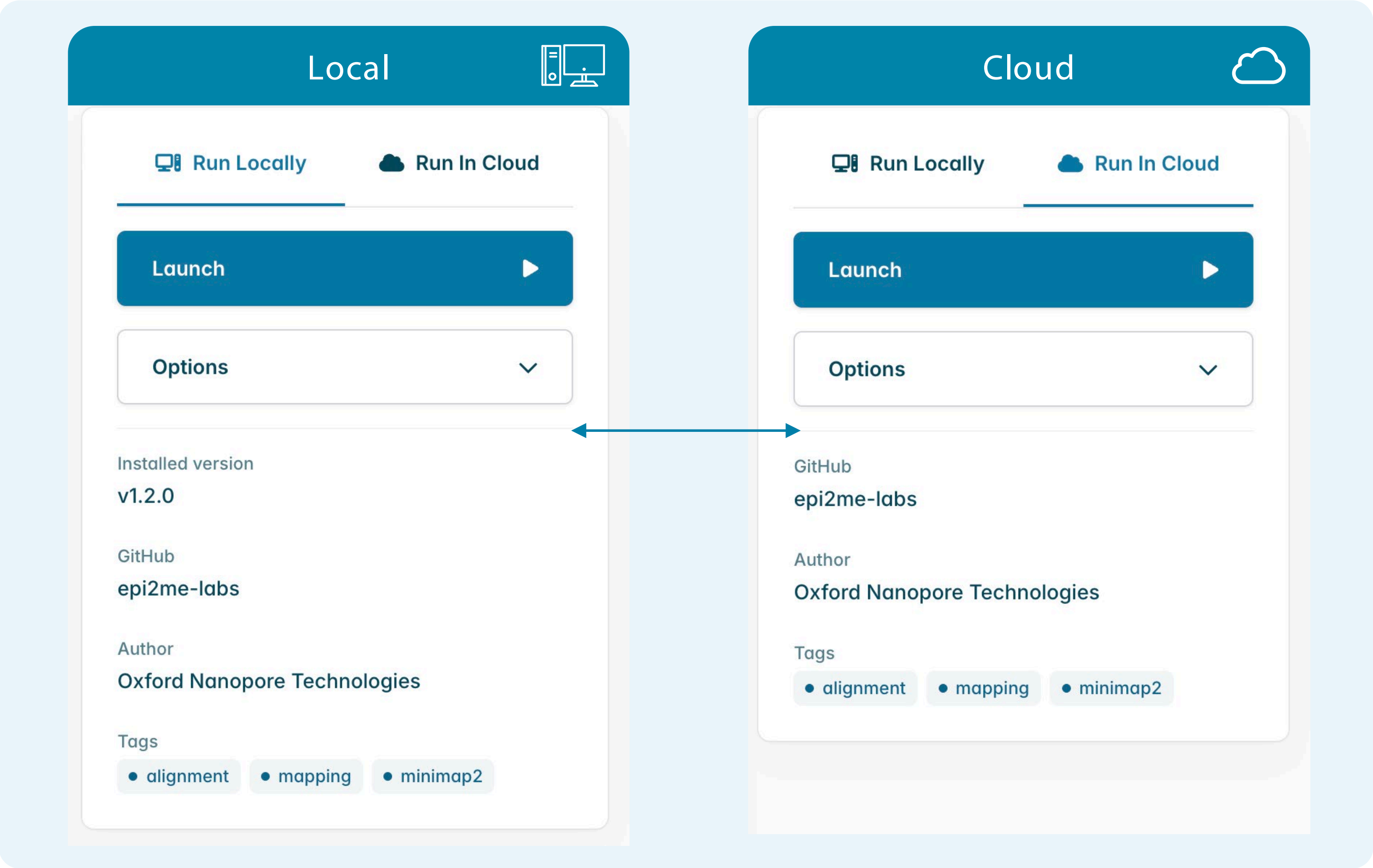
Report and output files are generated – results are downloaded to local system*

* Data is not stored in the cloud and need to be synced to local systems. Data in the cloud is automatically removed after 2 weeks from the end of the analysis.

Integrated local and cloud EPI2ME experience

Local and cloud analysis in a single software

- Location
- Access
- Configurability
- Reporting
- Focus
- Operating systems
- Bioinformatics expertise



Integrated cloud functionality
for selected workflows

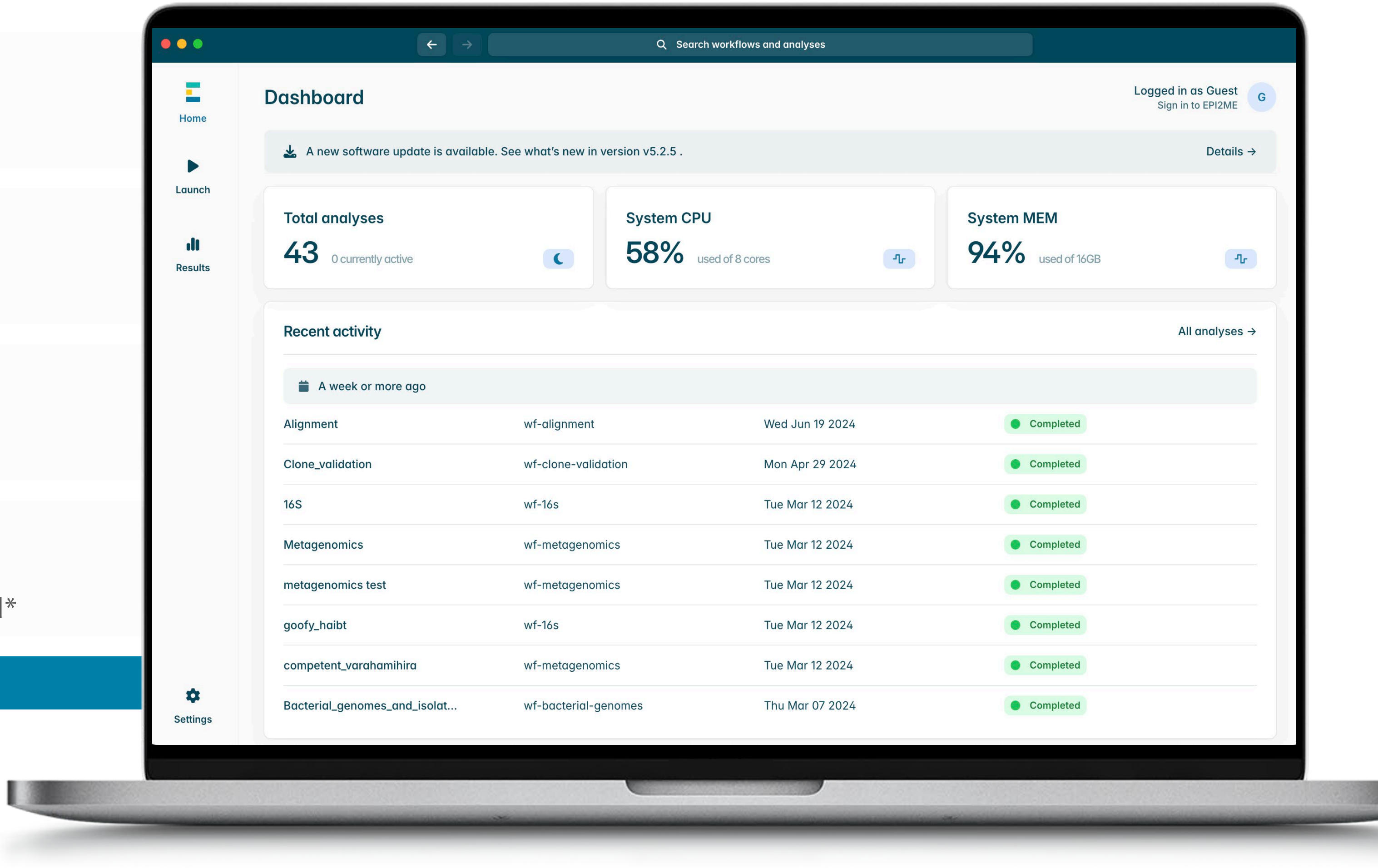
[non-human data]

EPI2ME desktop application: how does it work?

An open analysis platform

- Free to download standalone desktop application
- Pre-packaged with >15 workflows
- Minimal setup required
- Compatibility
  - Windows, macOS, Linux
  - Desktop computers, laptops, on-device, cluster or cloud*

*Must be logged in to access cloud functionality



Oxford Nanopore Technologies

Workflows available in EPI2ME

Human genetics
- wf-trio ●●● NEW!
- wf-human-variation ●●●
- wf-somatic-variation ●●●

Basic tasks
- wf-alignment ●● (cloud)
- wf-basecalling (large GPU required)

What's in my pot?
- wf-metagenomics ●● (cloud)
- wf-16S ● (cloud)

Quality control
- wf-clone-validation ● (cloud)
- wf-aav-qc ● (cloud)

Transcriptome sequencing
- wf-transcriptomes ●● (cloud)
- wf-single-cell ●●● (cloud)

Targeted sequencing
- wf-amplicon ● (cloud)

Infectious disease
- wf-bacterial-genomes ●● (cloud)
- wf-flu ● (cloud)

Genome structure
- wf-teloseq ●● NEW!
- wf-pore-c ●●●

(cloud) = cloud analysis available

(large GPU required) = large GPU required

CPU requirements
- ● standard laptop 8 GB RAM -- Grid / P2i / P24 (/ Laptop)
- ●● large memory computer 32 GB RAM –- Grid / P2i / P24 (/ Laptop)
- ●●● 16 CPUs + 64GB RAM required --- Prom24

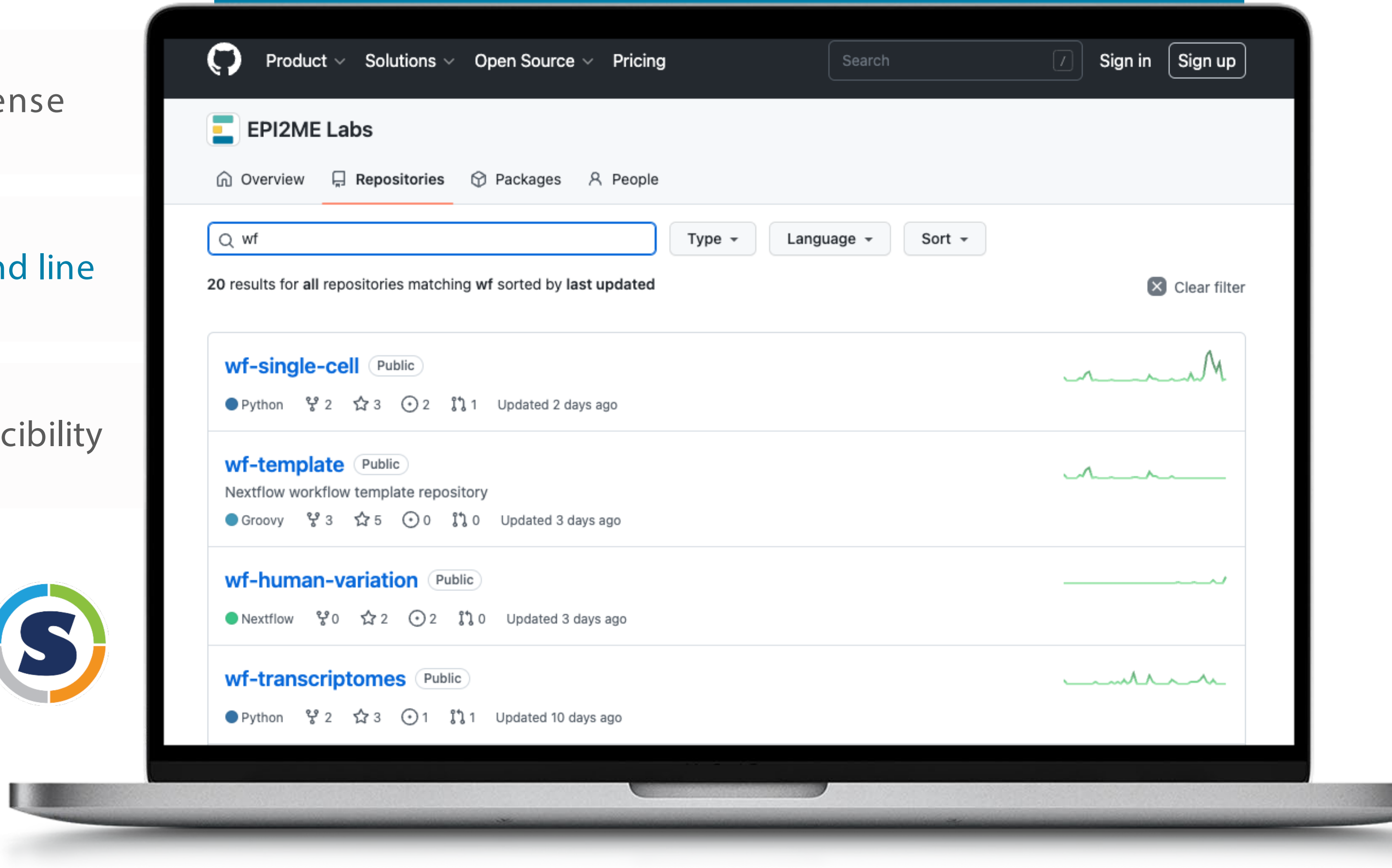
Interested in running **wf-human-variation** in the cloud? Register your interest here

