

TRAD2026

TREINAMENTO EM REPRODUTIBILIDADE
PARA ANÁLISE DE DADOS

Módulo III: NextFlow



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO





Esta apresentação

- Objetivos de Aprendizado
 - Contextualização
 - Reprodutibilidade
 - Dataflow paradigm
 - Process e Tasks
 - Channels
 - Inside Nextflow
- nf-core
 - Pipelines
 - Ampliseq
 - MAG
 - GenomeAssembler
 - RNASeq

Objetivos de Aprendizado

- Explorar o conceito de orquestrar pipelines
- Criar, configurar e aplicar Nextflow
- Explorar pipelines com nf-core
- Aplicar em dados ômicos do #TRAD2026





Contextualização

```
for sample in *.fastq.gz; do
    fastqc $sample
    bwa mem ref.fa $sample > sample.sam
    samtools sort sample.sam > sample.bam
done
```

- Como retomar após a falha sem reprocessar tudo?
- Como paralelizar por amostra sem controlar manualmente?
- Como definir recursos diferentes para cada etapa?
- Como garantir versões e ambiente computacional?
- Como auditar o que foi executado?



Reprodutibilidade

PERSPECTIVE

The Economics of Reproducibility in Preclinical Research

Leonard P. Freedman , Iain M. Cockburn, Timothy S. Simcoe

Published: June 9, 2015 • <https://doi.org/10.1371/journal.pbio.1002165>

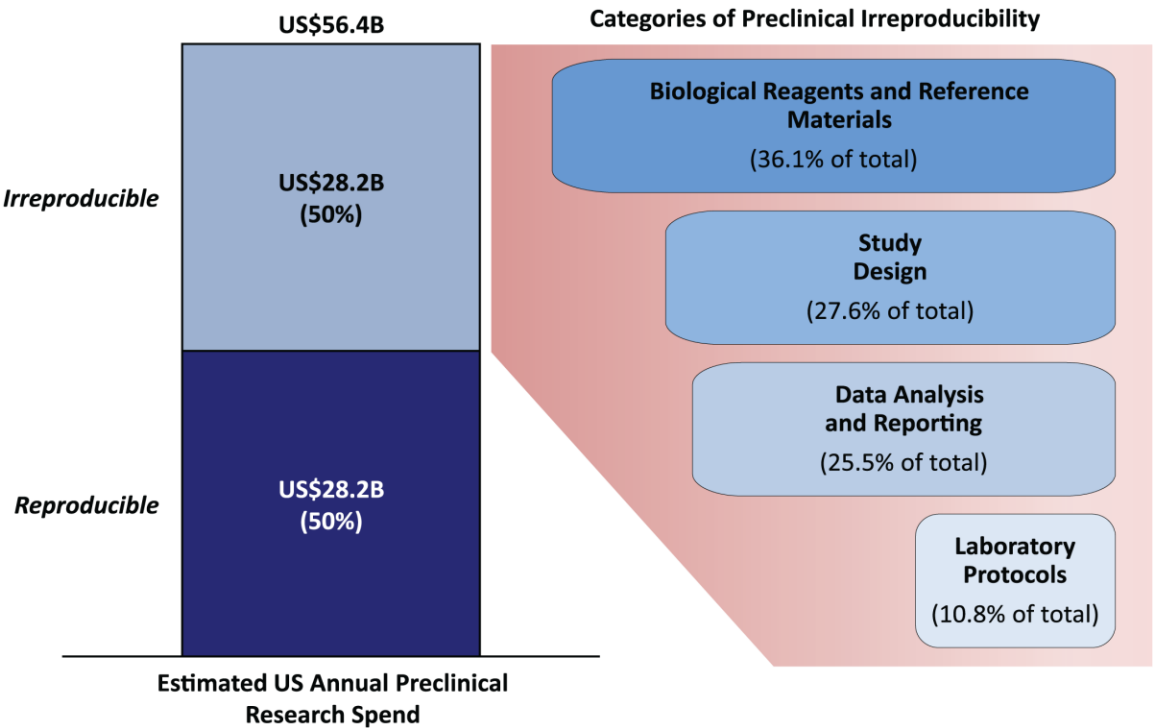
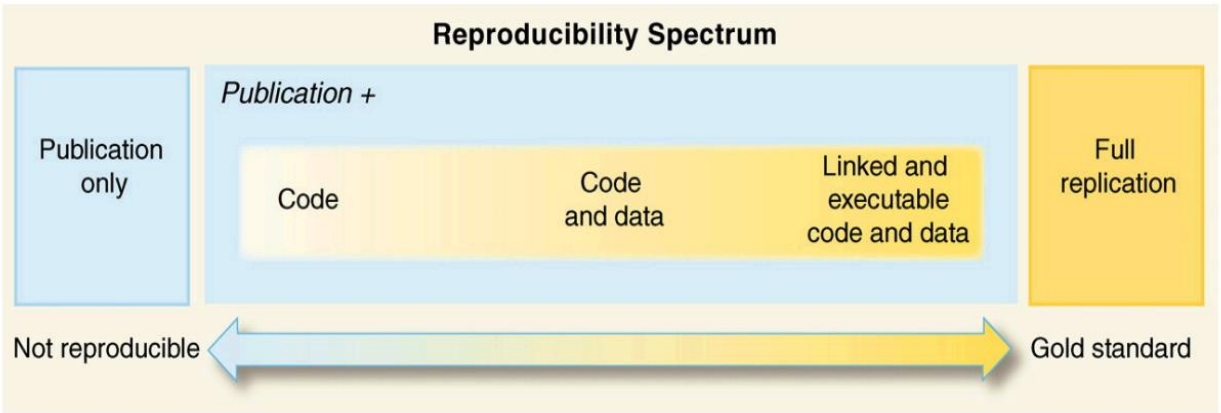


TABLE 1 Simple grid-based system for defining concepts that can be used to describe the validity of a result^a

Methods	Same experimental system	Different experimental system
Same methods	Reproducibility	Replicability
Different methods	Robustness	Generalizability

^aThis is a generalization of the approach used by Whitaker (9), who used it to describe computational analyses.

