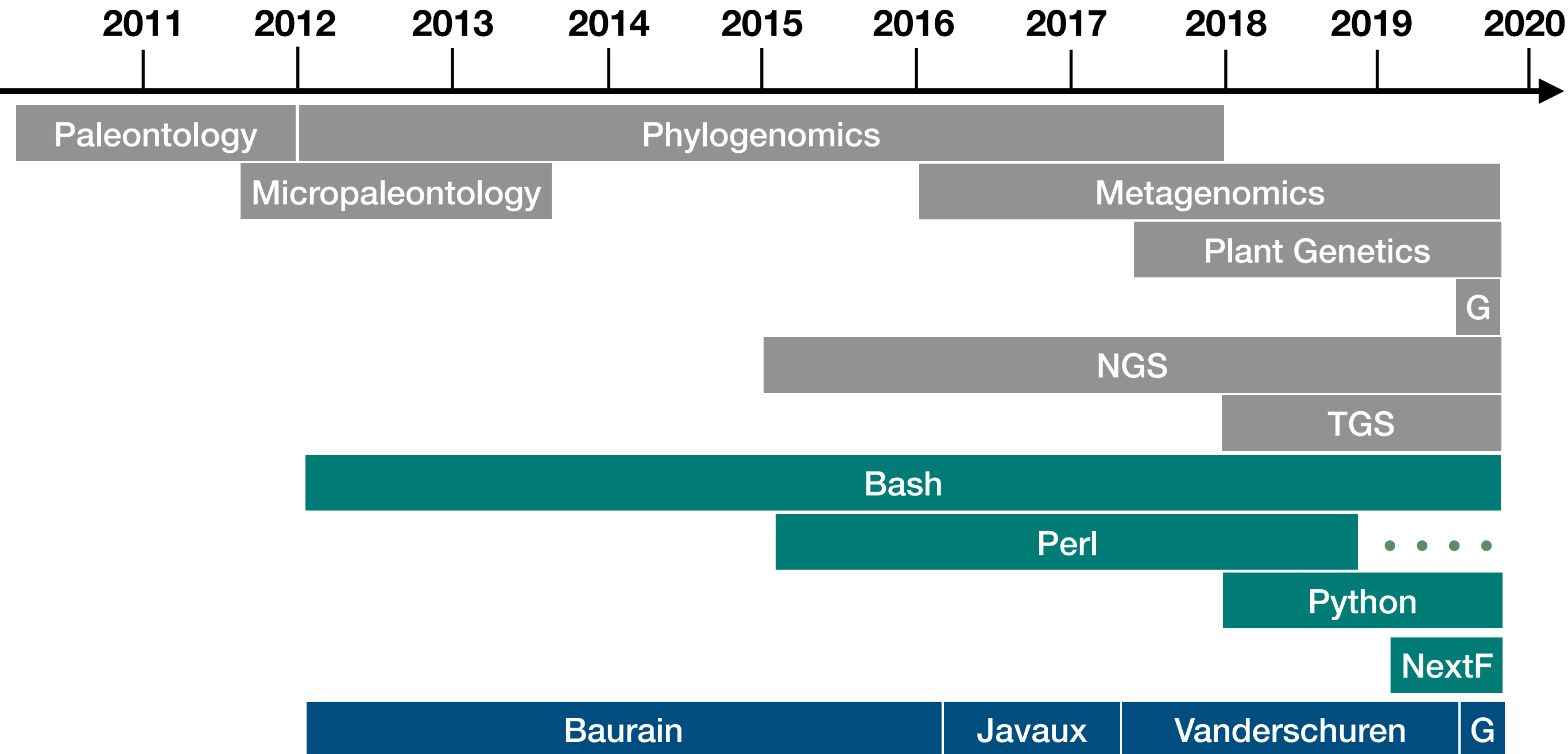


Luc Cornet

Bioinformatician



Genomics



A BAC-guided assembly of cassava using ultra long reads improves accuracy in repeat-rich regions

Cornet Luc, Syed Shan-e-Ali Zaidi, Marc Hanikenne, Denis Baurain, Hervé Vanderschuren



Importance of Cassava

- **Important food security crop in Africa**
- **Feeds more than 500 million people in Africa, Asia and Latin America**
- **Drive Africa economic development (industrial usage)**

Cassava mosaic disease (CMD)

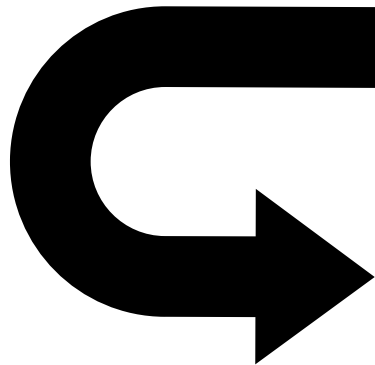


60444

TME3
CMD resistant

Cassava project

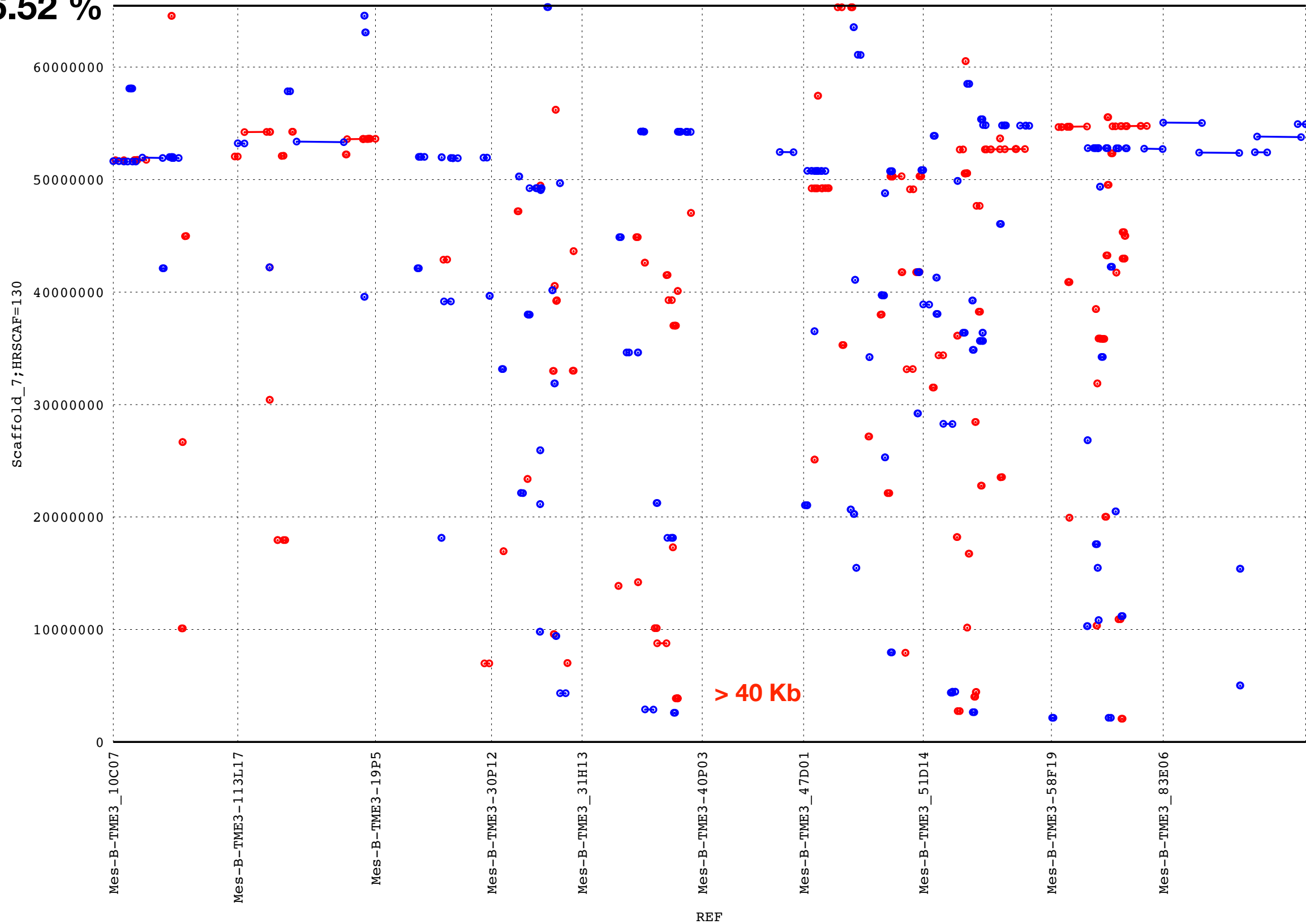
- PacBio sequencing
- Illumina HiSeq sequencing
- Optical map scaffolding
- Hi-C sequencing