EPI2ME workflows from command line

For deployment on own clusters or custom pipelines

Nextflow workflows available on GitHub

- EPI2ME code is deposited in GitHub under MPL-2.0 license
- Nextflow workflows can be executed from the command line
- Support containers Singularity and Docker for reproducibility



More at: github.com/orgs/epi2me-labs/repositories

EPI2ME integrations

Workflows now easier to install

Enabled by the 2ME format

Supports offline compute infrastructure

2ME format includes all workflow pre-requisites

One-click installation

Workflow bundle

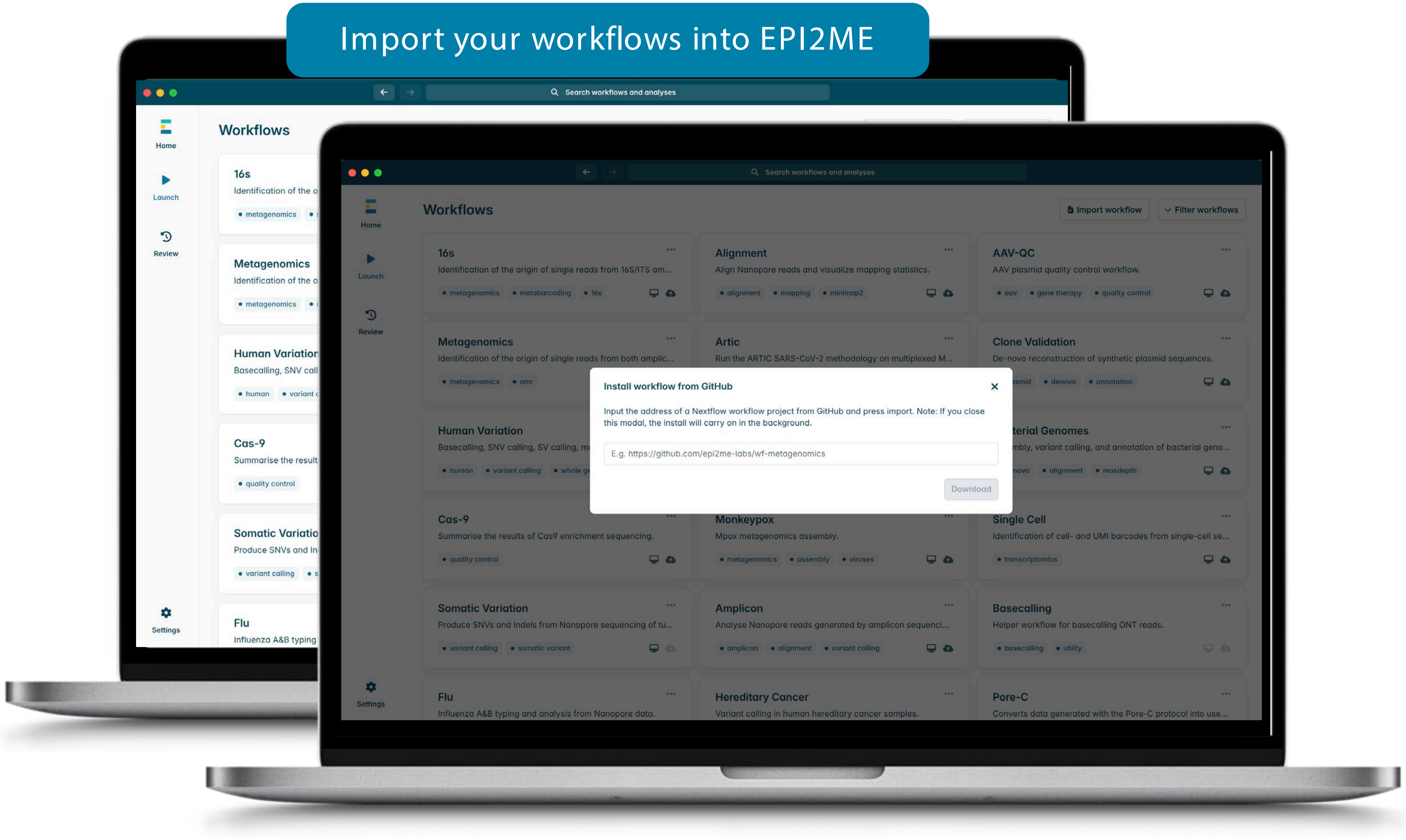
Code	Containers	Data
Reference data	Instances	

Oxford Nanopore Technologies

Integrate your workflows with EPI2ME desktop application

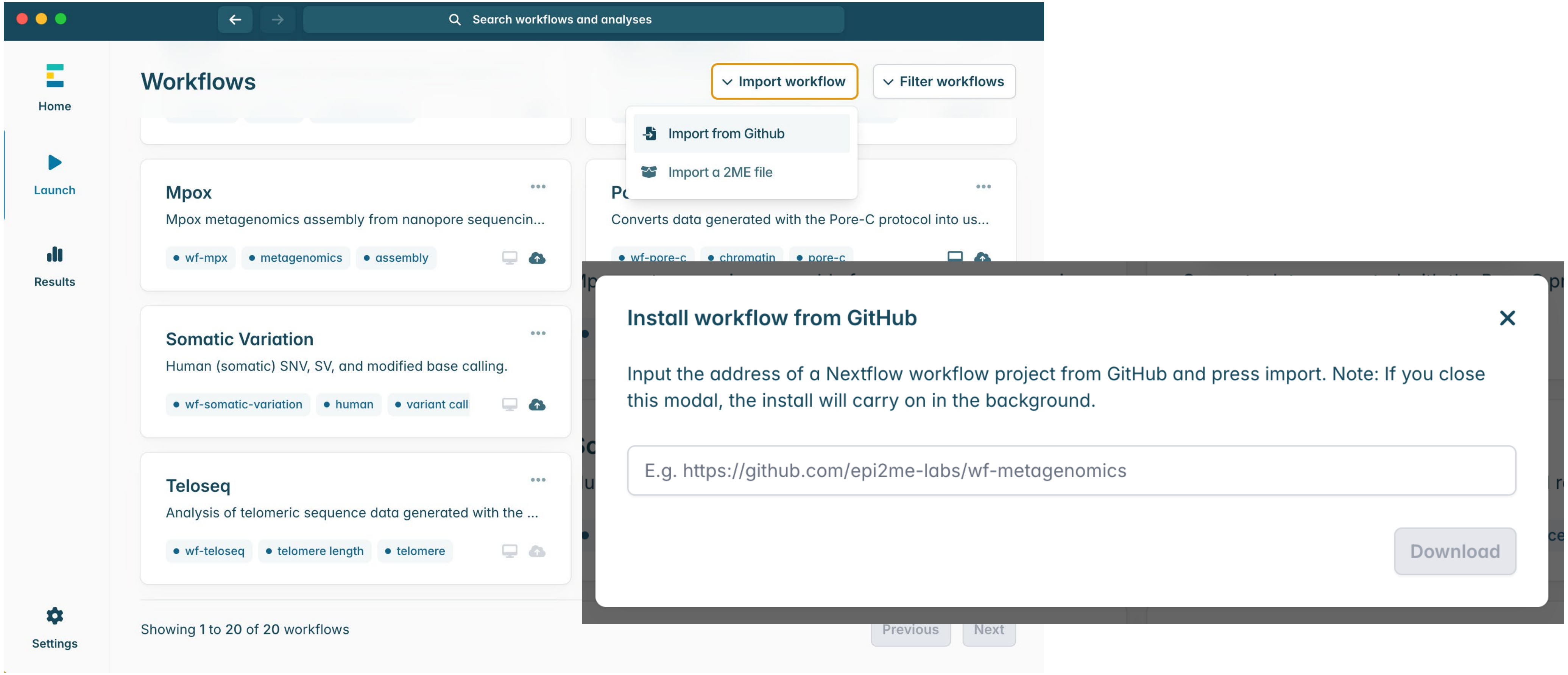
Install with the GitHub repository

- EPI2ME desktop application enables
  - Running locally community-developed bioinformatics workflows
  - Implemented in Nextflow as per nf-core standards
- Import straight from GitHub
- Share your workflows with the broader Oxford Nanopore community