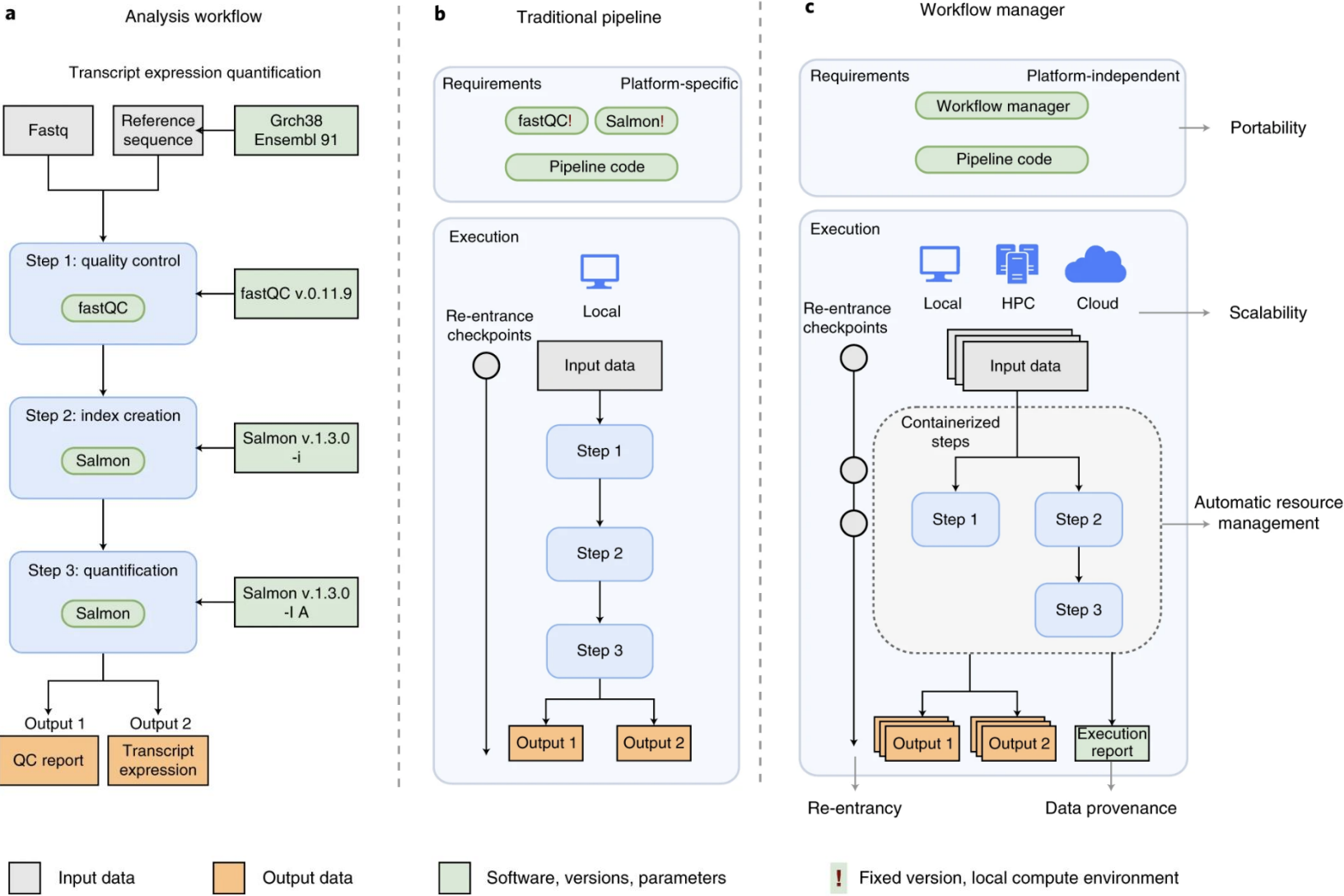
10.1128/mBio.00525-18



10.1126/science.1213847



Workflows





Por que Nextflow?





Reprodutibilidade

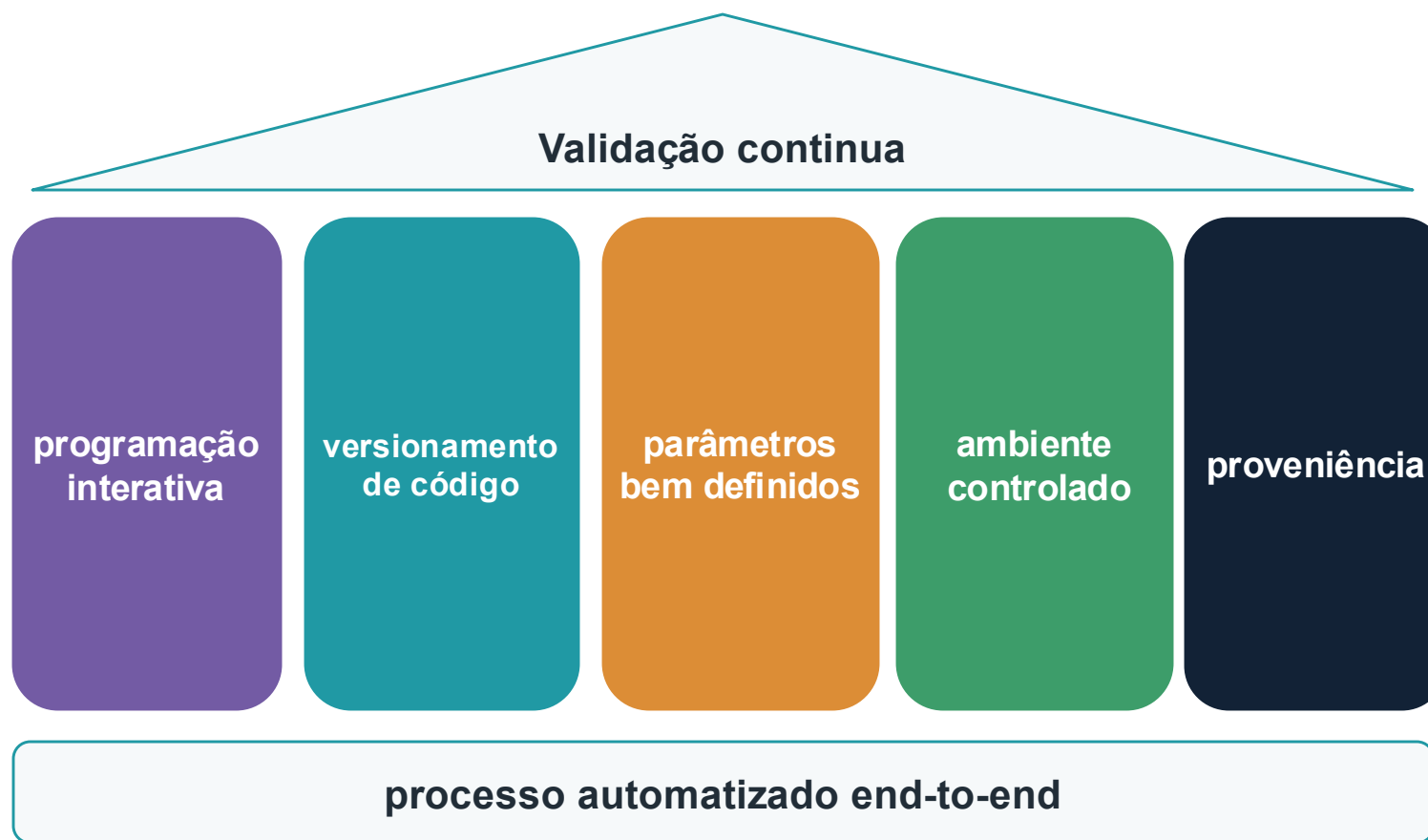


A análise só é reexecutável quando esses elementos são explícitos e recuperáveis.





Reprodutibilidade



JOURNAL ARTICLE

The five pillars of computational reproducibility: bioinformatics and beyond

Mark Ziemann , Pierre Poulain, Anusuiya Bora

Briefings in Bioinformatics, Volume 24, Issue 6, November 2023, bbad375,
<https://doi.org/10.1093/bib/bbad375>