New assembly

Breakpoints = 1108
Insertions = 253
Deletions = 146
Bac = 0/10
Id = 96.52 %

Zurich - TME3





New Cassava project

- **Assembly close to biological reality**
- **Identify resistant gene(s)**

5 clusters

**950 000 CPU hours
(108 years)**

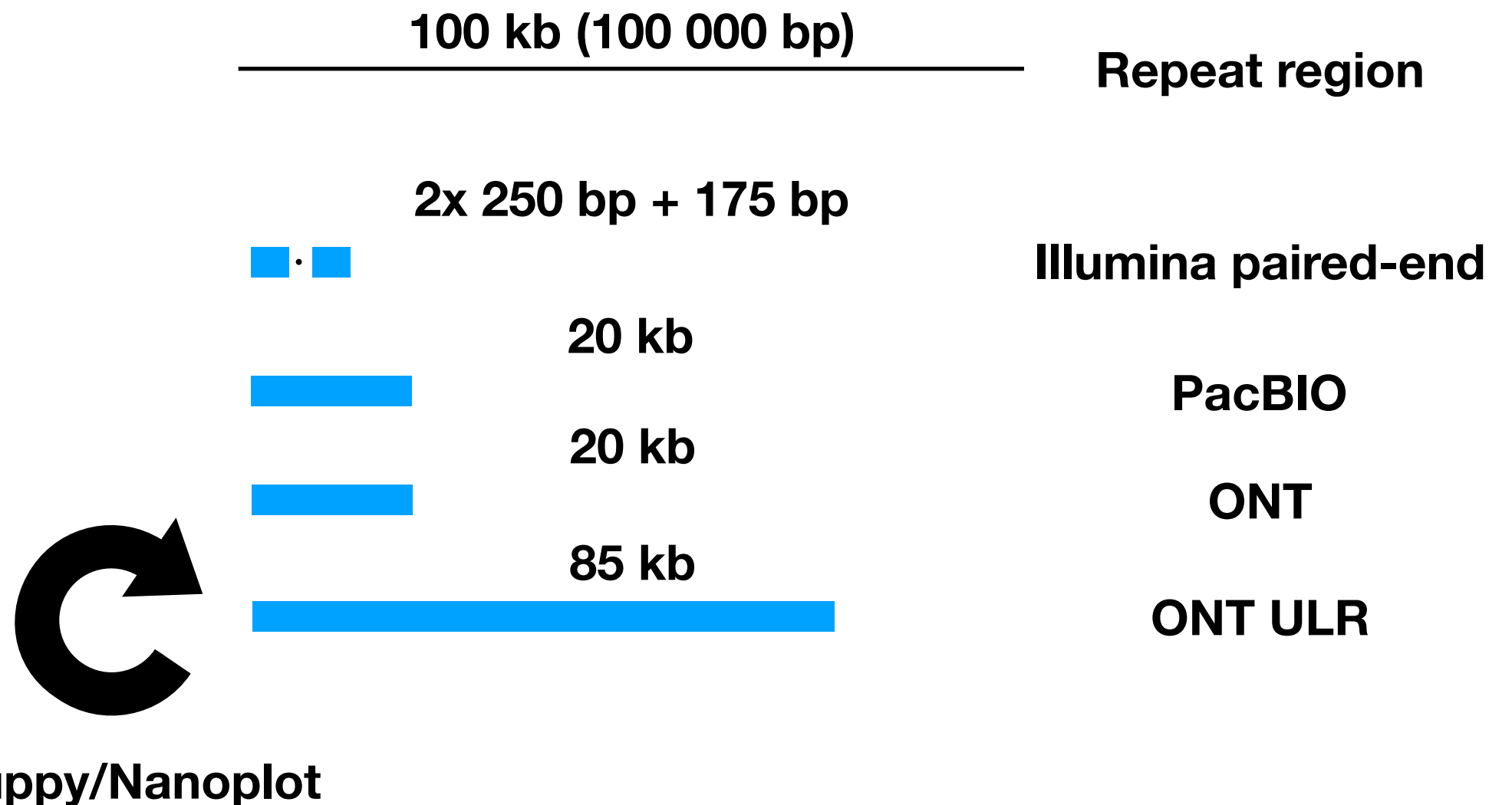
90 Tera

2 years

New Cassava project

Data	fastq ONT
Program	Guppy/ Nanoplot
Time	24 h
Storage	50 Gb
RAM	256 Gb
CPU	32 cpus
Computer	Working station

Selection of sequencing technology



New Cassava project

Data	fastq ONT
Program	Guppy/ Nanoplot
Time	24 h
Storage	500 Gb
RAM	256 Gb
CPU	32 cpus
Computer	Working station

Selection of sequencing technology

NanoPlot report

Summary statistics

General summary:

Mean read length: 3,949.6

Mean read quality: 7.4

Median read length: 1,276.0

Median read quality: 7.5

Number of reads: 677,772.0

Read length N50: 11,170.0

Total bases: 2,676,899,391.0

Number, percentage and megabases of reads above quality cutoffs

>Q5: 595747 (87.9%) 2573.9Mb

>Q7: 415180 (61.3%) 2010.2Mb

>Q10: 53062 (7.8%) 258.1Mb

>Q12: 328 (0.0%) 0.6Mb

>Q15: 0 (0.0%) 0.0Mb

Raw data from 10 flow cells -> 500 Gb

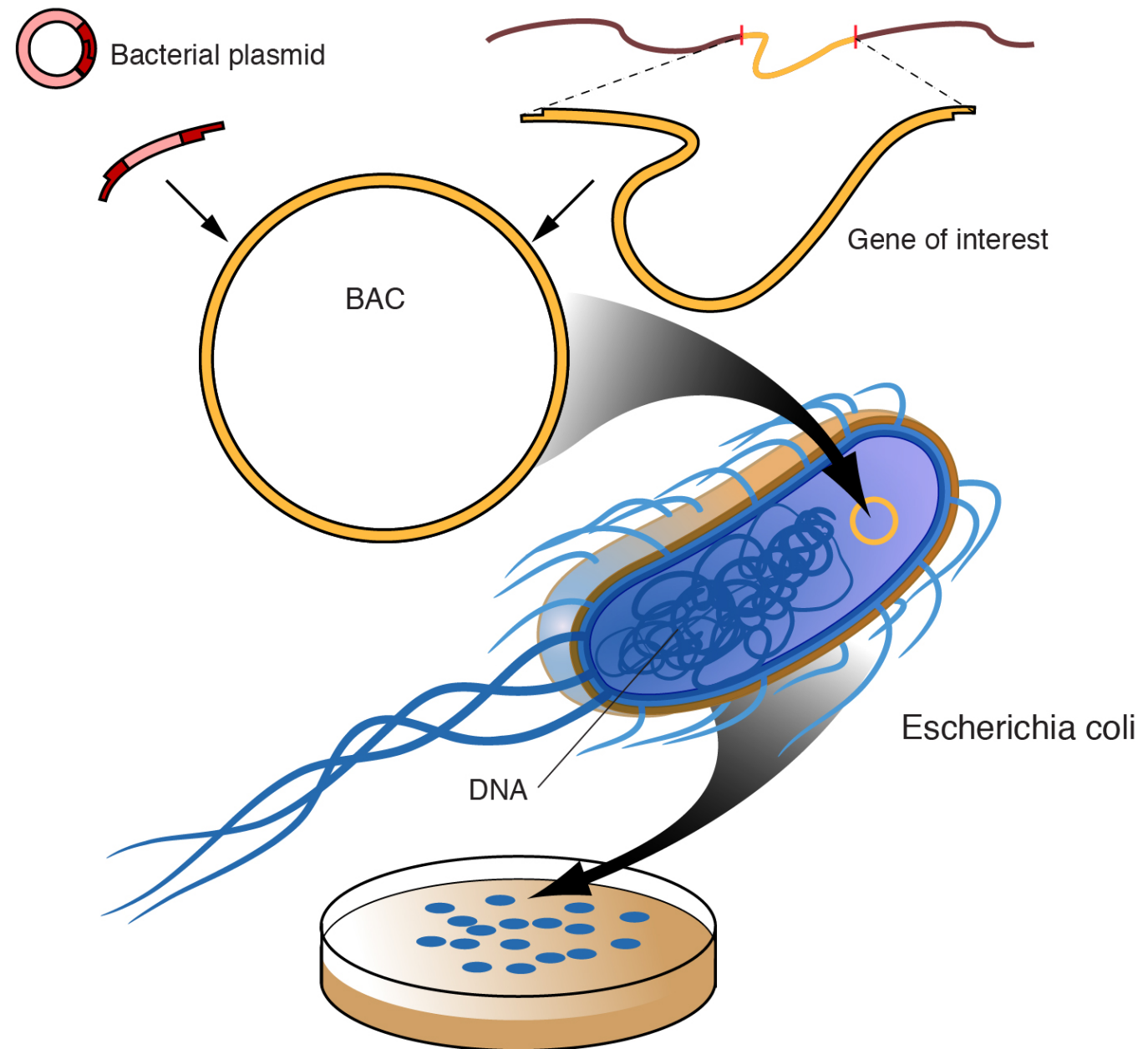


ONT raw data

New Cassava project

Data	fastq PacBio
Program	SMRT ccs
Time	24 h
Storage	25 Gb
RAM	256 Gb
CPU	32 cpus
Computer	HPC

Assembly validation : BAC





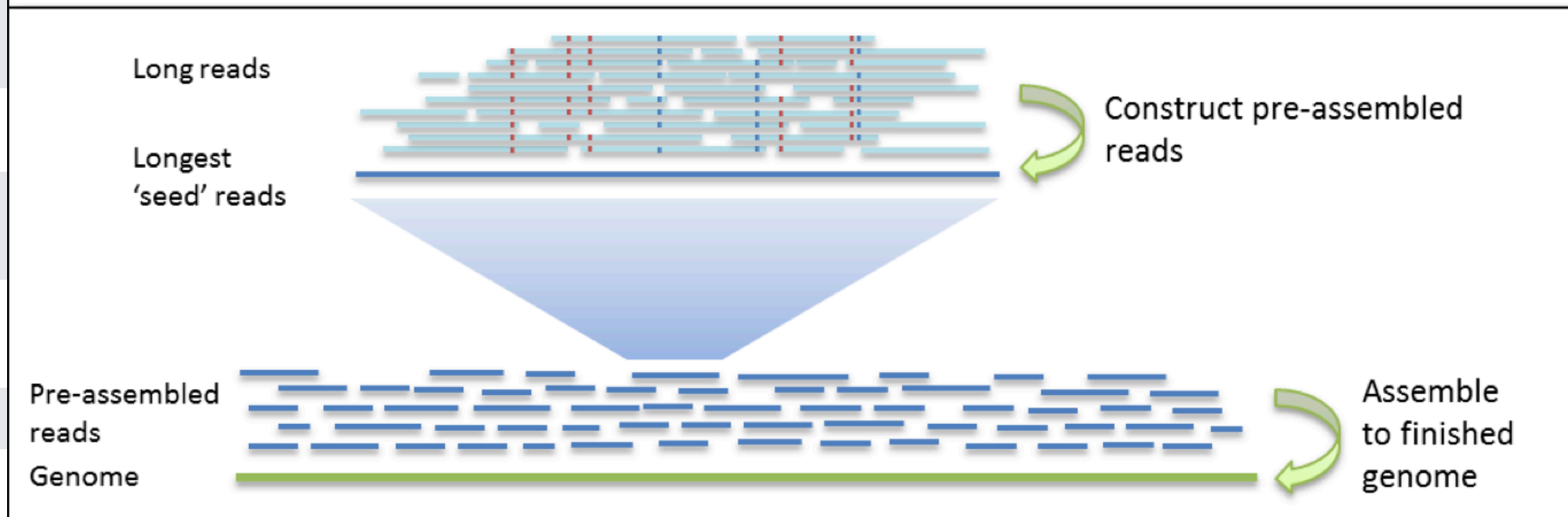
ONT raw data

BAC sequences

New Cassava project

Data	fastq/fastq
Program	CANU/Flye/ WTDBG2
Time	3 - 28 days
Storage	50 Gb -6 T
RAM	512 Gb
CPU	60 cpus
Computer	HPC Dragon2/Lm3/ Nic4/Durandal