Integrate your workflows with EPI2ME desktop application

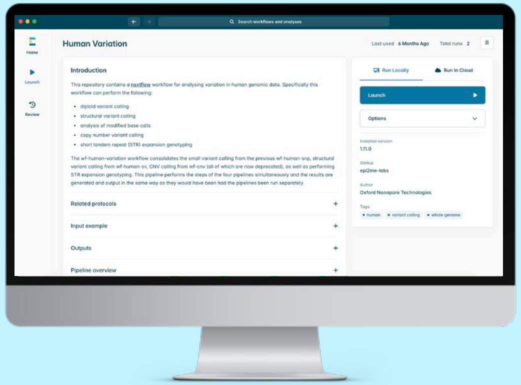
Install with the GitHub repository



Enabling nanopore data interpretation

Through tertiary analysis compatible partners

EPI2ME
Human variation workflow



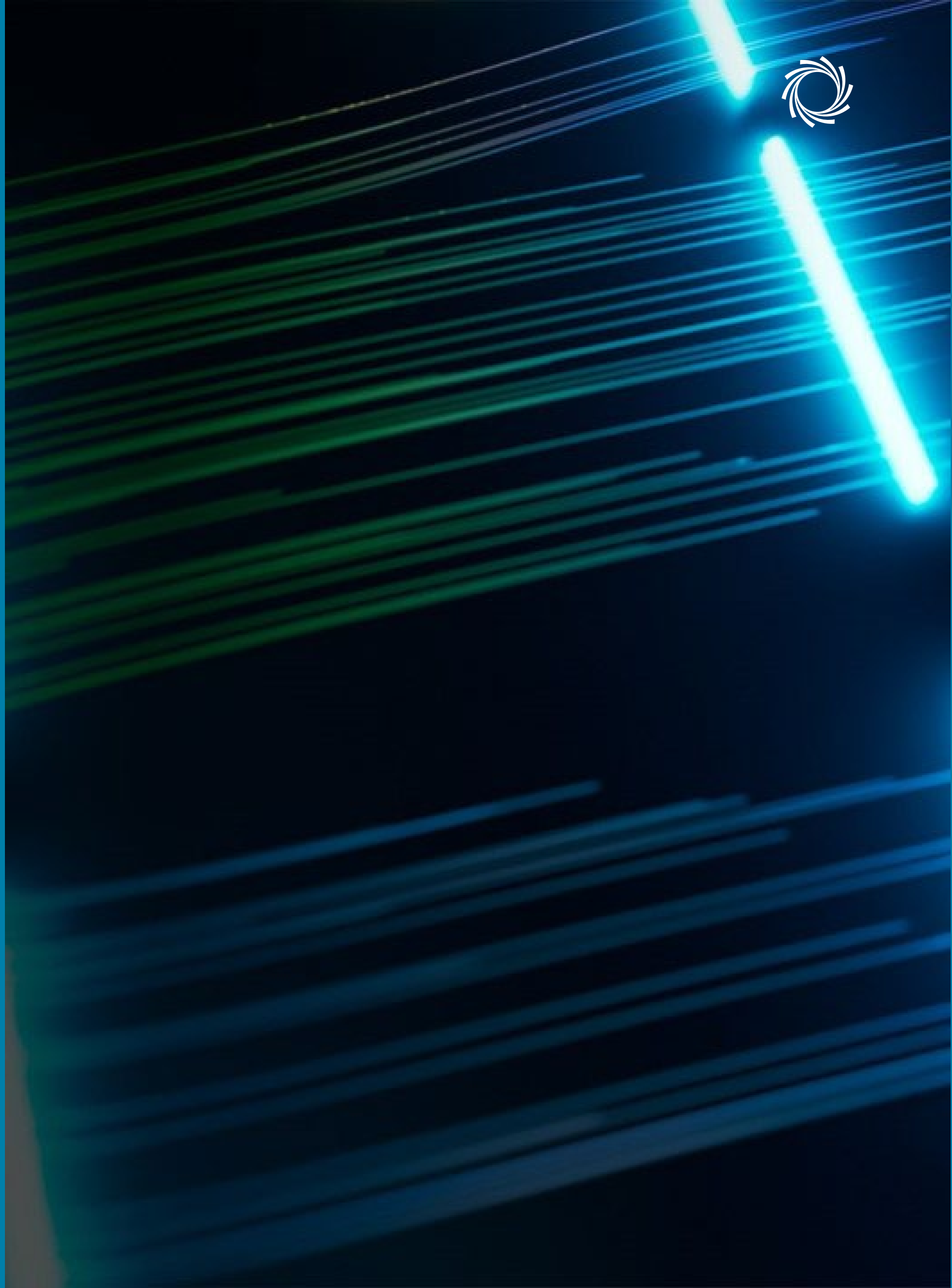
Secondary analysis

Compatible partners

Tertiary analysis and interpretation

Oxford Nanopore Technologies

Datasets



An increasing library of datasets

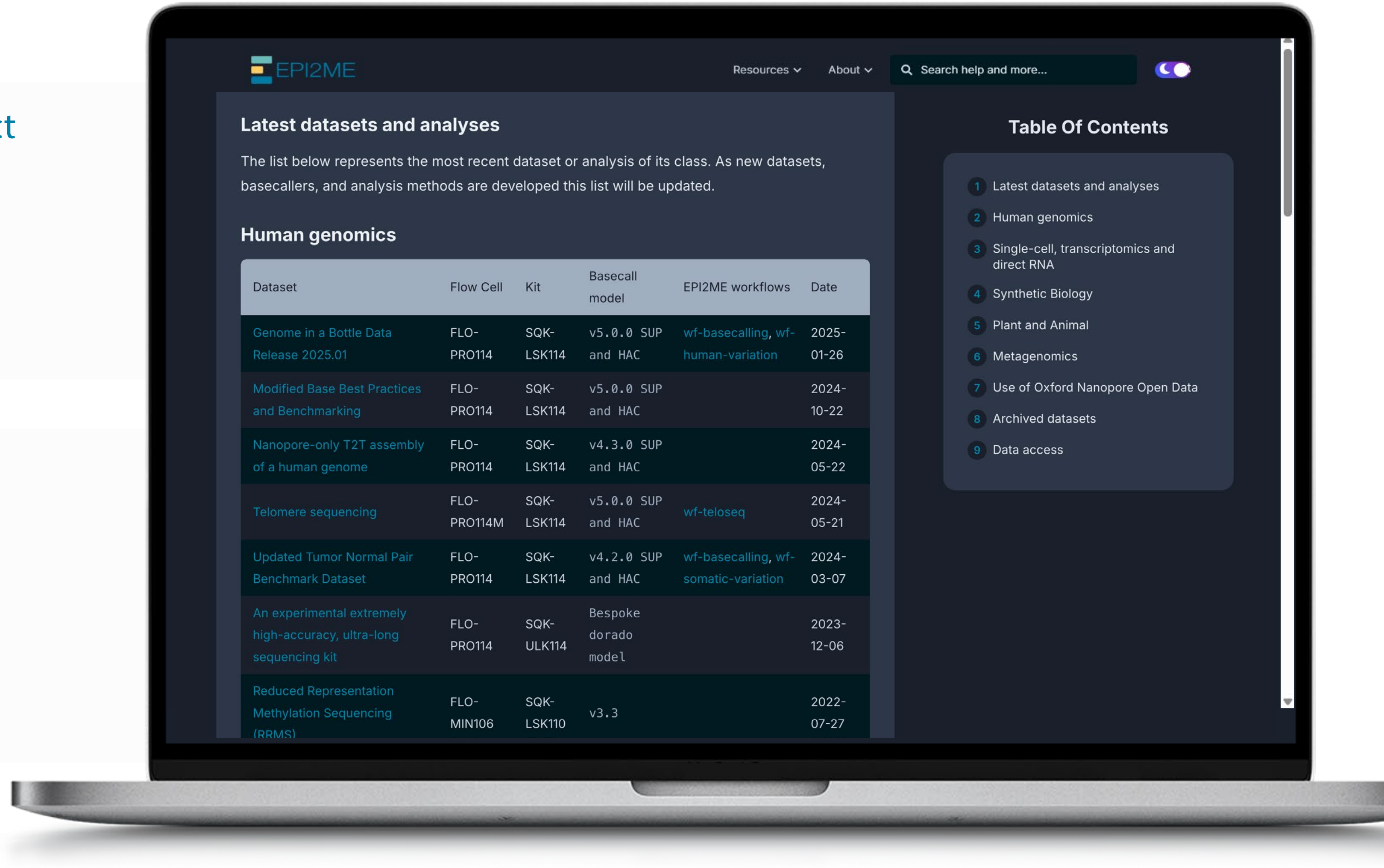
Oxford Nanopore open dataset project

- Test EPI2ME
- Tool development
- Reproducible benchmarks
- Explore characteristics of nanopore data

Over 20 datasets available e.g.

- Genome in a Bottle
- Plasmid Validation
- Metagenomics
- And much more

https://epi2me.nanoporetech.com/dataindex/



EPI2ME Resources About Search help and more...

Latest datasets and analyses

The list below represents the most recent dataset or analysis of its class. As new datasets, basecallers, and analysis methods are developed this list will be updated.

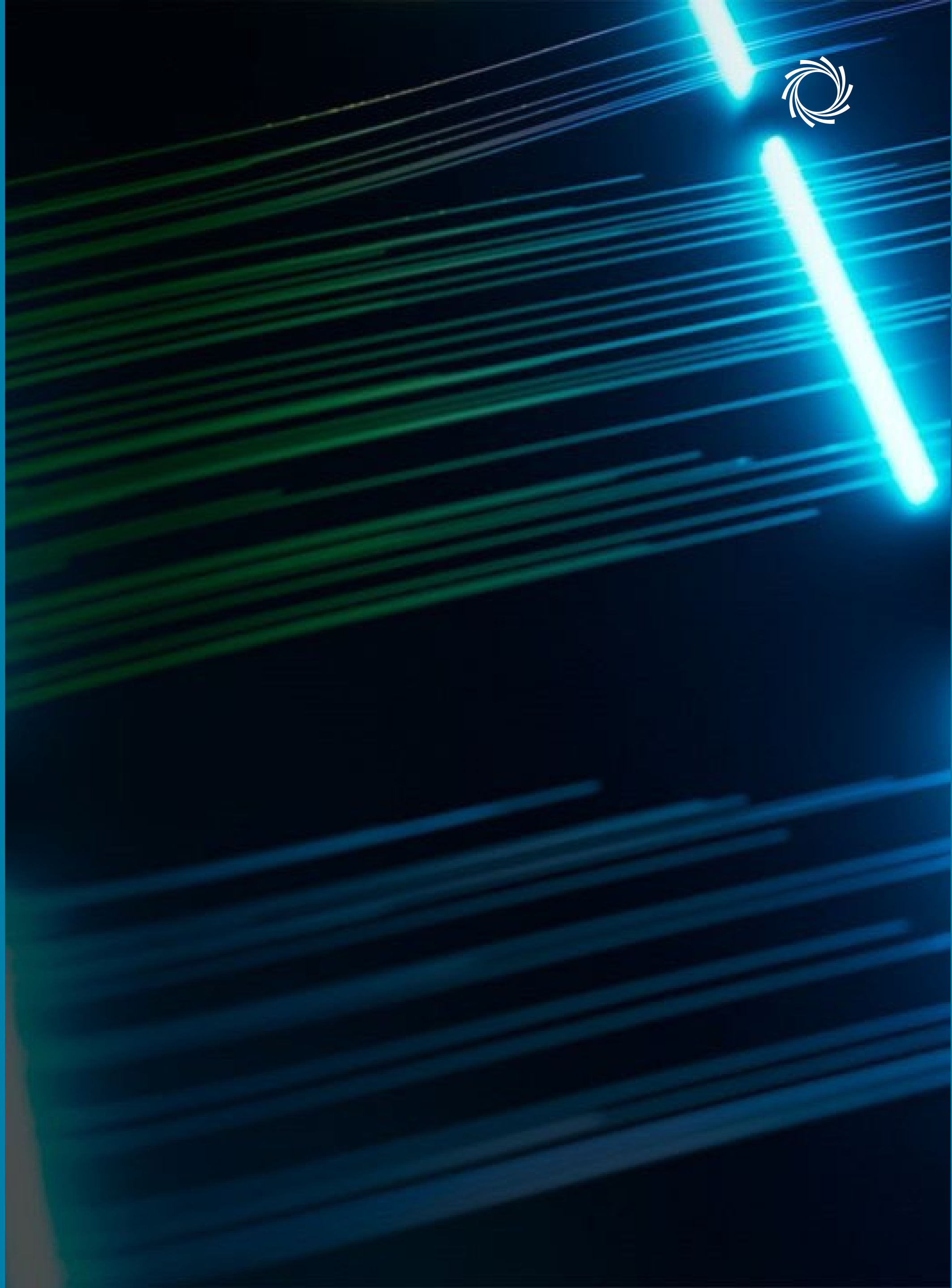
Human genomics

Dataset	Flow Cell	Kit	Basecall model	EPI2ME workflows	Date
Genome in a Bottle Data Release 2025.01	FLO-PRO114	SQK-LSK114	v5.0.0 SUP and HAC	wf-basecalling, wf-human-variation	2025-01-26
Modified Base Best Practices and Benchmarking	FLO-PRO114	SQK-LSK114	v5.0.0 SUP and HAC		2024-10-22
Nanopore-only T2T assembly of a human genome	FLO-PRO114	SQK-LSK114	v4.3.0 SUP and HAC		2024-05-22
Telomere sequencing	FLO-PRO114M	SQK-LSK114	v5.0.0 SUP and HAC	wf-teloseq	2024-05-21
Updated Tumor Normal Pair Benchmark Dataset	FLO-PRO114	SQK-LSK114	v4.2.0 SUP and HAC	wf-basecalling, wf-somatic-variation	2024-03-07
An experimental extremely high-accuracy, ultra-long sequencing kit	FLO-PRO114	SQK-ULK114	Bespoke dorado model		2023-12-06
Reduced Representation Methylation Sequencing (RRMS)	FLO-MIN106	SQK-LSK110	v3.3		2022-07-27

Table Of Contents

EPI2ME workflows

- Examples of streamlined analysis workflows
- Basecalling and mapping to a reference