Volume 24, Issue 6
November 2023



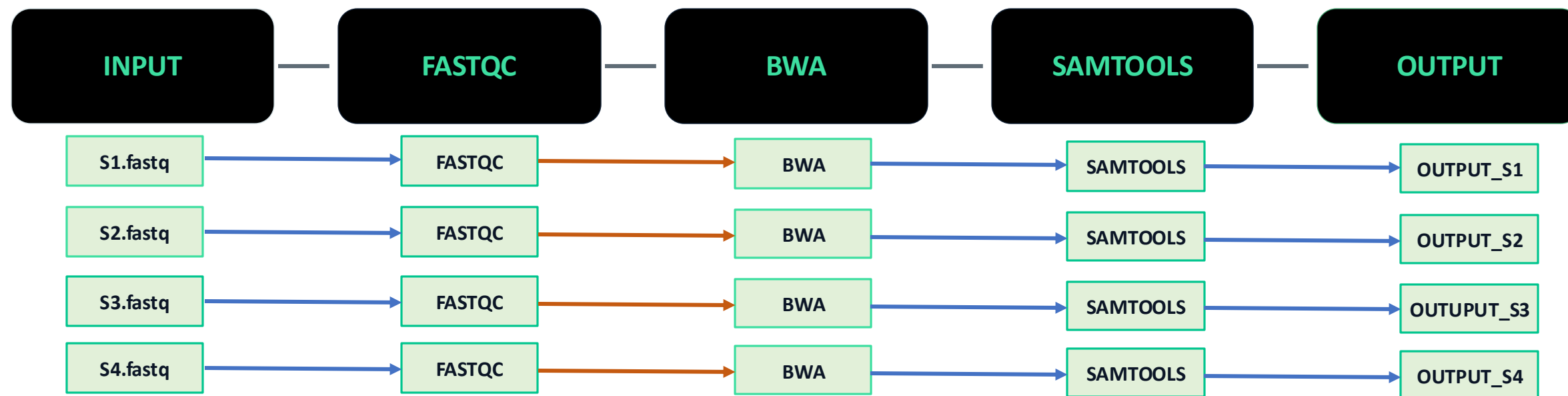
Hello Nextflow

dataflow, processes, channels e tasks



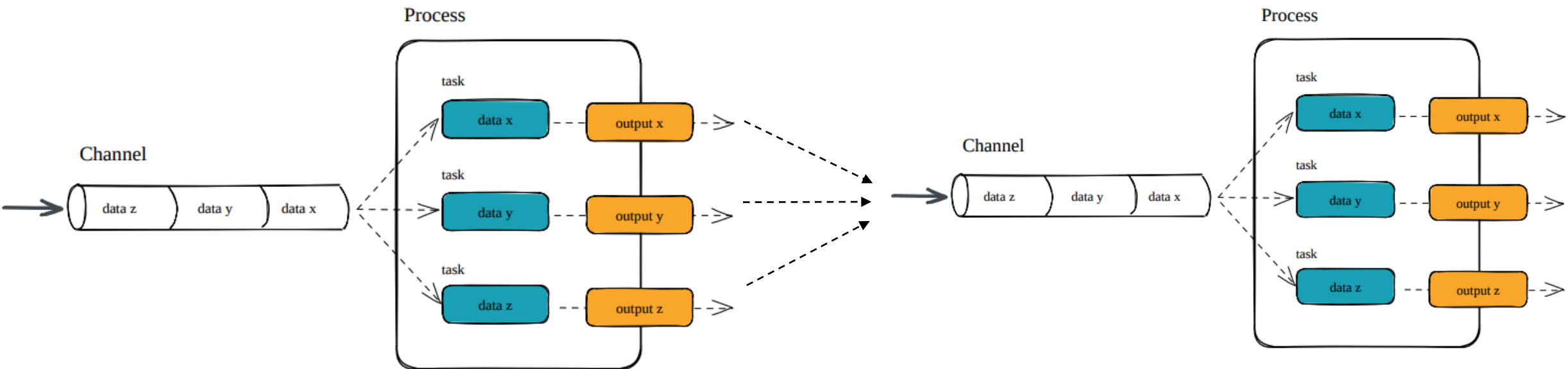


Dataflow paradigm





Process e Tasks



Orquestração de pipelines



Quem rege essa orquestra?

- A) O maestro
- B) O orquestrador do workflow
- C) O SLURM
- D) Java

Source: Open AI ChatGPT

Orquestração de pipelines

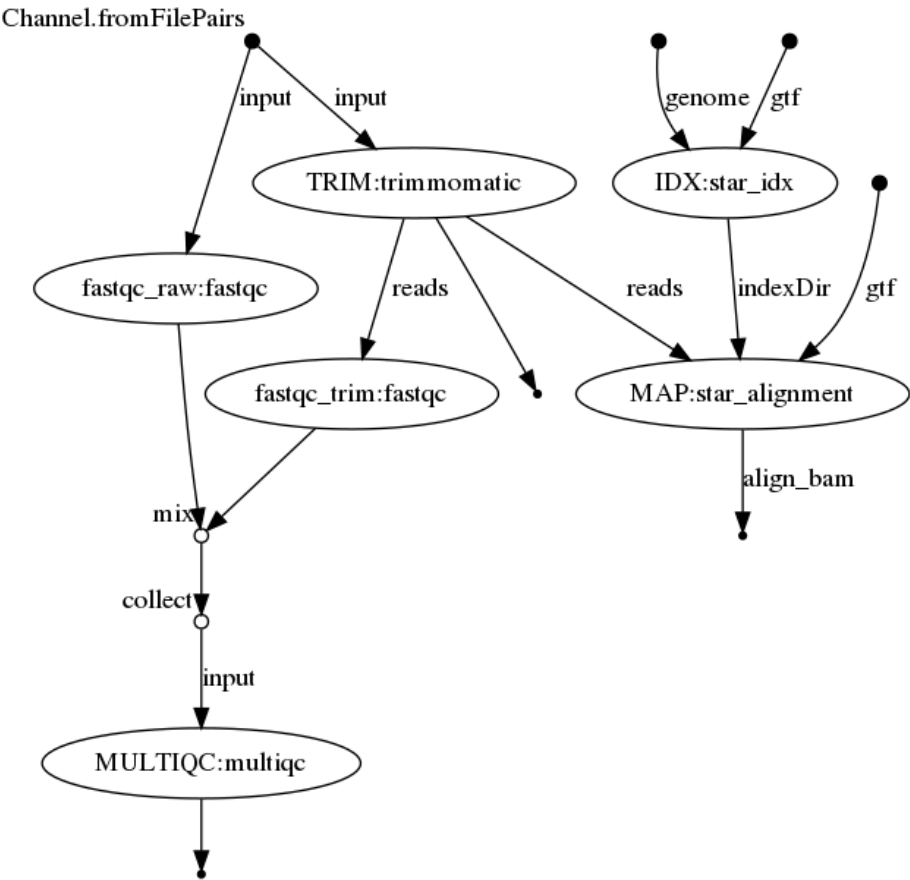


Quem rege essa orquestra?