ONT assembly



New Cassava project

Data	fastq/fastq
Program	CANU/Flye/ WTDBG2
Time	3 - 28 days
Storage	50 Gb -6 T
RAM	512 Gb
CPU	60 cpus
Computer	HPC Dragon2/Lm3/ Nic4/Durandal

750 000 CPU hours

ONT assembly

WTDBG2

Reads Min size 0 Kb
Overlap length 2 kb
#Overlapping reads 3

Reads Min size 3 Kb
Overlap length 2 kb
#Overlapping reads 3

Reads Min size 3 Kb
Overlap length 3 kb
#Overlapping reads 3

Reads Min size 5 Kb
Overlap length 2 kb
#Overlapping reads 3

Reads Min size 5 Kb
Overlap length 3 kb
#Overlapping reads 3

3 days

Flye

Reads Min size 0 Kb
#Reads polishing 1

Reads Min size 0 Kb
#Reads polishing 2

Reads Min size 5 Kb
#Reads polishing 1

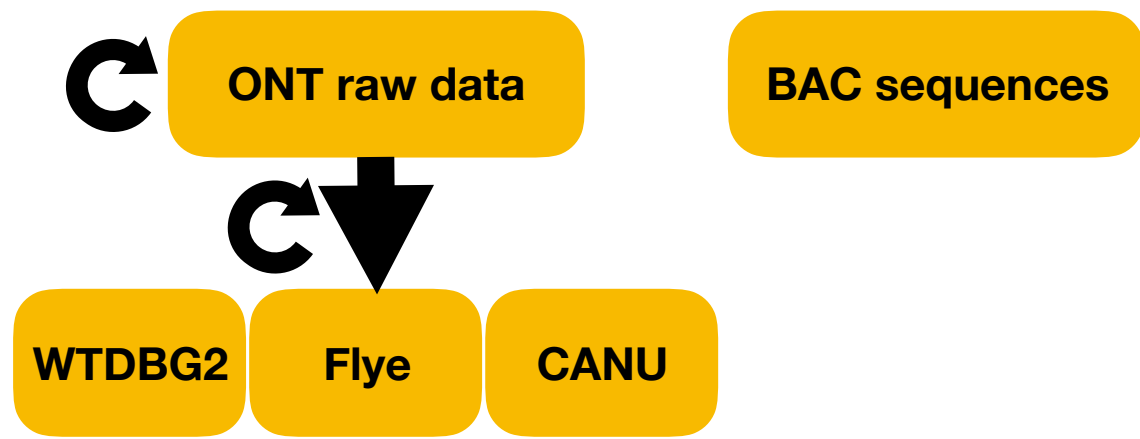
Reads Min size 8 Kb
#Reads polishing 1

7 days

CANU

Reads Min size 8 Kb
#Default values

28 days



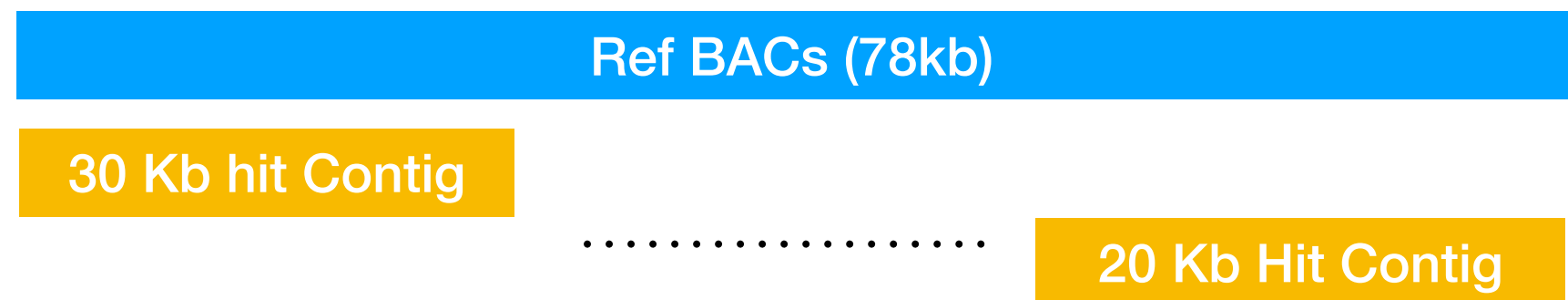
New Cassava project

Data	plots/txt
Program	Nucmer/ BLAST/BLASR
Time	3 days
Storage	NA
RAM	512 Gb
CPU	60 cpus
Computer	HPC