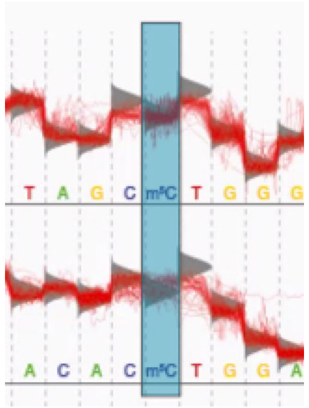
Oxford Nanopore Technologies

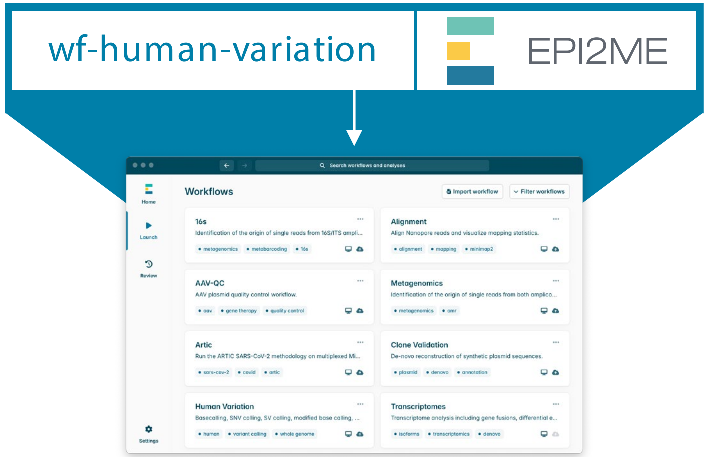
Human genetics

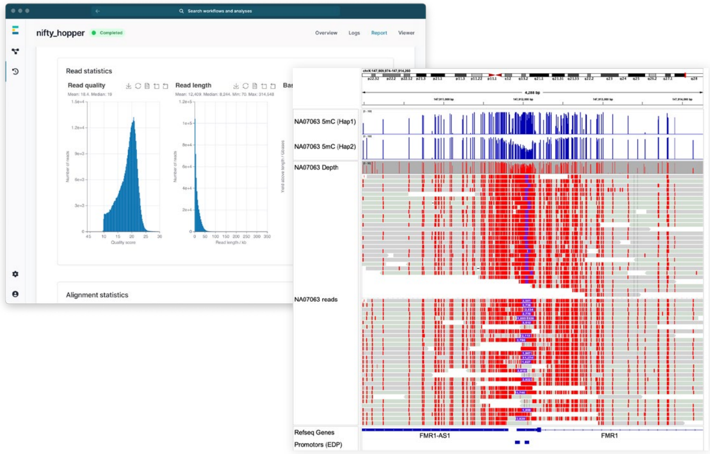
Investigate human variation with a single workflow

Investigation of structural variants, single nucleotide variants, and methylation using EPI2ME

5mC







Sample sequencing and basecalling in MinKNOW

wf-human-variation

Detect

SNVs, SVs, STR, CNVs, methylation, phasing

Intuitive reports

Visualise results

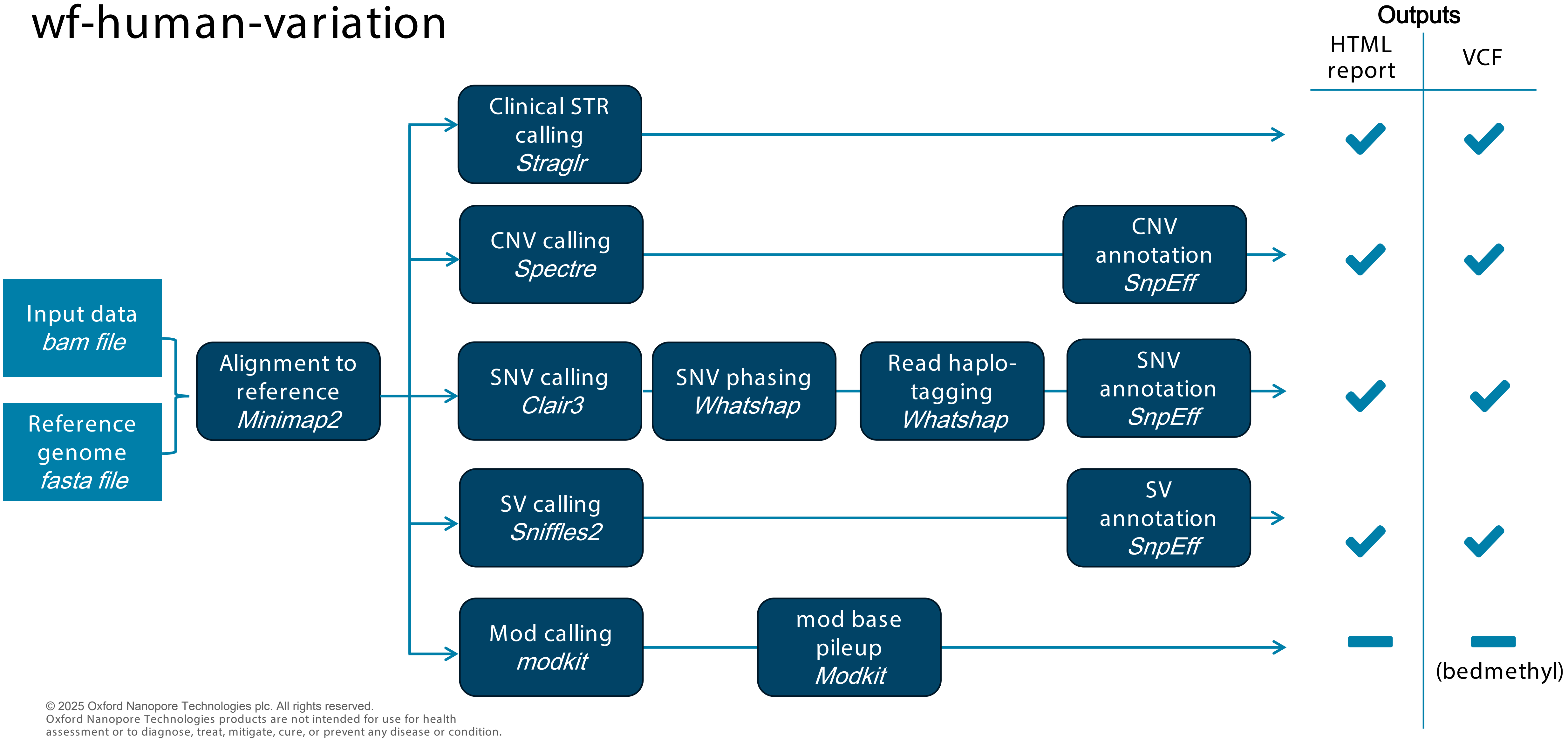
MinKNOW

EPI2ME

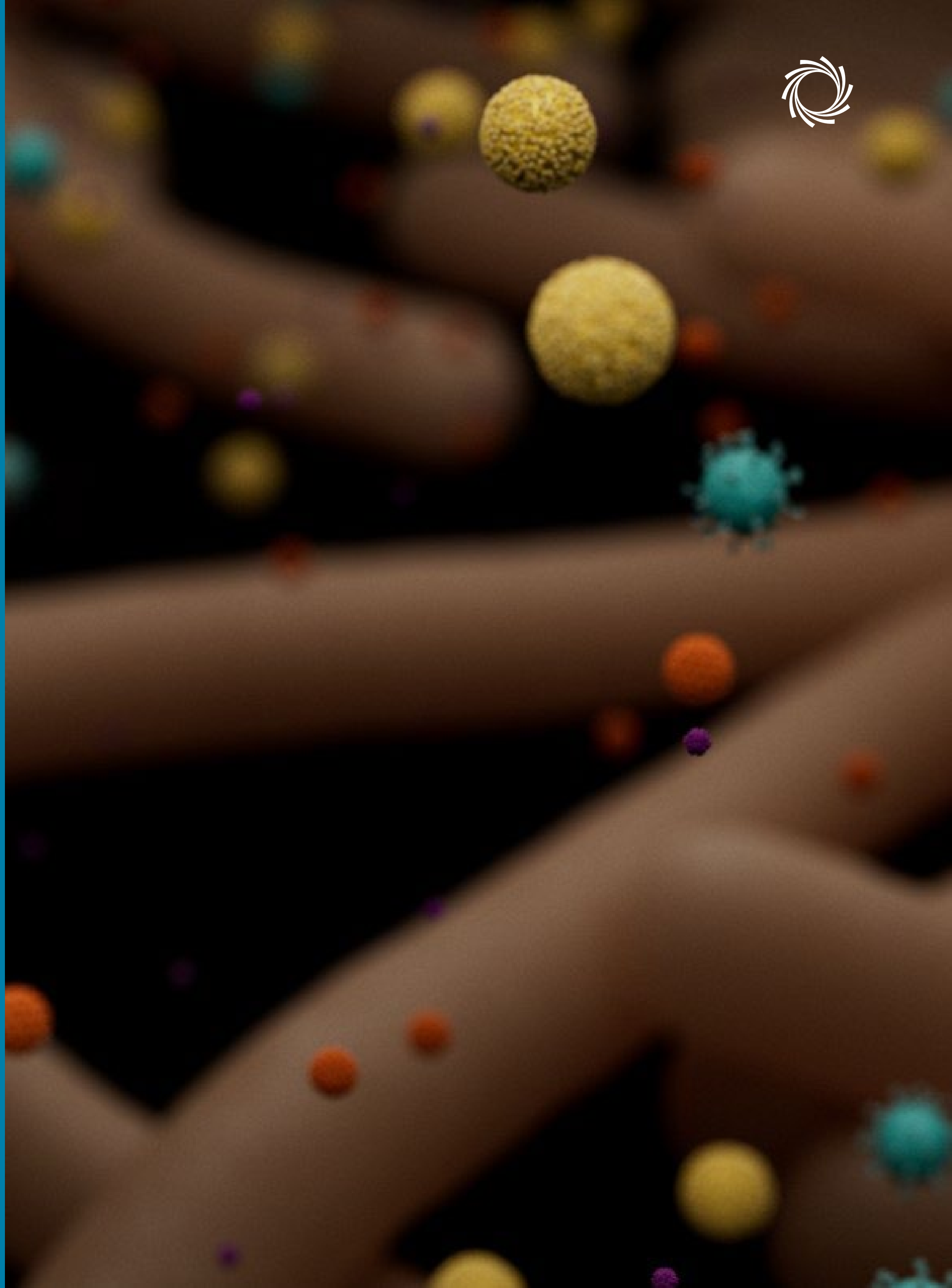
This workflow is accessible from both the intuitive graphical interface and the command line

Interested in running **wf-human-variation** in the cloud? Register your interest here

wf-human-variation



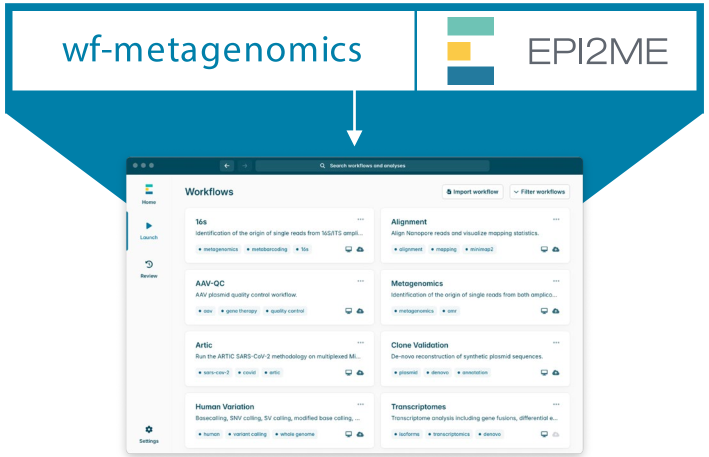
What's in my pot?