- A) O maestro
- B) O orquestrador do workflow 
- C) O SLURM
- D) Java



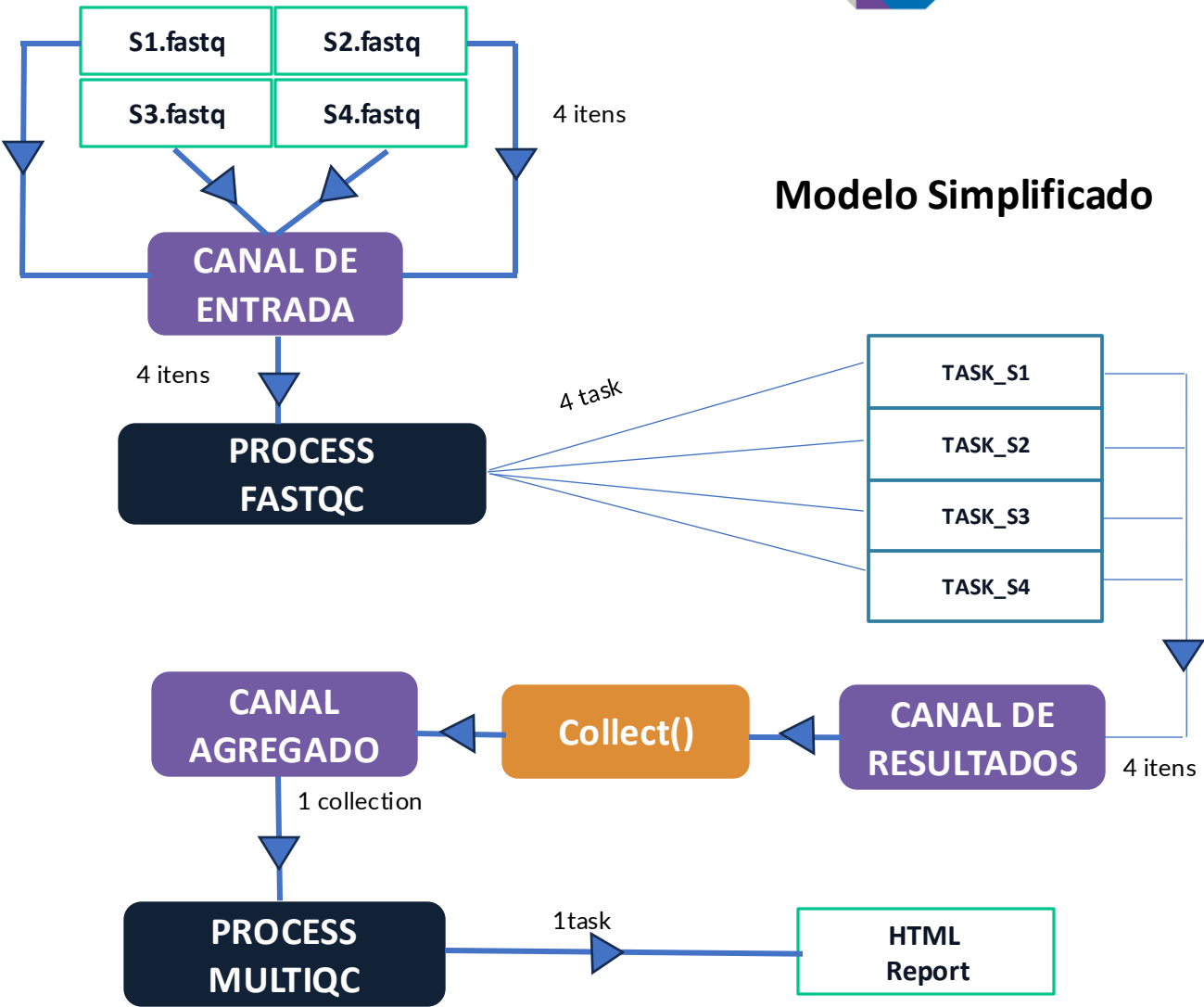
Channels



Grafo de Execução (DAG)