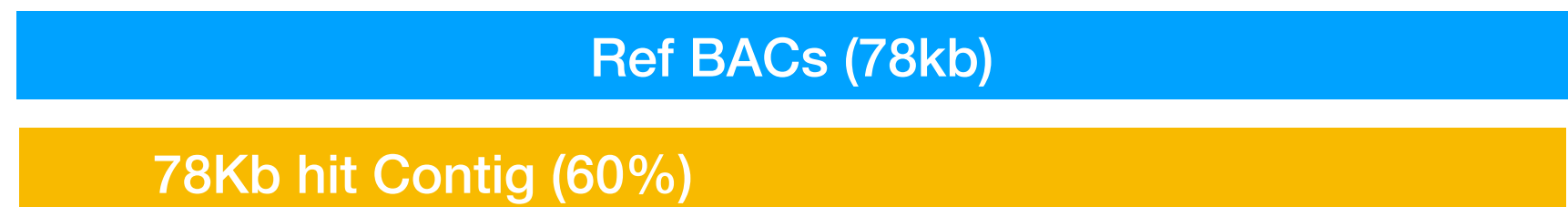
BAC mapping

Basic Local Alignment Search Tool
BLAST

← **AGGT** →

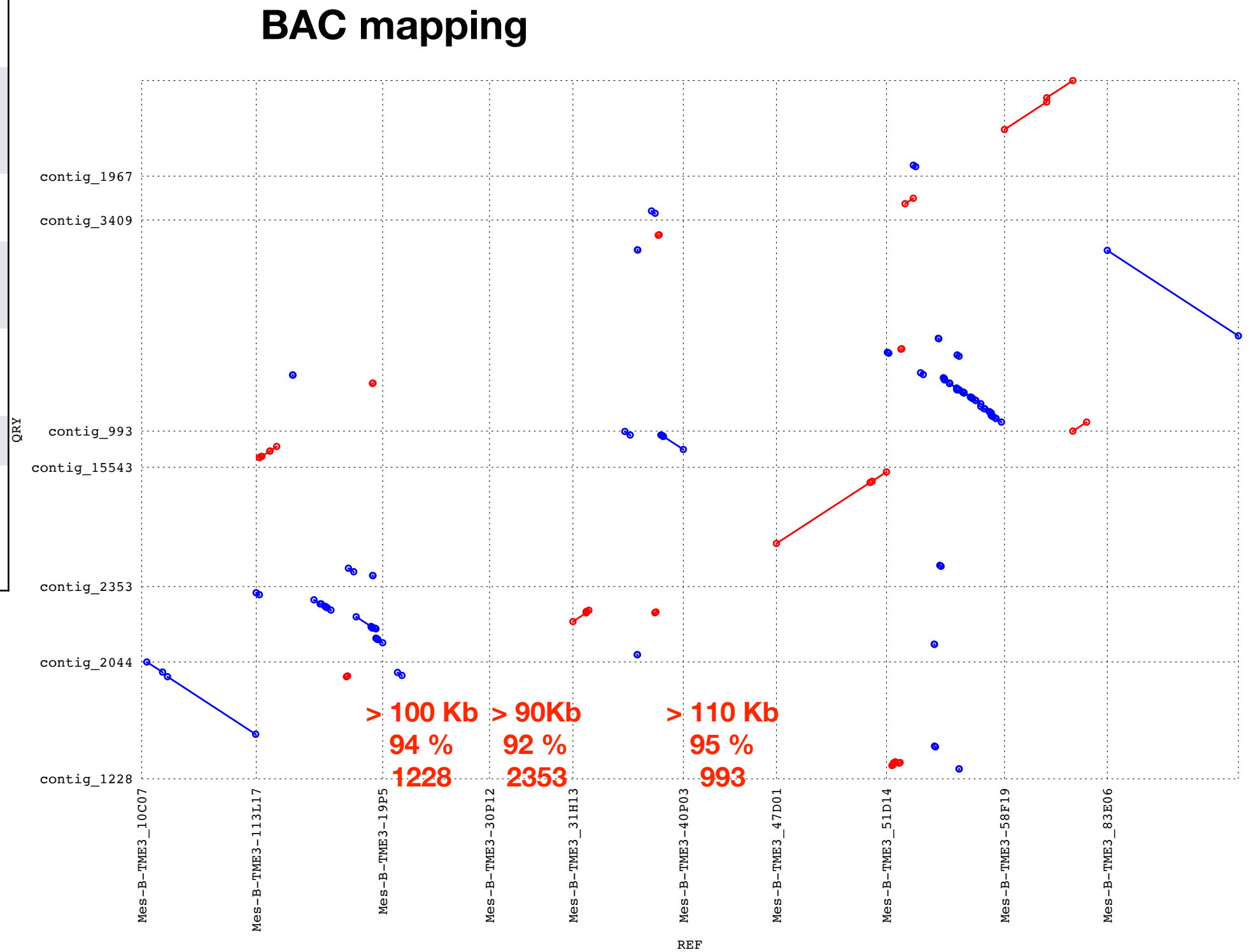


Basic Local Alignment with Successive Refinement
BLASR



New Cassava project

Data	plots/txt
Program	Nucmer/ BLAST/BLASR
Time	3 days
Storage	NA
RAM	512 Gb
CPU	60 cpus
Computer	HPC



New Cassava project

Data	plots/txt
Program	Nucmer/ BLAST/BLASR
Time	3 days
Storage	NA
RAM	512 Gb
CPU	60 cpus
Computer	HPC

BAC mapping

Zurich

Breakpoints = 1108

Insertions = 253

Deletions = 146

Bac = 0/10

Id = 96.52 %

Gembloux

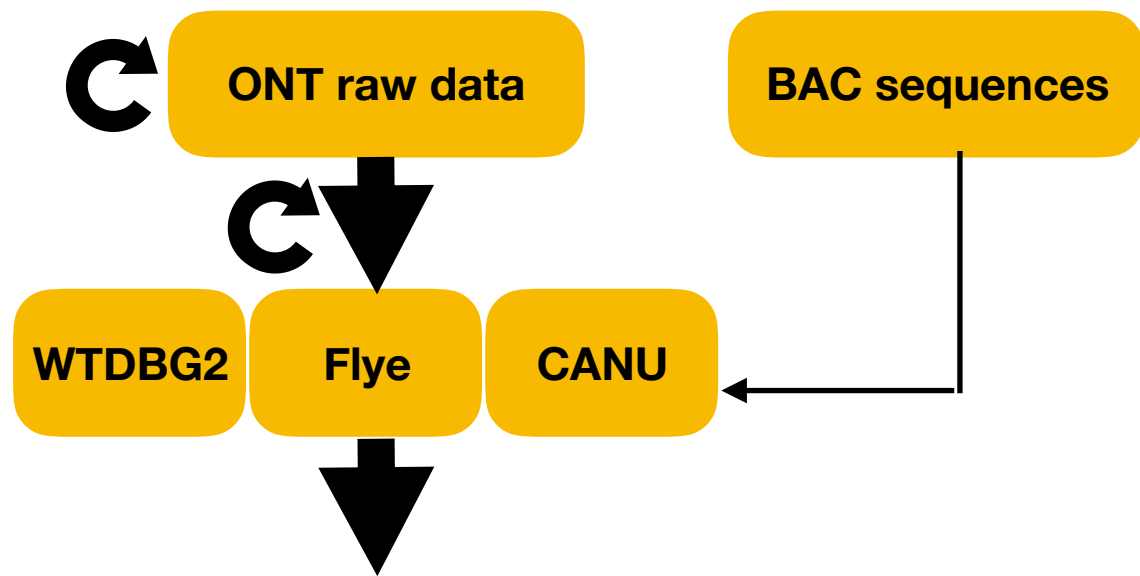
Breakpoints = 294

Insertions = 66

Deletions = 59

Bac = 7/10

Id = 93.61%

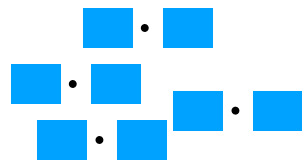


New Cassava project

Data	fasta/fastq
Program	Pilon
Time	7 days
Storage	35 Gb
RAM	512 Gb
CPU	32 cpus
Computer	HPC