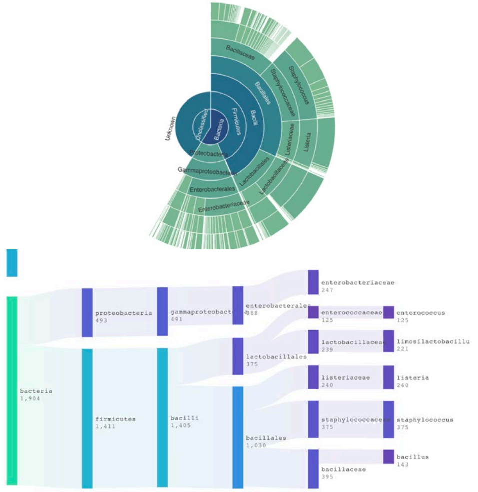


Get insights into your metagenomic samples: shotgun or amplicons

Intuitive visualisation of taxonomy, diversity, abundances, and more

Amplicons or Native DNA





Sequence and basecall amplicon-targeted and shotgun metagenomics samples in MinKNOW

wf-metagenomics

Quick classification with Kraken 2. Fine classification with minimap2

Real-time options

Intuitive report, classified and unclassified reads, and text files with lineage details

MinKNOW

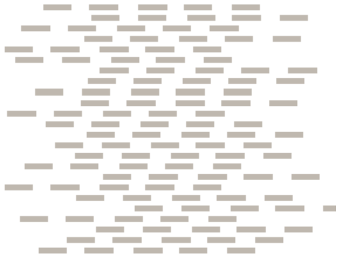
EPI2ME

This workflow is accessible from both the intuitive graphical interface and the command line

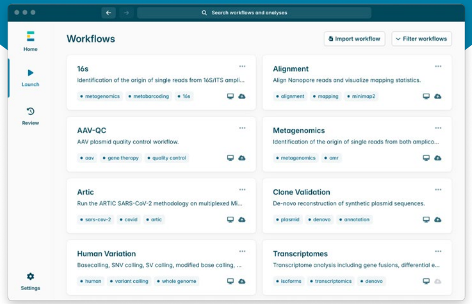
Identification based on 16S, 18S, and ITS amplicons

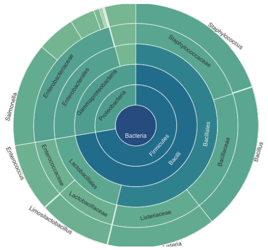
Intuitive visualisation of taxonomy, diversity, abundances, and more

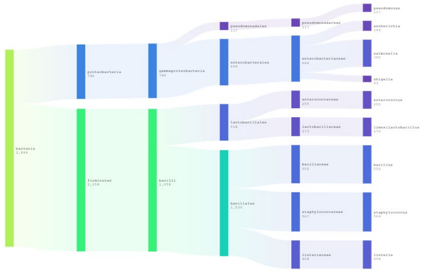
wf-16S | EPI2ME



16S / 18S / ITS Amplicons







Sequence and basecall amplicon-targeted 16S/18S/ITS in MinKNOW

wf-16S

Quick classification with Kraken 2. Fine classification with minimap2

Real-time options

Intuitive report, classified and unclassified reads, and text files with lineage details

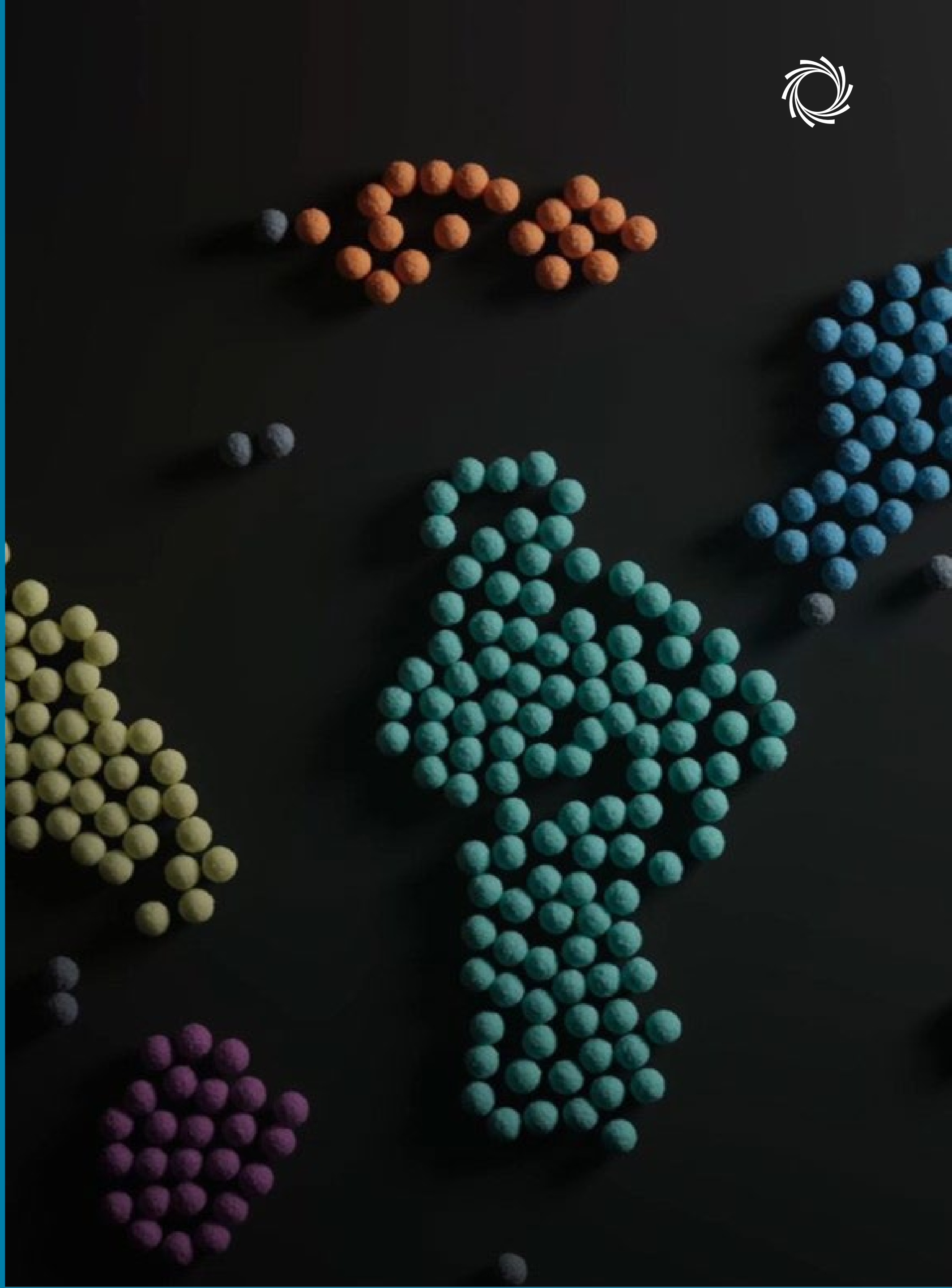
MinKNOW

EPI2ME

This workflow is accessible from both the intuitive graphical interface and the command line

Oxford Nanopore Technologies

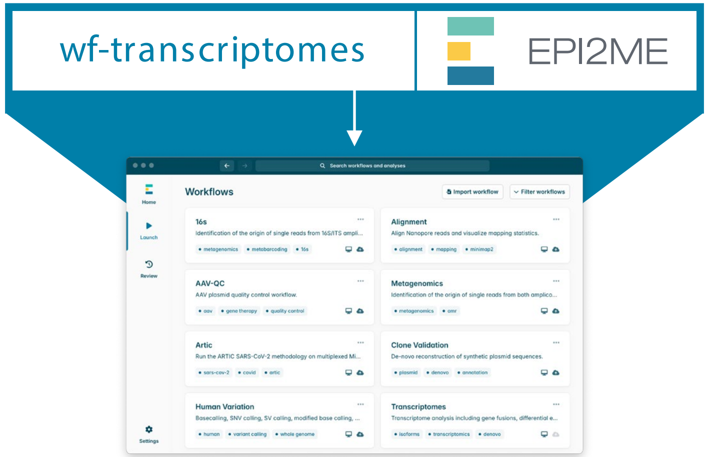
Transcriptome sequencing

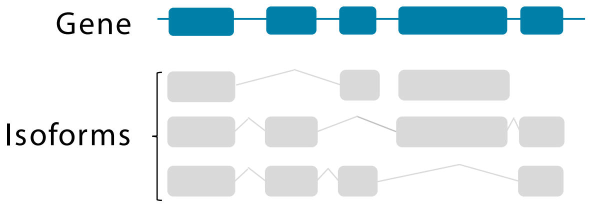


Transcriptomics workflow for cDNA or native RNA

Identify isoforms, measure gene expression, and detect fusion genes

Native RNA or cDNA reads





1. Full-length native RNA reads or cDNA transcripts sequenced and basecalled in MinKNOW
2. wf-transcriptomics
3. Input FASTQ file(s) and reference genome (if available). Additional reference files needed for gene fusion and gene expression*
4. Assembled transcriptome (annotated). Optional: fusion transcript sequences and differential transcript plots

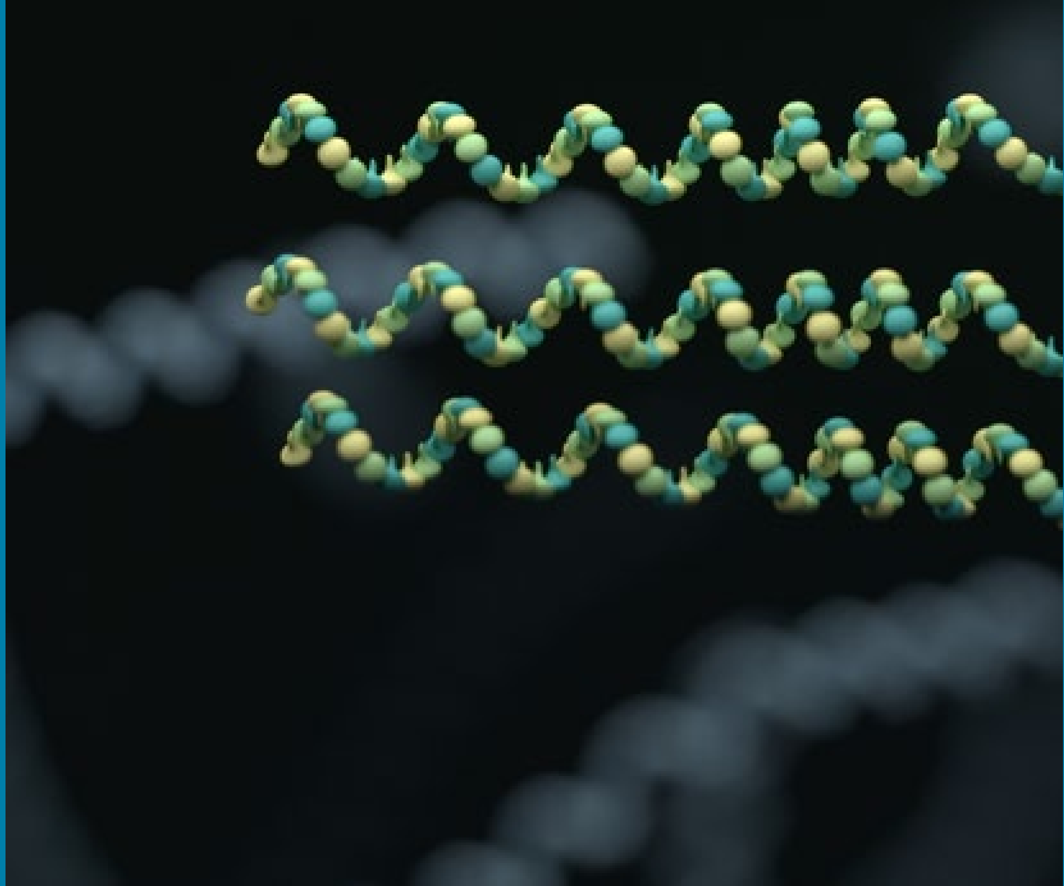
MinKNOW

EPI2ME

*For differential expression analysis provide reference annotation in GFF2/3 format, and for fusion detection, provide the JAFFAL reference files.
This workflow is accessible from both the intuitive graphical interface and the command line.