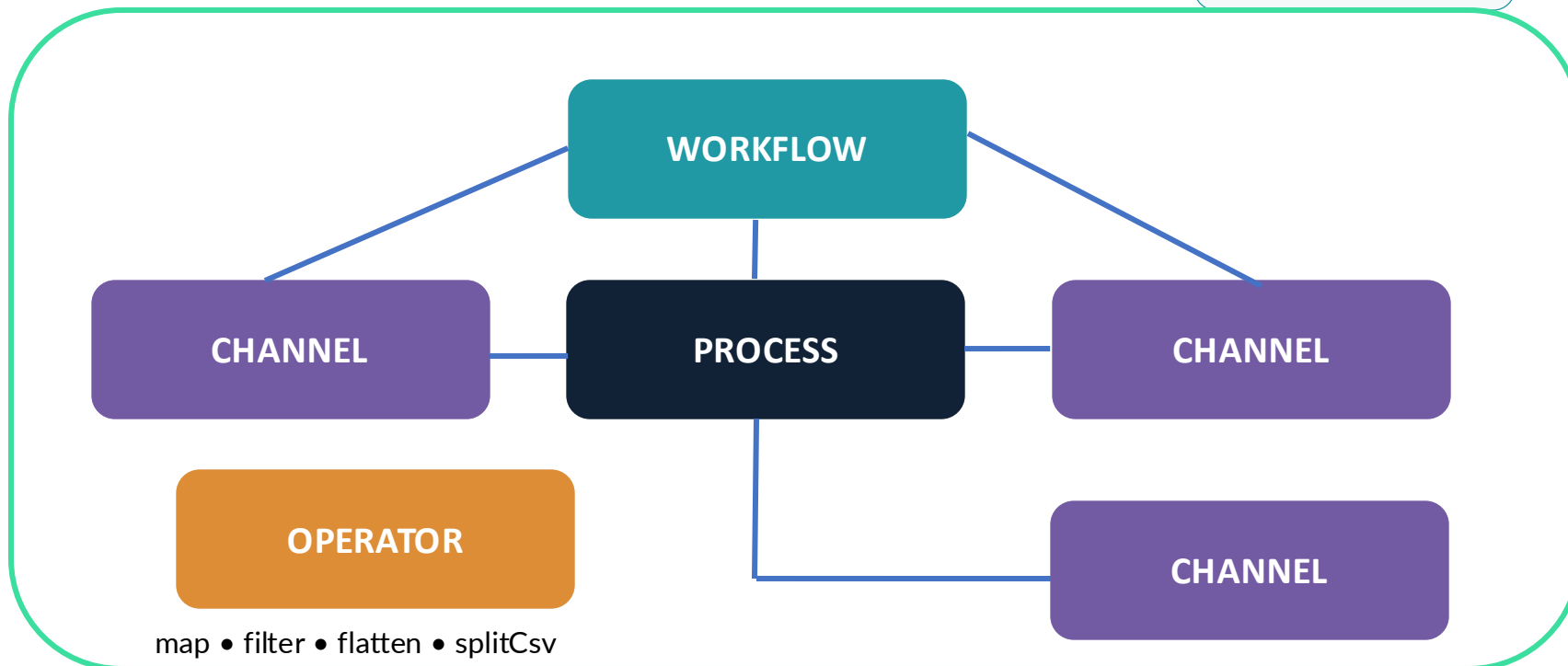


Modelo Simplificado



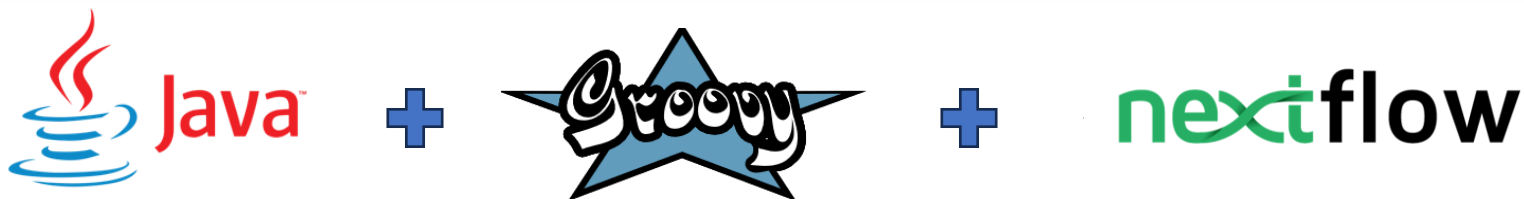
Inside Nextflow

Nextflow DSL2



CONFIG

Nextflow.config
Profiles
Params
Executor
CPU
Memory
Containers



Nextflow Engine



Workflow manager

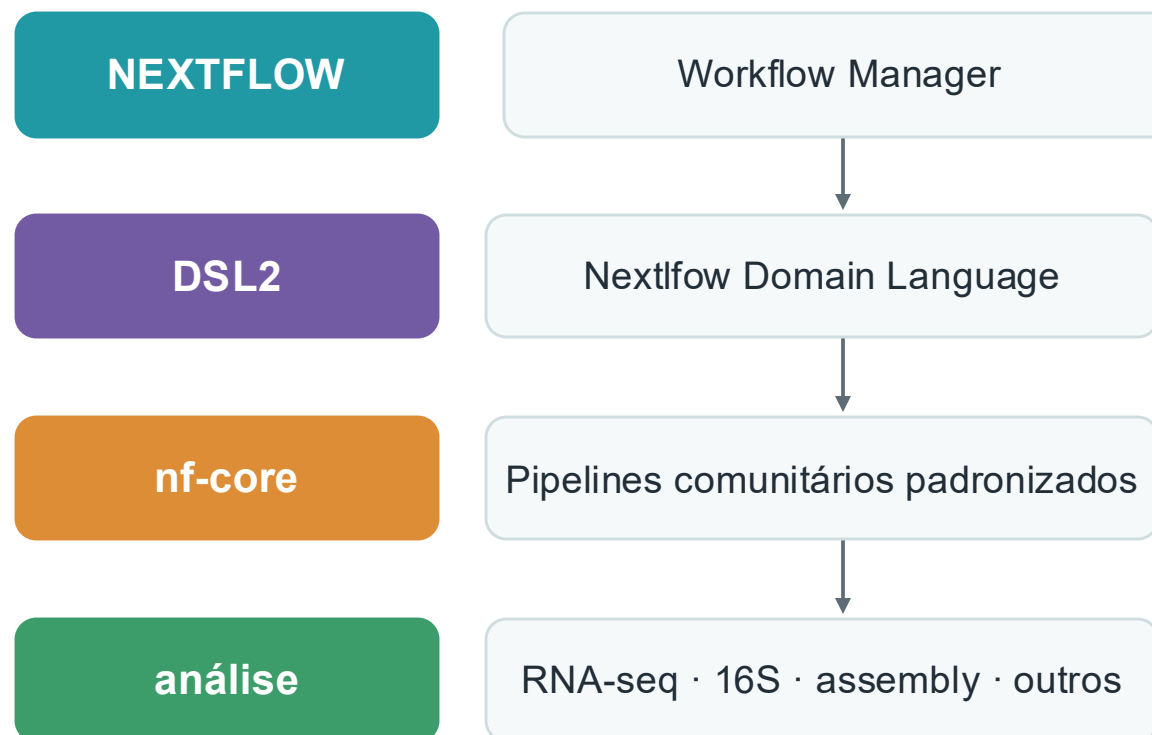
Tabela 01 – Visão geral das ferramentas de gerenciamento de workflows para Bioinformática

Tool	Class	Ease of use ^a	Expressiveness ^b	Portability ^c	Scalability ^d	Learning resources ^e	Pipeline initiatives ^f
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●	●●○
KNIME	Graphical	●●●	●○○	○○○	●●●	●●●	●●○
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●●	●●●	●●○	●●●
GenPipes	DSL	●●○	●●●	●●○	●●○	●●○	●●○
bPipe	DSL	●●○	●●●	●●○	●●●	●●○	●○○
Pachyderm	DSL	●●○	●●●	●○○	●●○	●●●	○○○
SciPipe	Library	●●○	●●●	○○○	○○○	●●○	○○○
Luigi	Library	●●○	●●●	●○○	●●●	●●○	○○○
Cromwell + WDL	Execution + workflow specification	●○○	●●○	●●●	●●●	●●○	●●○
cwltool + CWL	Execution + workflow specification	●○○	●●○	●●●	○○○	●●●	●●○
Toil + CWL/WDL/Python	Execution + workflow specification	●○○	●●●	●●○	●●●	●●○	●●○

Fonte: Wratten et al 2021. *Nature Methods* V.18, p1161–1168.



Visão Geral





nf-core

A global community collaborating to build open-source
Nextflow components and pipelines



nf-core pipelines



<https://nf-co.re/pipelines/>

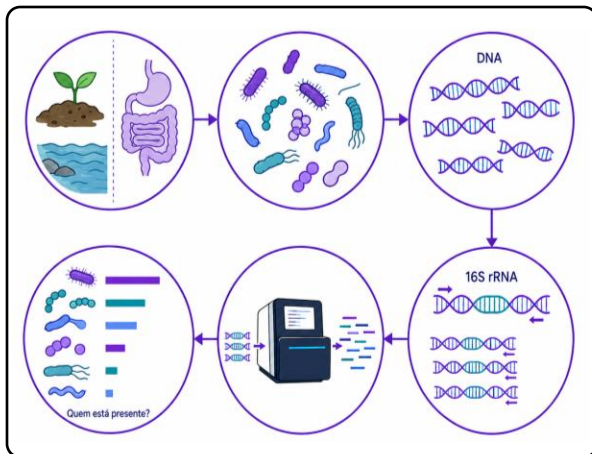
genomeqc ✓ ☆ 26 Compare the quality of multiple genomes, along with their annotations. genome-assembly-evaluation genomics phylogenetic-trees quality-control 1.0.0 released 2 days ago	oncoanalyser ✓ ☆ 120 A comprehensive cancer DNA/RNA analysis and reporting pipeline cancer clinical ctdna dna exome longitudinal ngs panel rna targeted wgs wigits wts 3.0.0 released 2 days ago	fastqrepair ✓ ☆ 9 A pipeline that can be used to recover corrupted FASTQ.gz files, drop or fix uncompliant reads, remove unpaired reads, and settles reads that became disordered corruption data-cleaning data-recovery fastq fastq-corrupted fastq-format reads-interleaving recovery-tool unpaired-reads well-formed 1.1.1 released 3 days ago	airrflow ✓ ☆ 81 B-cell and T-cell Adaptive Immune Receptor Repertoire (AIRR) sequencing analysis pipeline using the Immcantation framework airr b-cell immcantation immunorepertoire repseq 5.1.1 released 6 days ago	pairgenomealign ✓ ☆ 12 Pairwise genome comparison pipeline using the LAST software to align a list of query genomes to a target genome, and plot the results comparative-genomics dot-plot genomics last pairwise-alignment synteny whole-genome-alignment 3.0.4 released 8 days ago	sopa ✓ ☆ 19 Nextflow version of Sopa - spatial omics pipeline and analysis segmentation spatial-omics spatial-proteomics spatial-transcriptomics spatialdata 1.0.1 released 9 days ago
pixelator ✓ ☆ 14 Pipeline to generate Proximity Network Assay data with Pixelator (Pixelgen Technologies AB) molecular-pixelation pixelator pixelgen-technologies proteins single-cell single-cell-omics 5.0.1 released 14 days ago	seqinspector ✓ ☆ 28 Dedicated QC-only pipeline for sequencing data. The pipeline will run a (potentially large) set of QC tools and can output global and group specific Multiqc reports. The pipeline is targeting core facilities or research groups with larger sequencing throughput. genomics quality-control 1.1.2 released 15 days ago	sarek ✓ ☆ 598 Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing annotation cancer gatk4 genomics germline pre-processing somatic target-panels variant-calling whole-exome-sequencing whole-genome-sequencing 3.10.0 released 22 days ago	magmap ✓ ☆ 12 Best-practice analysis pipeline for mapping reads to a (large) collections of genomes 1.2.0 released 23 days ago	proteinfamilies ✓ ☆ 23 Generation and updating of protein families metagenomics protein-families 2.5.0 released 23 days ago	ismquant ✓ ☆ 6 A pipeline for processing and analysis of light-sheet microscopy images. 3dunet image-analysis image-processing light-sheet-microscopy segmentation stitching 1.0.3 released 24 days ago
epitopeprediction ✓ ☆ 56 A bioinformatics best-practice analysis pipeline for epitope prediction and annotation epitope epitope-prediction mhc-binding-prediction 3.1.1 released 27 days ago	spatialaxe ✓ ☆ 46 A bioinformatics best-practice processing and quality control pipeline for Xenium and Artera data 10x-genomics atera image-processing spatial spatial-data-analysis spatial-transcriptomics transcriptomics xenium 1.0.1 released 28 days ago	seqsubmit ✓ ☆ 18 nf-core pipeline for data submission to ENA 1.0.0 released about 1 month ago	phyloplace ✓ ☆ 12 nf-core/phyloplace is a bioinformatics best-practice analysis pipeline that performs phylogenetic placement with EPA-NG. evolution evolutionary-tree phylogenetic-placement phylogenetics sequence-classification taxonomy-assignment 2.1.0 released about 1 month ago	mag ✓ ☆ 319 Assembly and binning of metagenomes annotation assembly binning long-read-sequencing metagenomes metagenomics nanopore nanopore-sequencing 5.5.0 released about 1 month ago	raredisease ✓ ☆ 123 Call and score variants from WGS/WES of rare disease patients. diagnostics rare-disease srvc structural-variants variant-annotation variant-calling wes wgs 3.1.2 released about 2 months ago