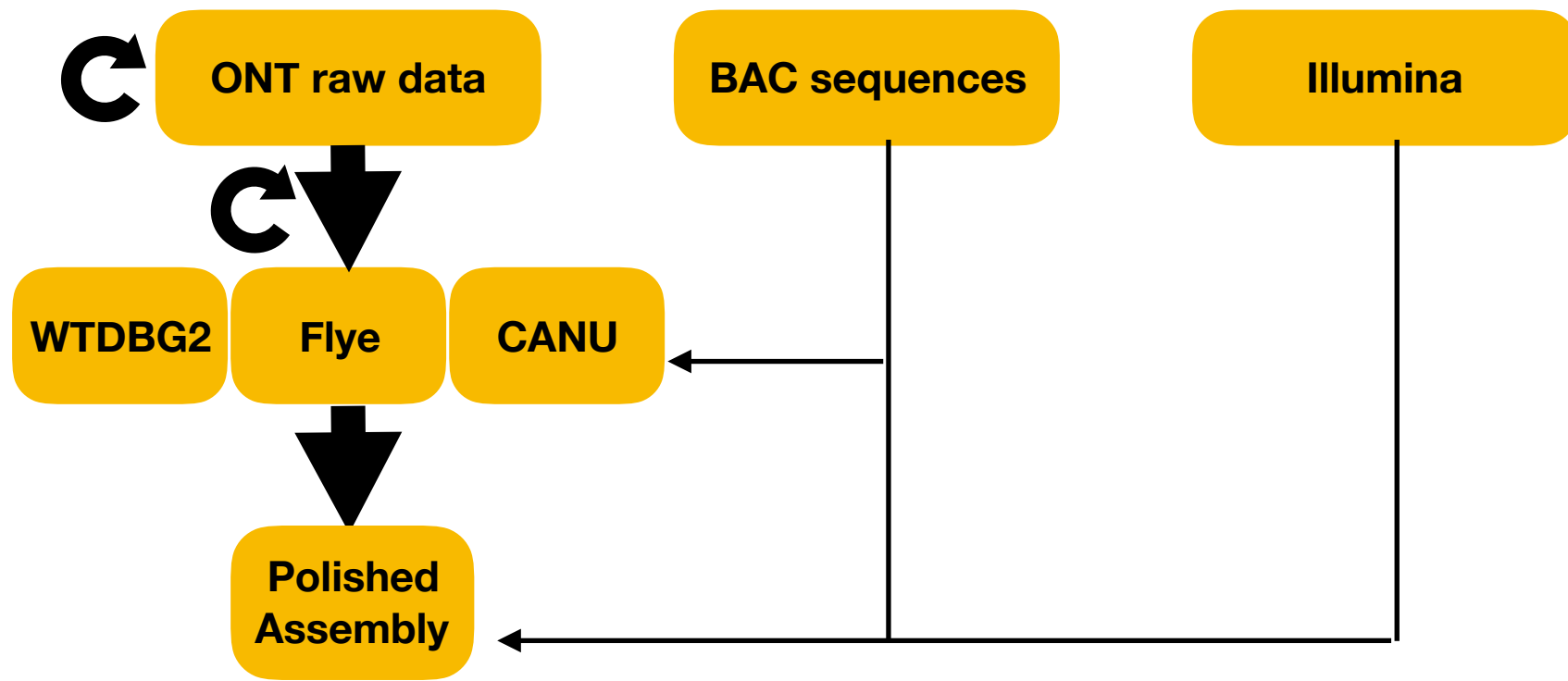
Illumina polishing

Flye contigs (94%)



Illumina (250 fold)

Flye contigs (97%)