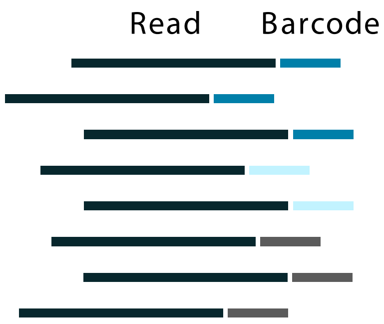
Oxford Nanopore Technologies

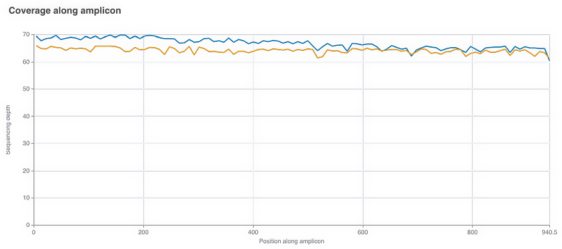
Targeted sequencing



Get high-quality amplicon consensus & detect variants

Barcode and pool samples & targets





Sample	Amplicon	Position	Ref. allele	Alt. allele	Type	Depth	Allelic balance
barcode01	katG__NC_000962.3_2154725-2155670	443	C	G	SNP	68	95.6%
barcode01	rpoB__NC_000962.3_760285-761376	870	C	T	SNP	79	94.9%
barcode02	katG__NC_000962.3_2154725-2155670	443	C	G	SNP	65	96.9%
barcode02	rpoB__NC_000962.3_760285-761376	870	C	T	SNP	60	98.3%

Sequence and basecall in MinKNOW

wf-amplicon

De novo assembly of consensus or variant calling against reference

Intuitive report

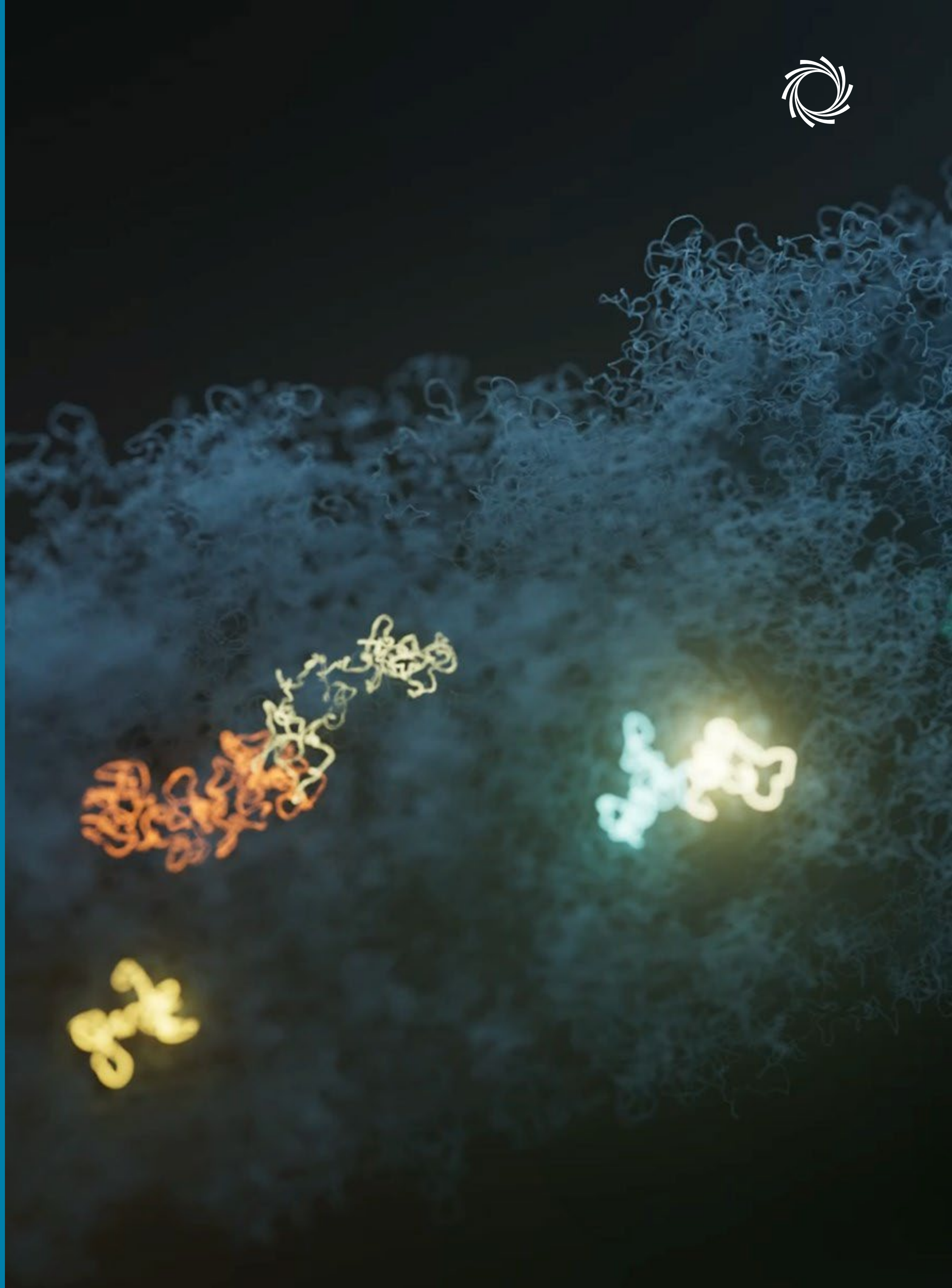
Includes pre-processing, QC, depth of coverage, etc.

MinKNOW

EPI2ME

This workflow is accessible from both the intuitive graphical interface and the command line

Infectious disease research

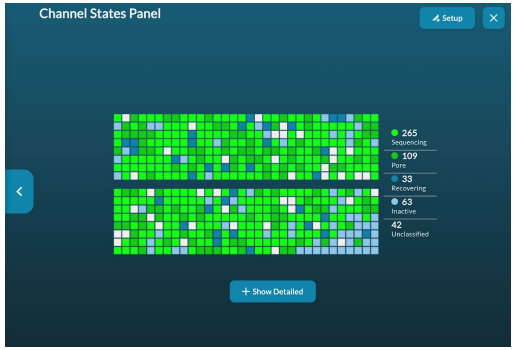


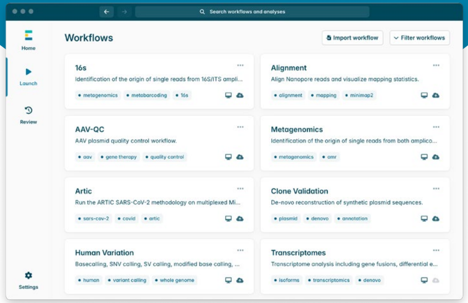
Bacterial genomes alignment or assembly

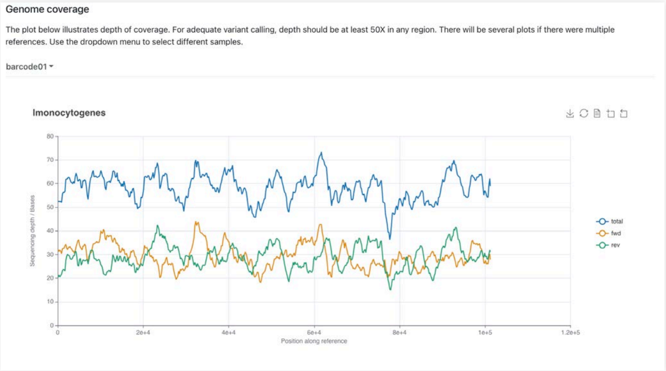
Including bacterial annotation & antimicrobial resistance genes

wf-bacterial-genomes

EPI2ME







Sample sequencing and basecalling in MinKNOW

wf-bacterial-genomes

Alignment to reference or assembly. Consensus annotation and resistance genes

Intuitive report, FASTA consensus and annotation

MinKNOW

EPI2ME

This workflow is accessible from both the intuitive graphical interface and the command line

Emerging tools for advanced users

- Open access software
- Bioinformatics expertise needed

The latest emerging tools from Oxford Nanopore

Open access software on GitHub

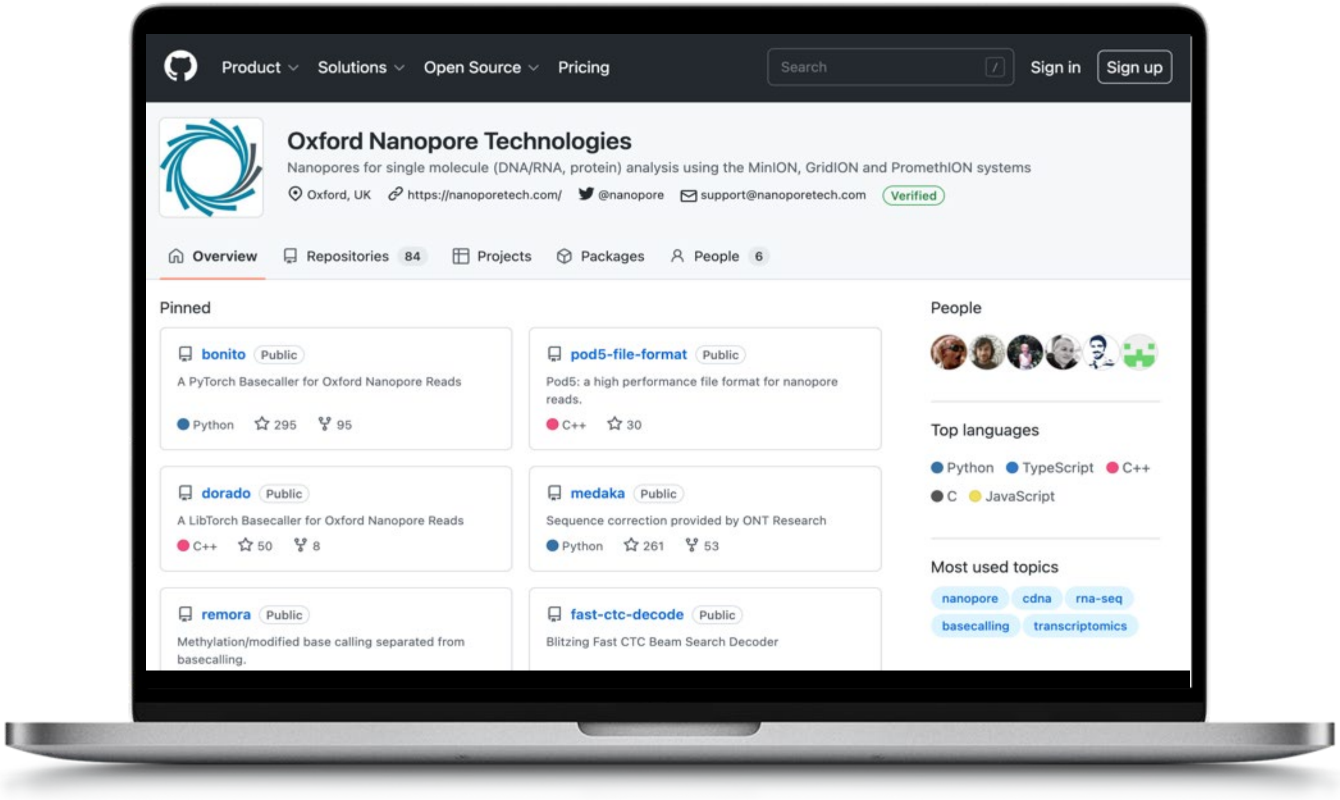
The latest tools and algorithms developed by Oxford Nanopore are available on the Oxford Nanopore GitHub repository

- From basecallers to modified base analysis
- Ready for your custom pipelines
- Command-line experience required
- Limited support due to rapid evolution

Key tools available

- Dorado: latest basecaller and other tools available standalone
- Modkit: toolbox for working with modified bases

Oxford Nanopore GitHub repository

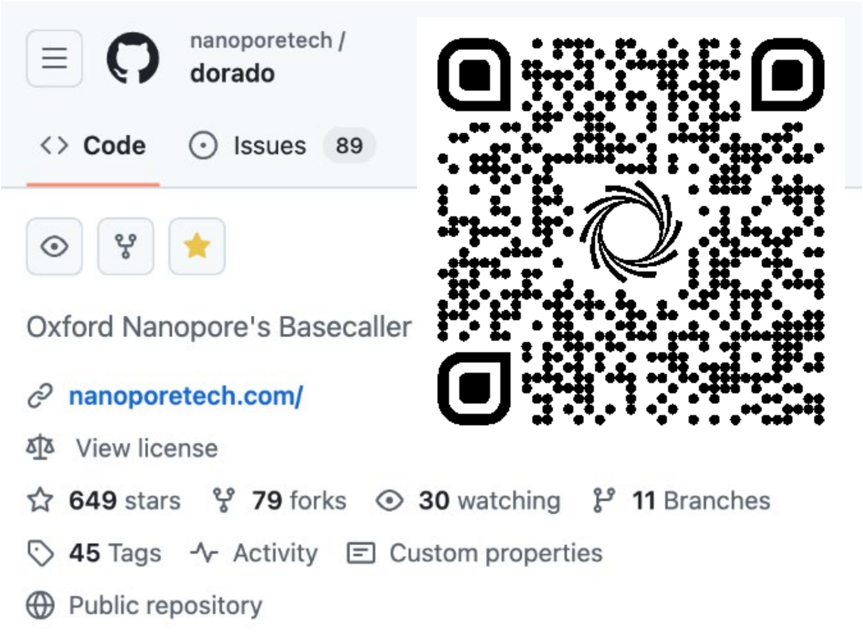


More at: github.com/nanoporetech

What is Dorado?