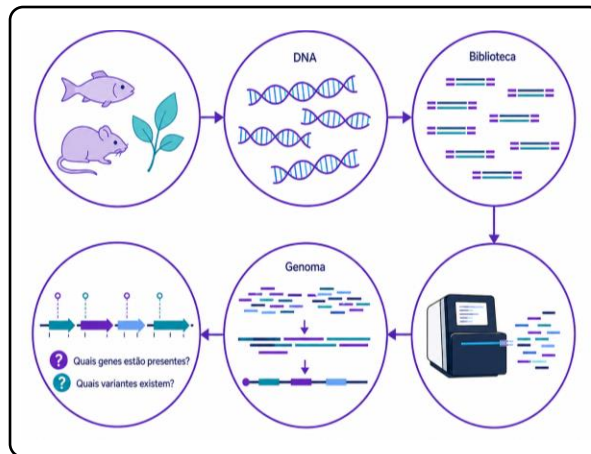


O conteúdo pode ser acessado no GitHub: [TRAD2026/DATASETS.md](https://github.com/Trad2026/DATASETS.md) e [TRAD2026/OMICS_DATA_TYPES.md](https://github.com/Trad2026/OMICS_DATA_TYPES.md)

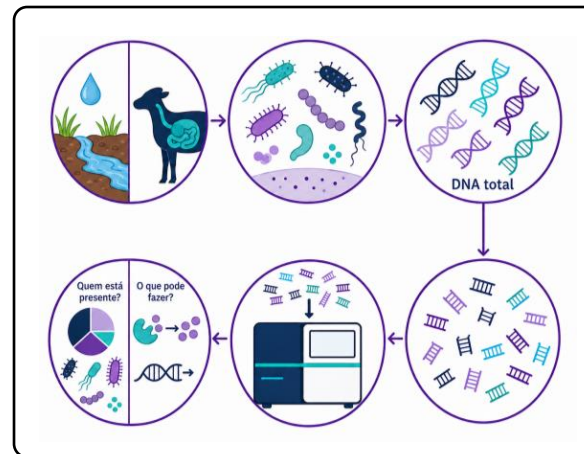
AMPLICON



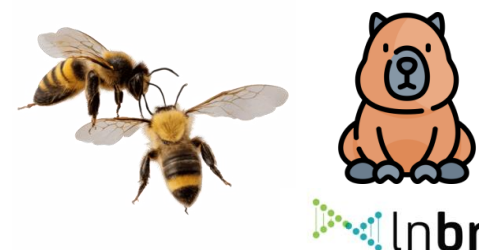
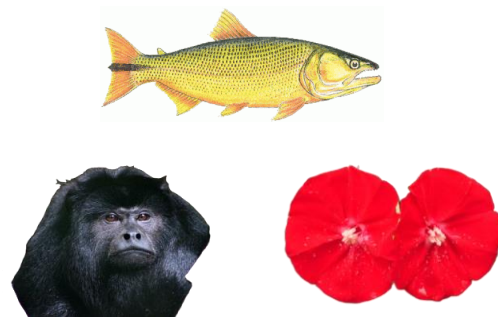
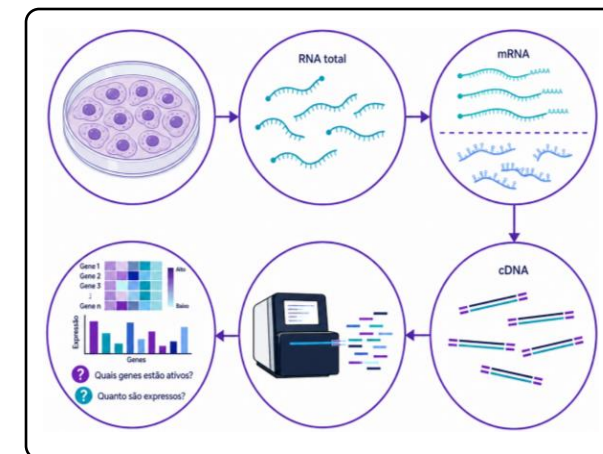
GENÔMICA