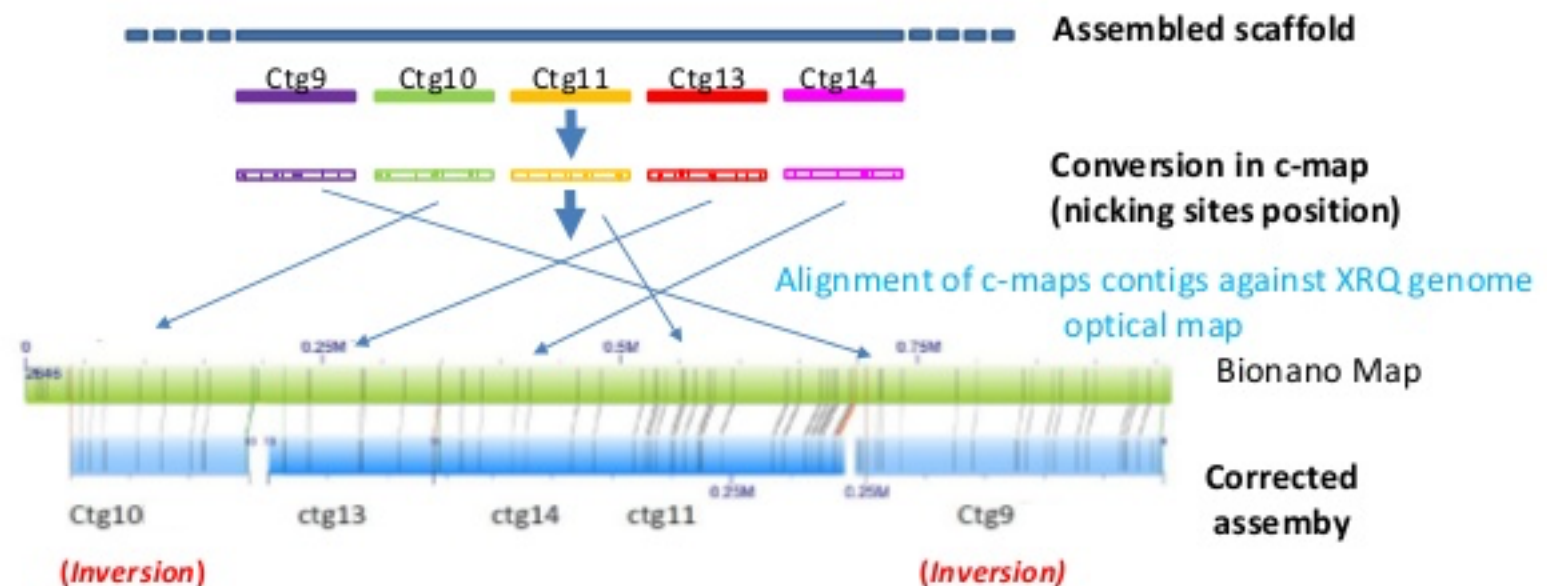
New Cassava project

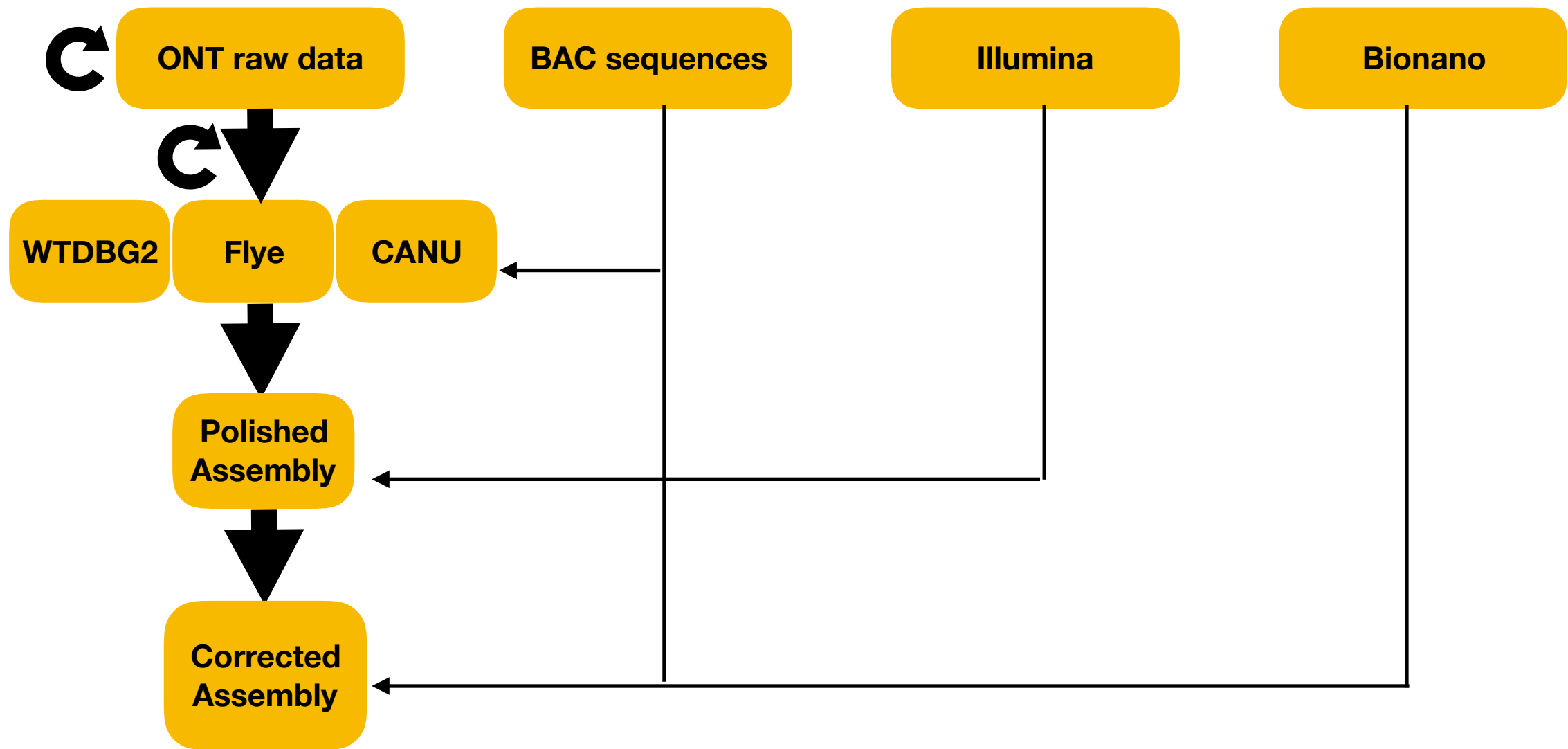
Data	fasta/cmap
Program	Bionano camp
Time	7 days
Storage	250 Gb
RAM	1T
CPU	60 cpus
Computer	AWS