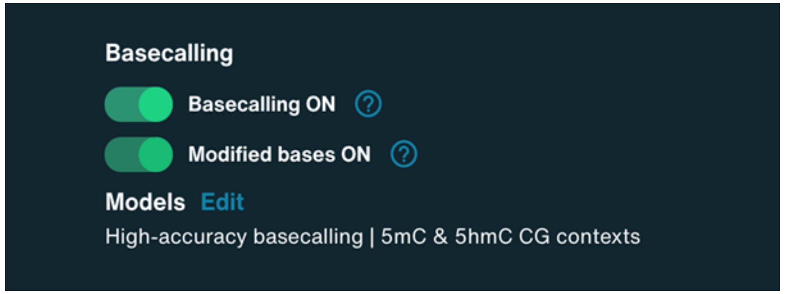
Dorado powers Oxford Nanopore basecalling and more

Latest features available on GitHub

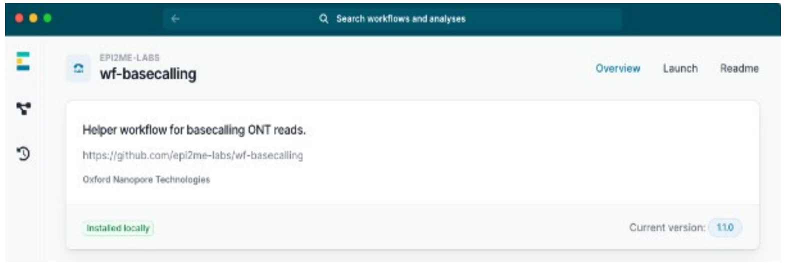


Integrated in MinKNOW and EPI2ME

MinKNOW



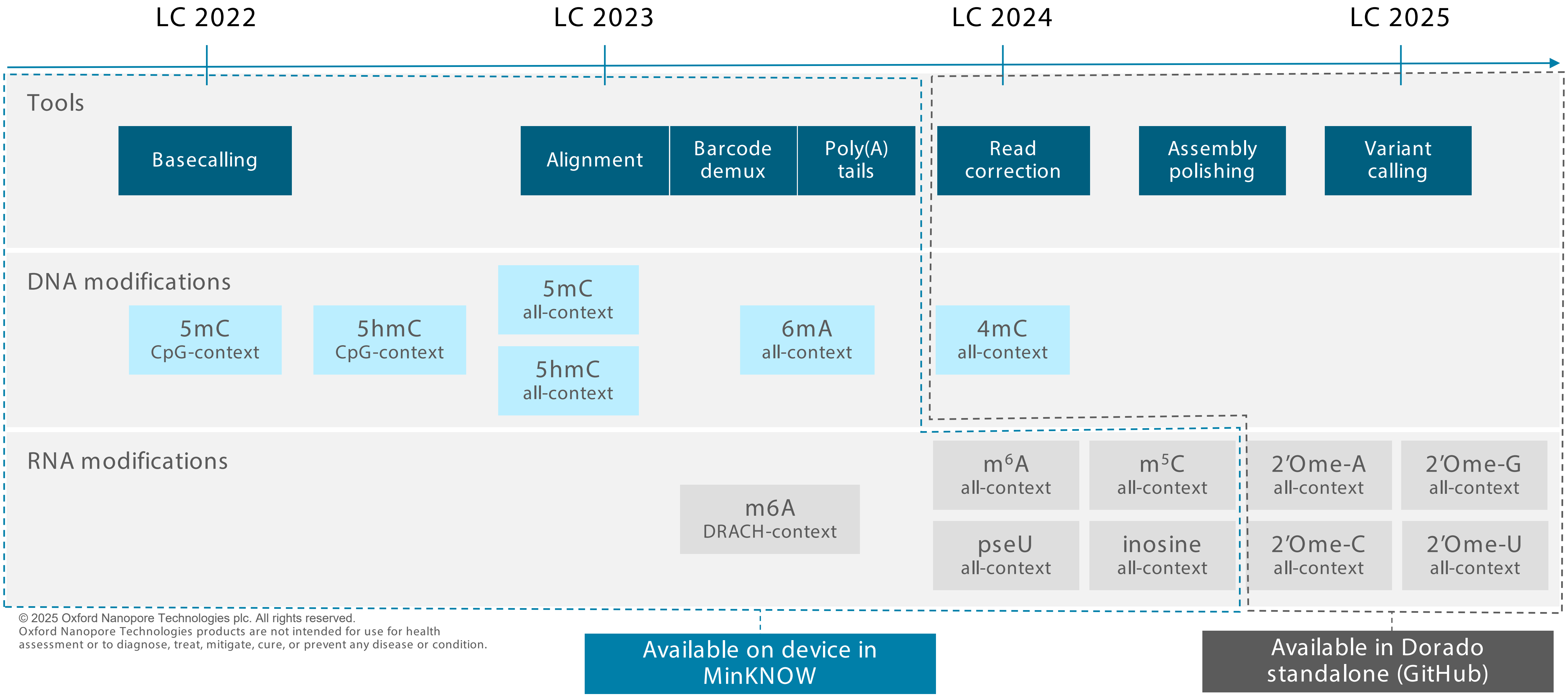
EPI2ME



Oxford Nanopore Technologies

Dorado: the engine for Oxford Nanopore data analysis

From basecalling through to assembly polishing and small variant calling

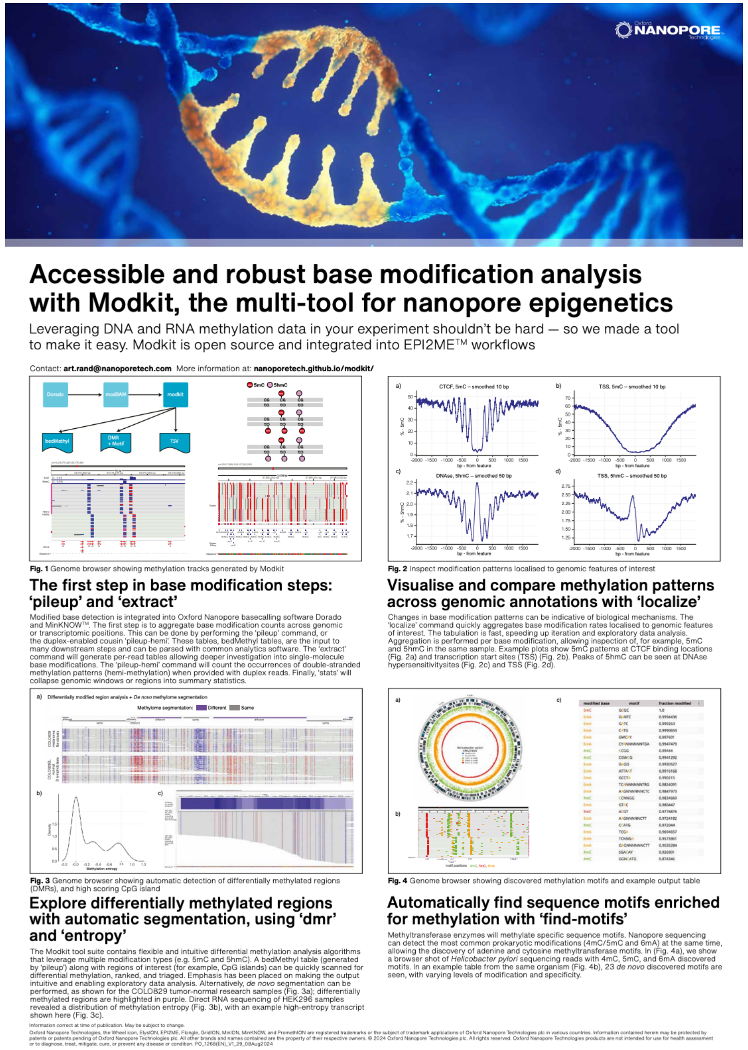


Methylation analysis with Modkit

Available in GitHub – implemented in EPI2ME workflows

- Modkit is a suite of tools for manipulating modified-base data
- Integrated in EPI2ME wf-human-variation
- Best-practices filtering implemented
- Suitable for the analysis of 5mC and 5hmC
- Compatible with chromatin stenciling (6mA): find and qualify promoter regions
- Compatible with hemi-methylation
- Input: BAM files (with MM and ML tags)
- Output: summary counts of modified and unmodified bases (bedMethyl)

More at: github.com/nanoporetech/modkit



Download poster at: https://shorturl.at/oDIIY

Our user community continually releases analysis tools

User-developed tools and algorithms tailored to Oxford Nanopore data

Common tools

- E.g., demultiplexing, filtering, mapping, assembling, variant calling, etc.

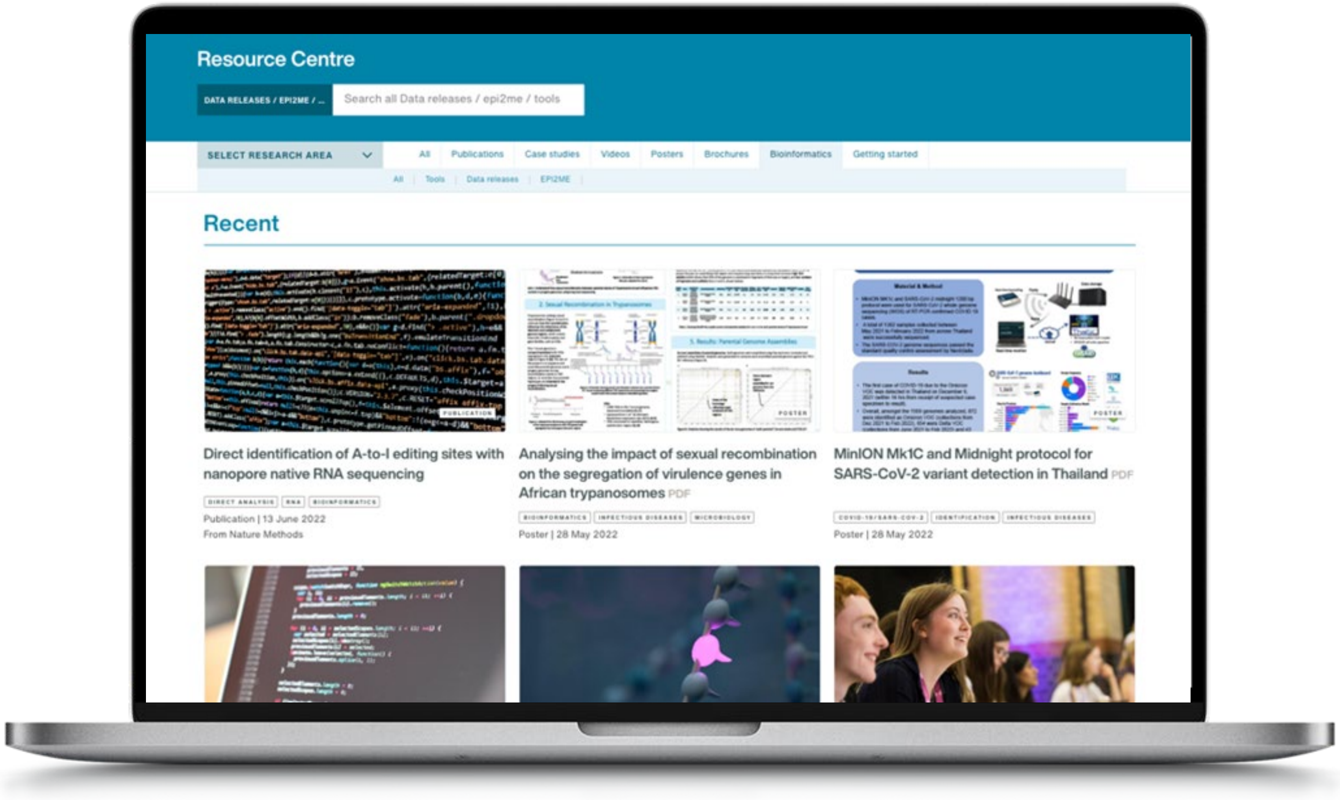
Application-specific tools

- E.g., bacterial genome assembly, SARS-CoV-2 monitoring, hepatitis C genome sequencing, mitochondrial DNA haplogroup classification, novel pathogen detection, transposon insertion identification, etc.