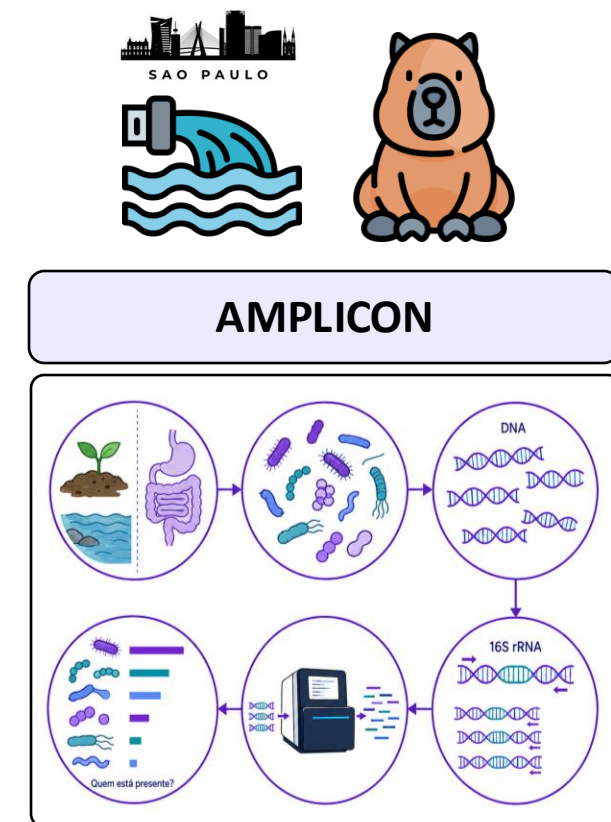


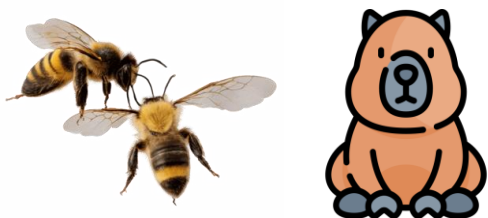
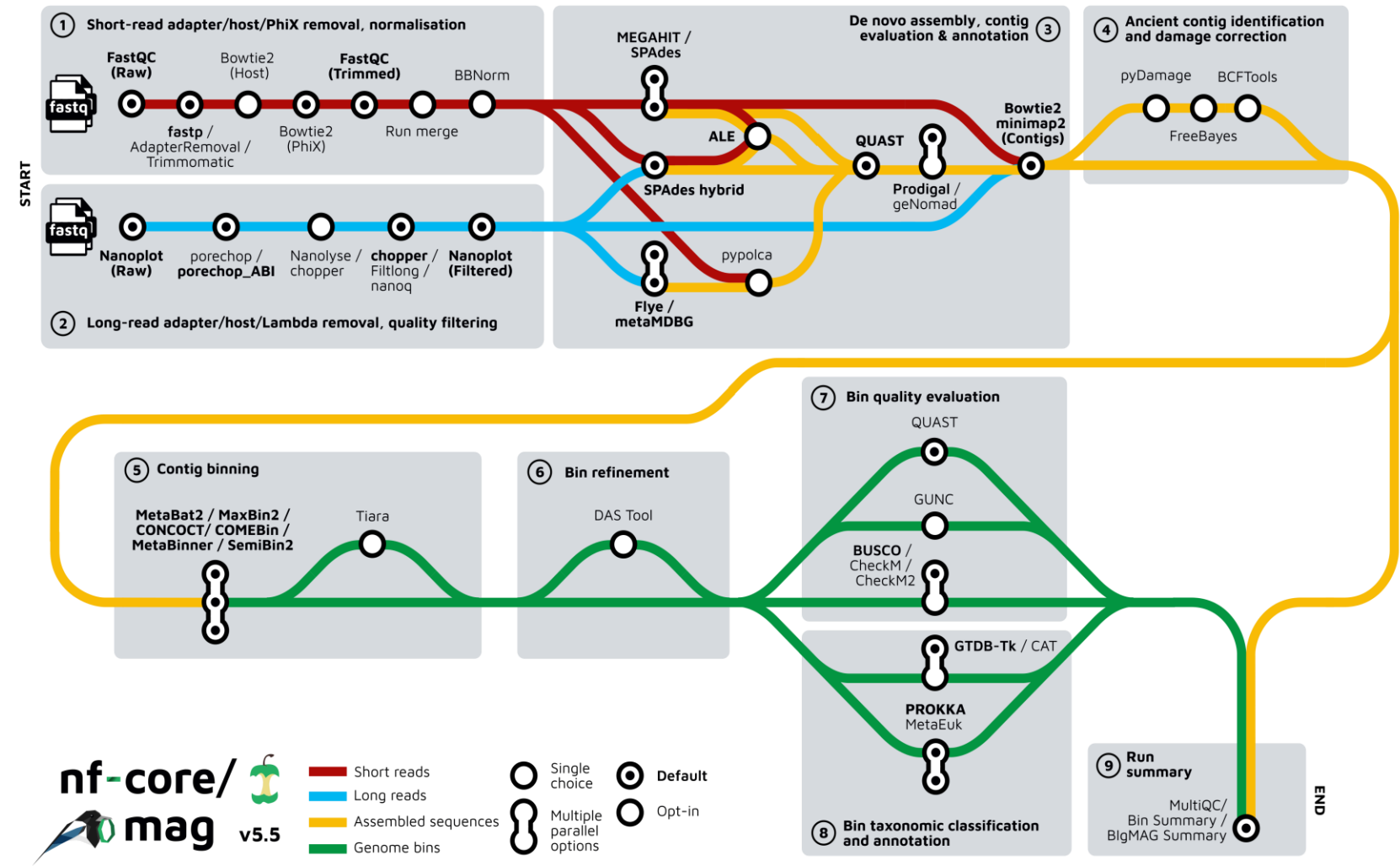
METAGENÔMICA



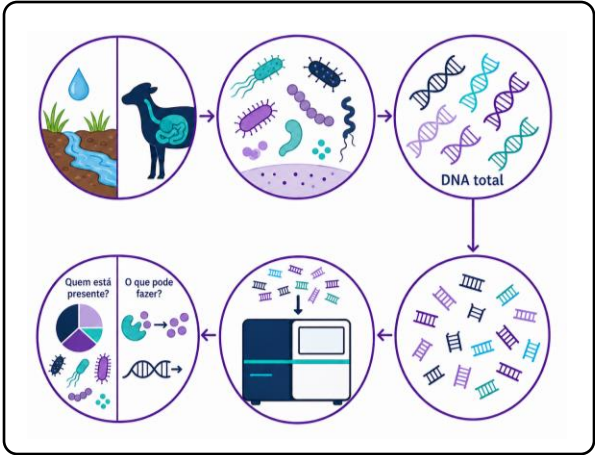
RNA-SEQ



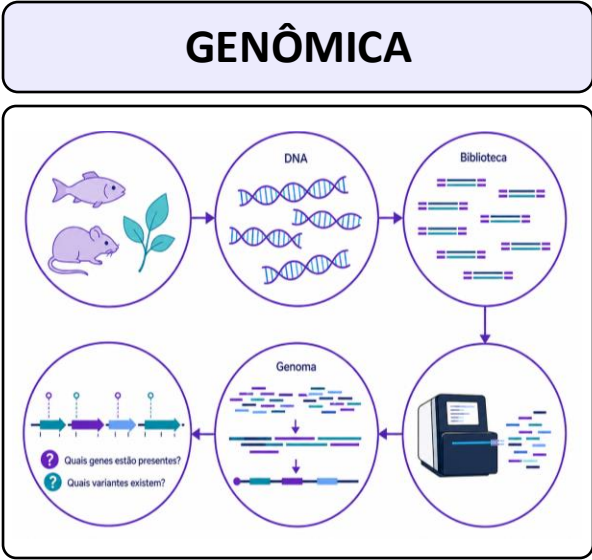
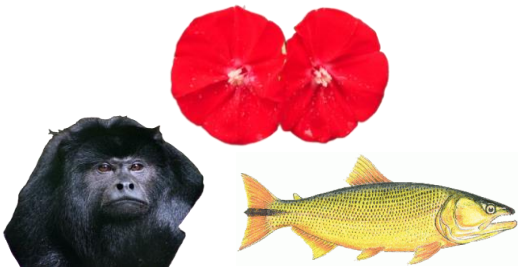
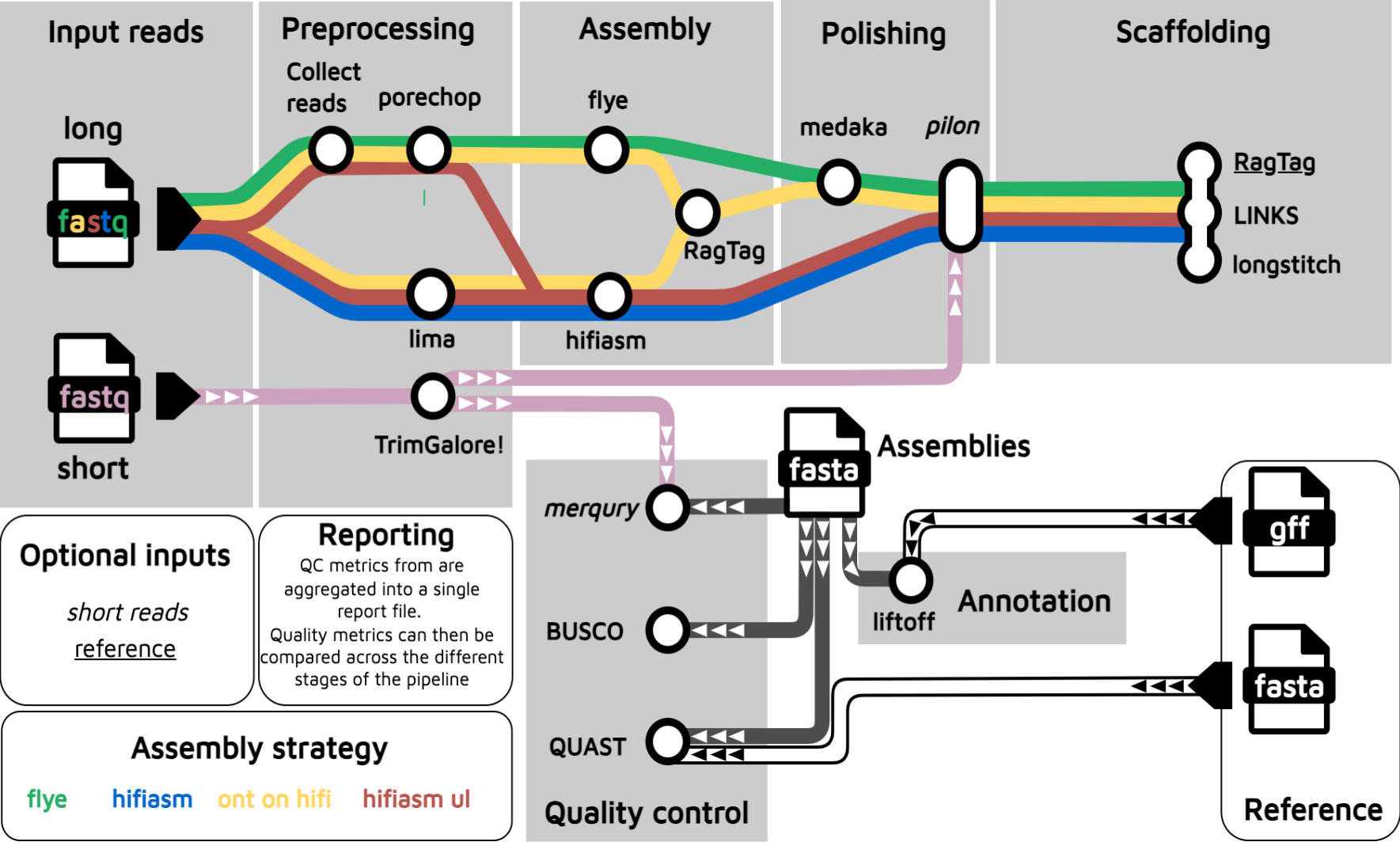