Extended functionality

- E.g., visualization tools for methylated data, RNA methylation prediction, barcode aware adaptive sampling

The bioinformatics section of the nanoporetech.com resource center



More at: nanoporetech.com/resource-centre

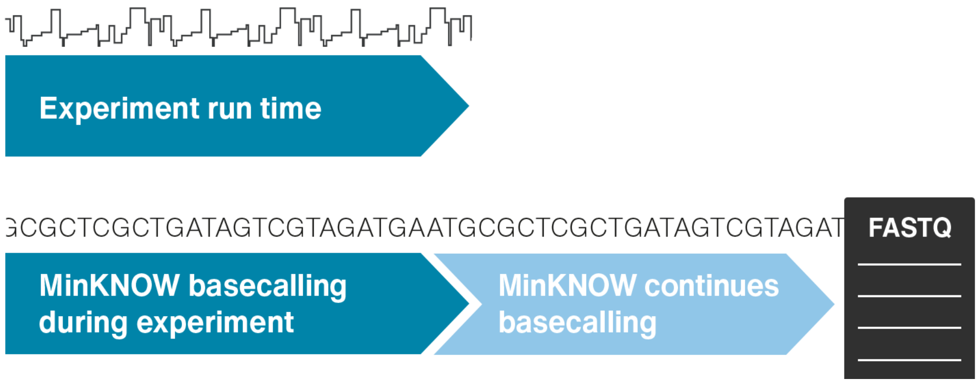
Basecalling with Dorado

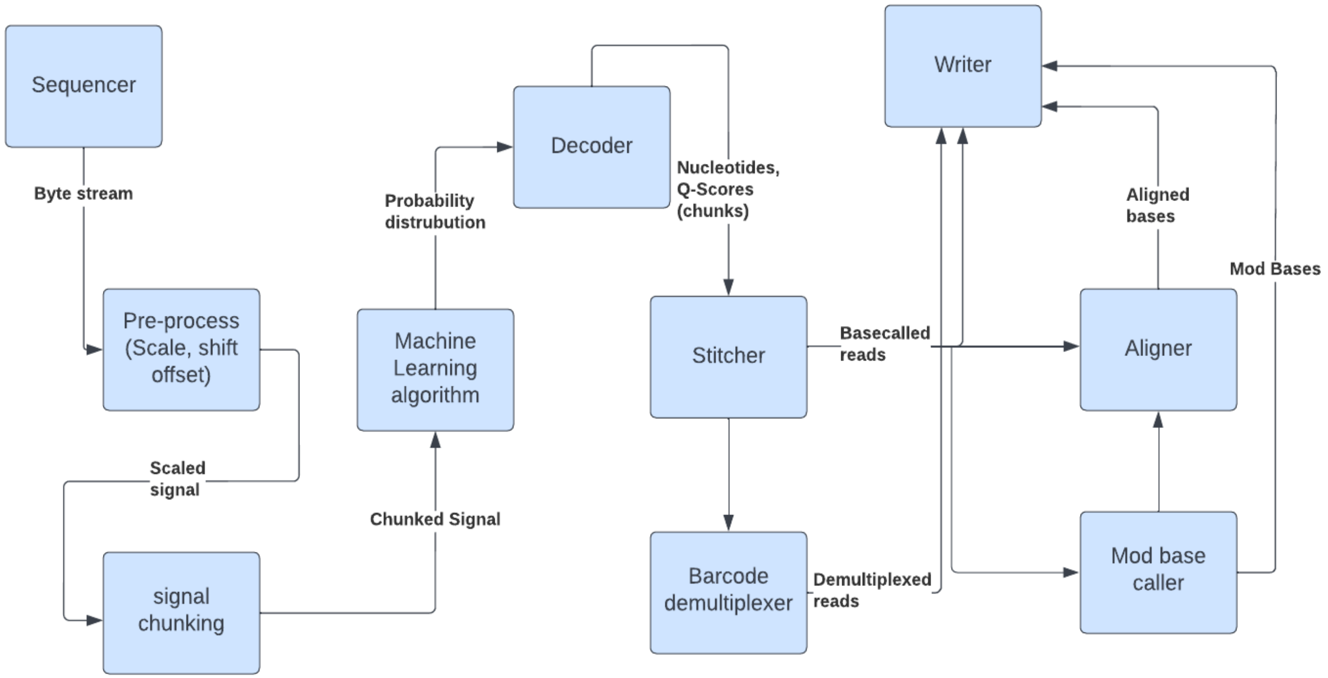
Dorado – a new basecaller

Dorado is the successor to our previous basecaller known as Guppy

- Improvements to speed & usability (equivalent accuracy)

The basecaller uses machine learning models (e.g *dna_r10.4.1_e8.2_400bps_sup@v4.3.0* to translate raw signal data (squiggles) into nucleotide sequence data

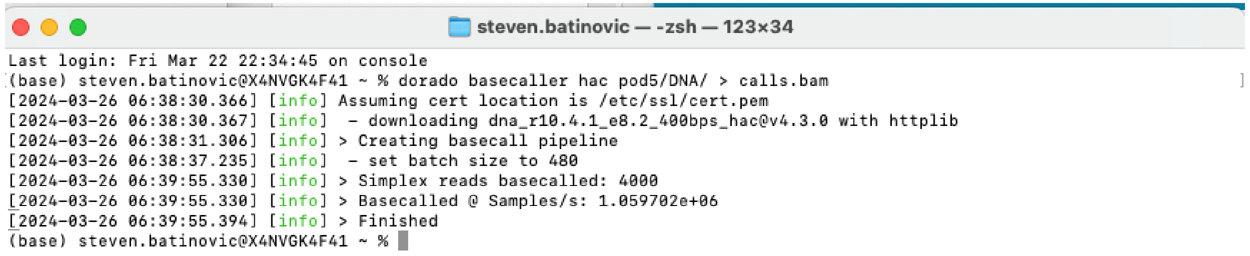




Dorado – a new basecaller

operates as a stand-alone tool, and a MinKNOW-integrated basecaller

- Dorado is a stand-alone, command line application (www.github.com/nanoporetech/dorado)
  - — Stand-alone version is usually a version ahead (and often access to newer models/feature-sets)
- Dorado is also integrated into MinKNOW for live-basecalling and in both MinKNOW and EPI2ME (Linux-only) for post-run basecalling.



Decoding basecaller model names

Basecalling Models	Compatible Modifications	Modifications Model Version	Data Sampling Frequency
dna_r10.4.1_e8.2_400bps_fast@v5.2.0			5 kHz
dna_r10.4.1_e8.2_400bps_hac@v5.2.0	4mC_5mC 5mCG_5hmCG 5mC_5hmC 6mA	v1 v1 v1 v1	5 kHz
dna_r10.4.1_e8.2_400bps_sup@v5.2.0	4mC_5mC 5mCG_5hmCG 5mC_5hmC 6mA	v1 v1 v1 v1	5 kHz

Decoding basecaller model names

- Dorado models names are systematically structured. Example: dna_r10.4.1_e8.2_400bps_sup@v4.3.0
- **Analyte Type (dna)**: denotes the type of analyte being sequenced. For DNA sequencing, it is represented as dna. If you are using the Direct RNA Sequencing Kit, this will be rna.
- **Pore Type (r10.4.1)**: This section corresponds to the type of flow cell used. For instance, FLO-MIN114/FLO-FLG114 is indicated by r10.4.1, while FLO-MIN106D/FLO-FLG001 is signified by r9.4.1.
- **Chemistry Type (e.8.2)**: This represents the chemistry type, which corresponds to the kit used for sequencing. For example, Kit 14 chemistry is denoted by e.8.2.
- **Translocation Speed (400bps)**: speed of translocation.
- **Model Type (sup)**: This represents the size of the model, where larger models yield more accurate basecalls but take more time. The three types of models are fast, hac, and sup. The fast model is the quickest, sup is the most accurate, and hac provides a balance between speed and accuracy. For most users, the hac model is recommended.
- **Model Version Number (v4.3.0)**: This denotes the version of the model. Model updates are regularly released, and higher version numbers typically signify greater accuracy.

dna_r10.4.1_e8.2_400bps_sup@v5.2.0

How to use stand-alone Dorado -> simplex basecalling

- First, download Dorado from www.github.com/nanoporetech/dorado
- Basecalling simplex data from a run (e.g LSK114) using high-accuracy basecalling (hac)
  - dorado basecaller hac /path/to/pod5/data > calls.bam

```
(base) steven.batinovic@X4NVGK4F41 ~ % dorado basecaller hac pod5/DNA/ > calls.bam
[2024-03-26 06:38:30.366] [info] Assuming cert location is /etc/ssl/cert.pem
[2024-03-26 06:38:30.367] [info]  - downloading dna_r10.4.1_e8.2_400bps_hac@v4.3.0 with httplib
[2024-03-26 06:38:31.306] [info] > Creating basecall pipeline
[2024-03-26 06:38:37.235] [info]  - set batch size to 480
[2024-03-26 06:39:55.330] [info] > Simplex reads basecalled: 4000
[2024-03-26 06:39:55.330] [info] > Basecalled @ Samples/s: 1.059702e+06
[2024-03-26 06:39:55.394] [info] > Finished
```

How to use stand-alone Dorado -> methylation detection

- What if you would like to pull out base modification data during basecalling?

- Basecalling simplex data using high-accuracy basecalling (hac) with 5mC and 5hmC detection (in CpG contexts)
  - dorado basecaller hac,5mCG_5hmCG /path/to/pod5/data > calls.bam

```
(base) steven.batinovic@X4NVGK4F41 ~ % dorado basecaller hac,5mCG_5hmCG pod5/DNA/ > calls.bam
[2024-03-26 06:58:45.227] [info] Assuming cert location is /etc/ssl/cert.pem
[2024-03-26 06:58:45.228] [info]  - downloading dna_r10.4.1_e8.2_400bps_hac@v4.3.0 with httplib
[2024-03-26 06:58:46.400] [info]  - downloading dna_r10.4.1_e8.2_400bps_hac@v4.3.0_5mCG_5hmCG@v1 with httplib
[2024-03-26 06:58:47.478] [info] > Creating basecall pipeline
[2024-03-26 06:58:52.977] [info]  - set batch size to 480
[2024-03-26 07:00:32.501] [info] > Simplex reads basecalled: 4000
[2024-03-26 07:00:32.501] [info] > Basecalled @ Samples/s: 8.315003e+05
[2024-03-26 07:00:32.650] [info] > Finished
```

dorado basecaller sup@v5.2.0,m5C_2OmeC,inosine_m6A_2OmeA,pseU_2OmeU,2OmeG pod5s/ > calls.bam

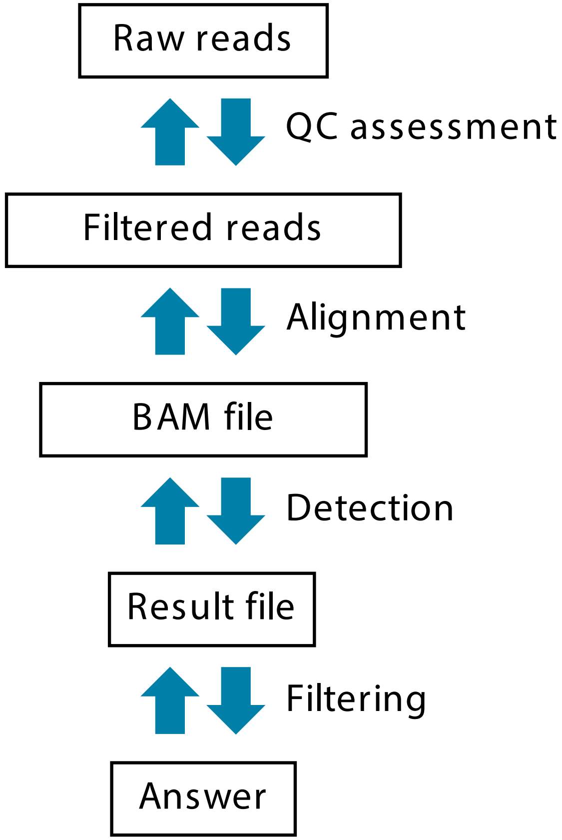
Bioinformatics workflows

Bioinformatics features many file formats

- Raw data (POD5, FASTQ, BAM)
- Alignment (SAM/BAM/)
- Reference genomes (FASTA)
- Annotation files (GTF/GFF/BED)
- Result files (VCF/TSV/BCF/)

Automated bioinformatics

- Pipelines are multi-step analysis workflows
- May involve many software
- Analyses can be automated and run using single command
- Automatically use output from one command as input to another
- Generate reports and results in reproducible way



Oxford Nanopore Technologies

Thank you