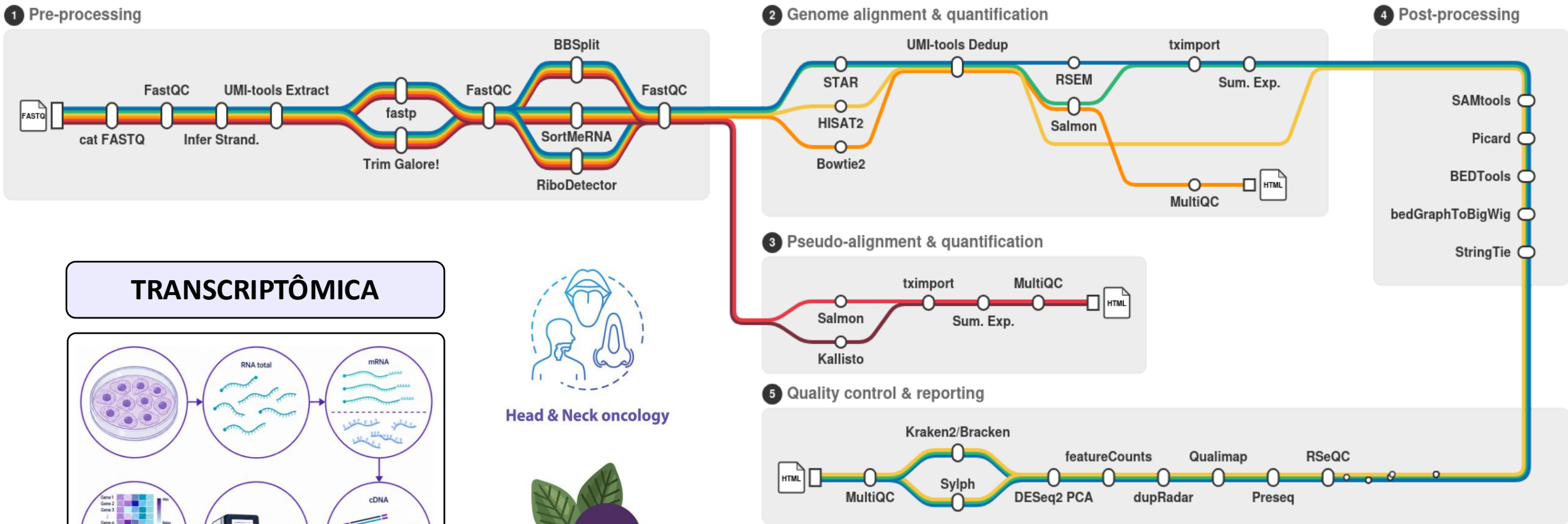


METAGENÔMICA

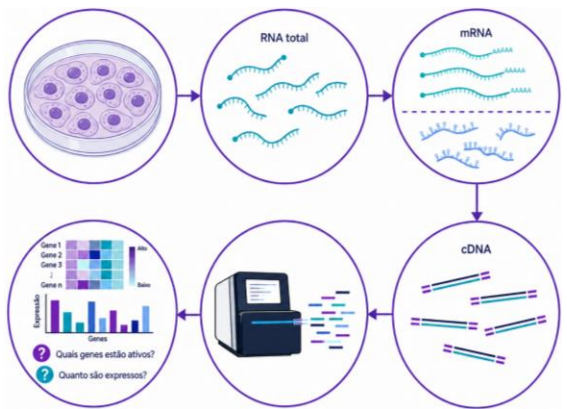


nf-core /genomeassembler





TRANSCRIPTÔMICA



Head & Neck oncology



- Aligner: STAR, Quantification: RSEM
- Aligner: STAR, Quantification: Salmon (default)
- Aligner: HISAT2, Quantification: None
- Aligner: Bowtie2, Quantification: Salmon
- Pseudo-aligner: Salmon, Quantification: Salmon
- Pseudo-aligner: Kallisto, Quantification: Kallisto

created with nf-metro v0.5.4





Hands-on

O conteúdo desta sessão pode ser acessado no GitHub: [TRAD2026/Nextflow](#)



OBIGADO



Nilson Coimbra

nilson.coimbra@lnbio.cnpem.br

[@nilsoncoimbra](#) 



Monyque Karolinie Silva

monyque.silva@lnbr.cnpem.br

[@kpsmonyque](#) 



Inscreva-se para receber
comunicados sobre o **CNPEM**
e suas unidades

cnpem.br

Material do curso

<https://github.com/TRAD2026>

Laboratório Nacional de Biociências

João V. S. Guerra (EDB)

José G. C. Pereira (EDB)

Nilson A. R. Coimbra (MAS)

Rick H. Hokama (EDB)

Laboratório Nacional de Biorrenováveis

Joaquim M. Junior (Ômicas Integrativas)

Monyque K. P. Silva (Ômicas Integrativas)



CNPEM

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