Optical map

Optical maps to solve conflicts in the assembly

Alignment of the contig against the BioNano assembly of XRQ genome

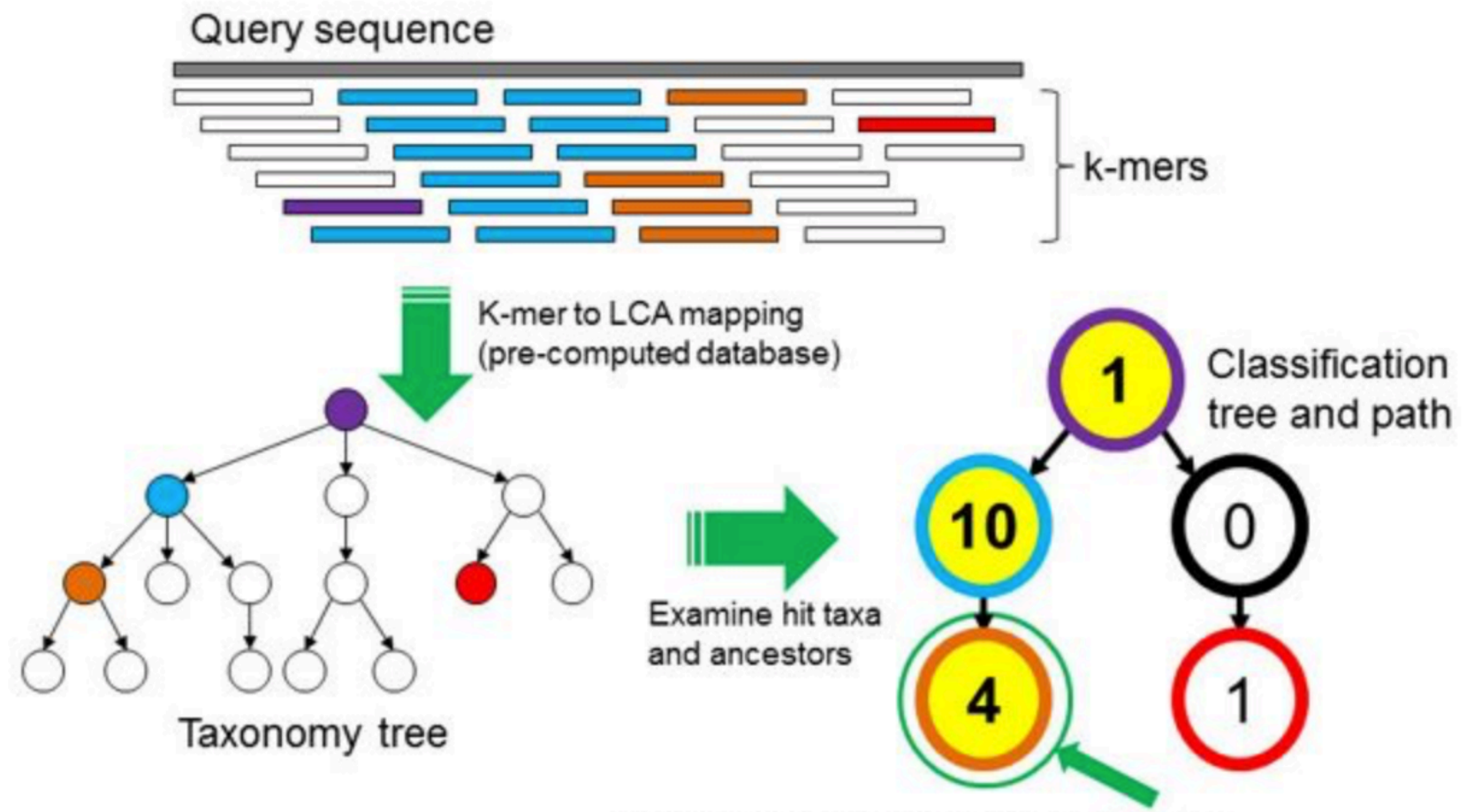


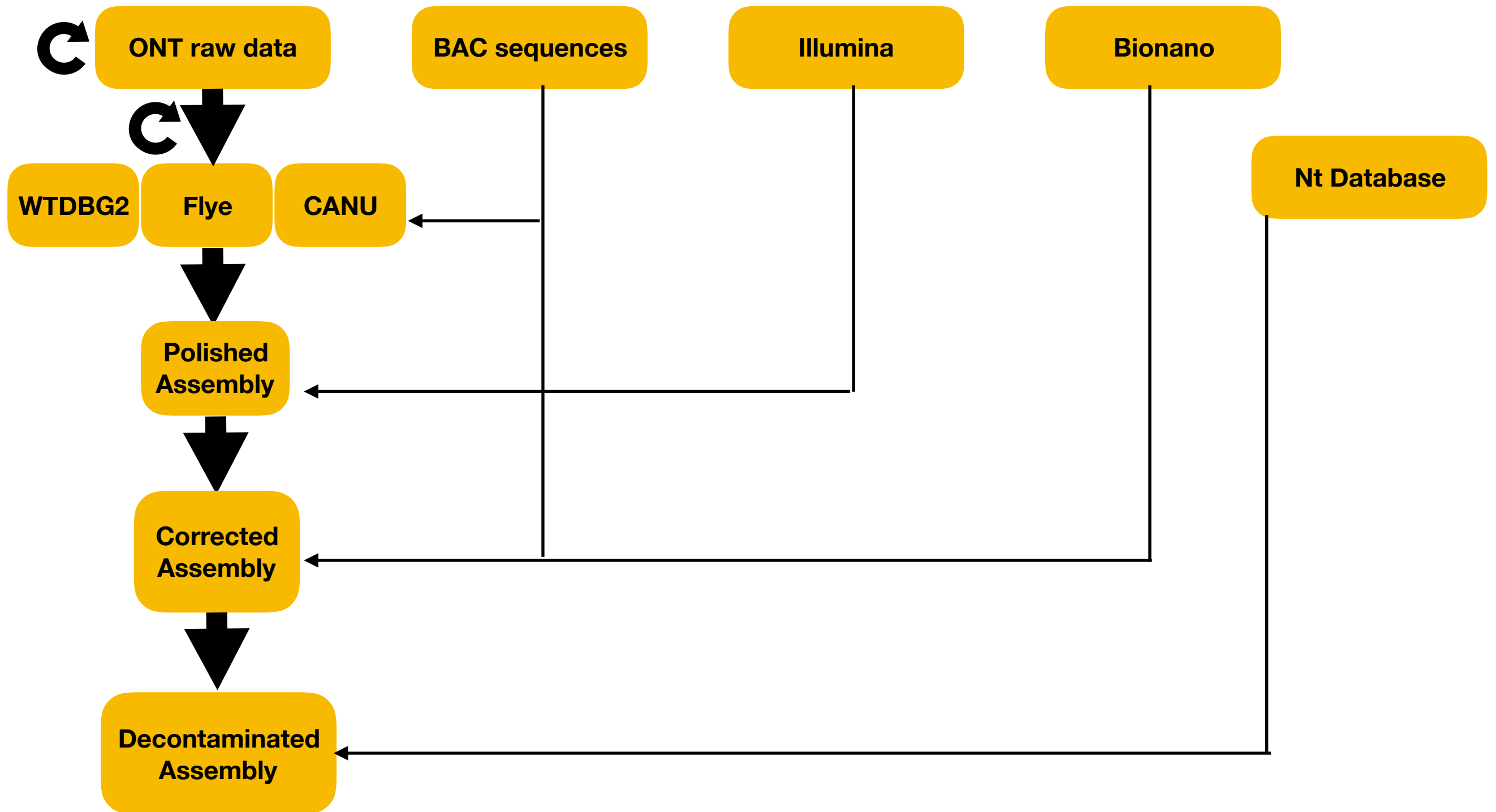


New Cassava project

Data	Blast report
Program	Kraken
Time	4 days
Storage	300 Gb
RAM	256 Gb
CPU	32
Computer	HPC

Contamination check

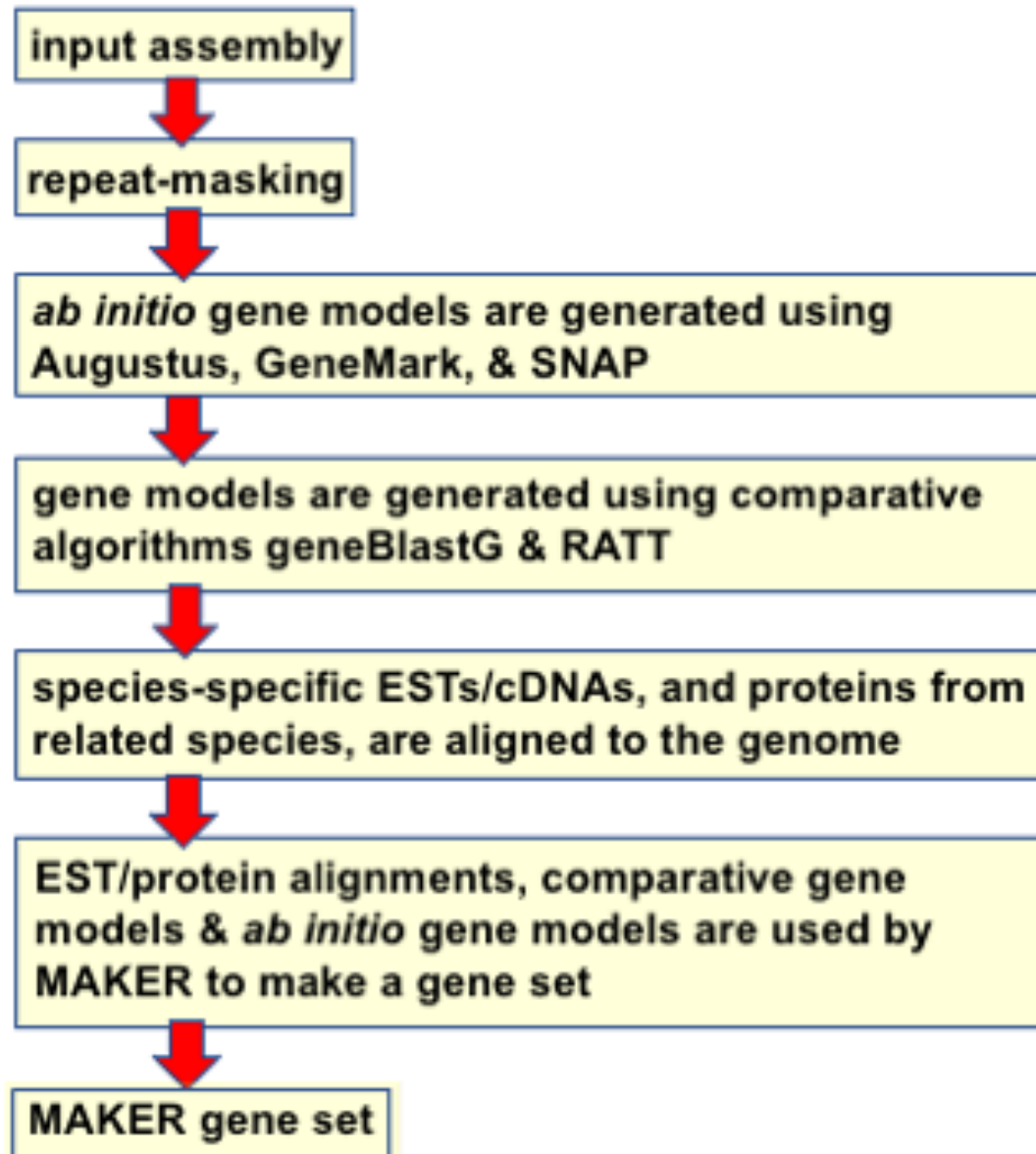


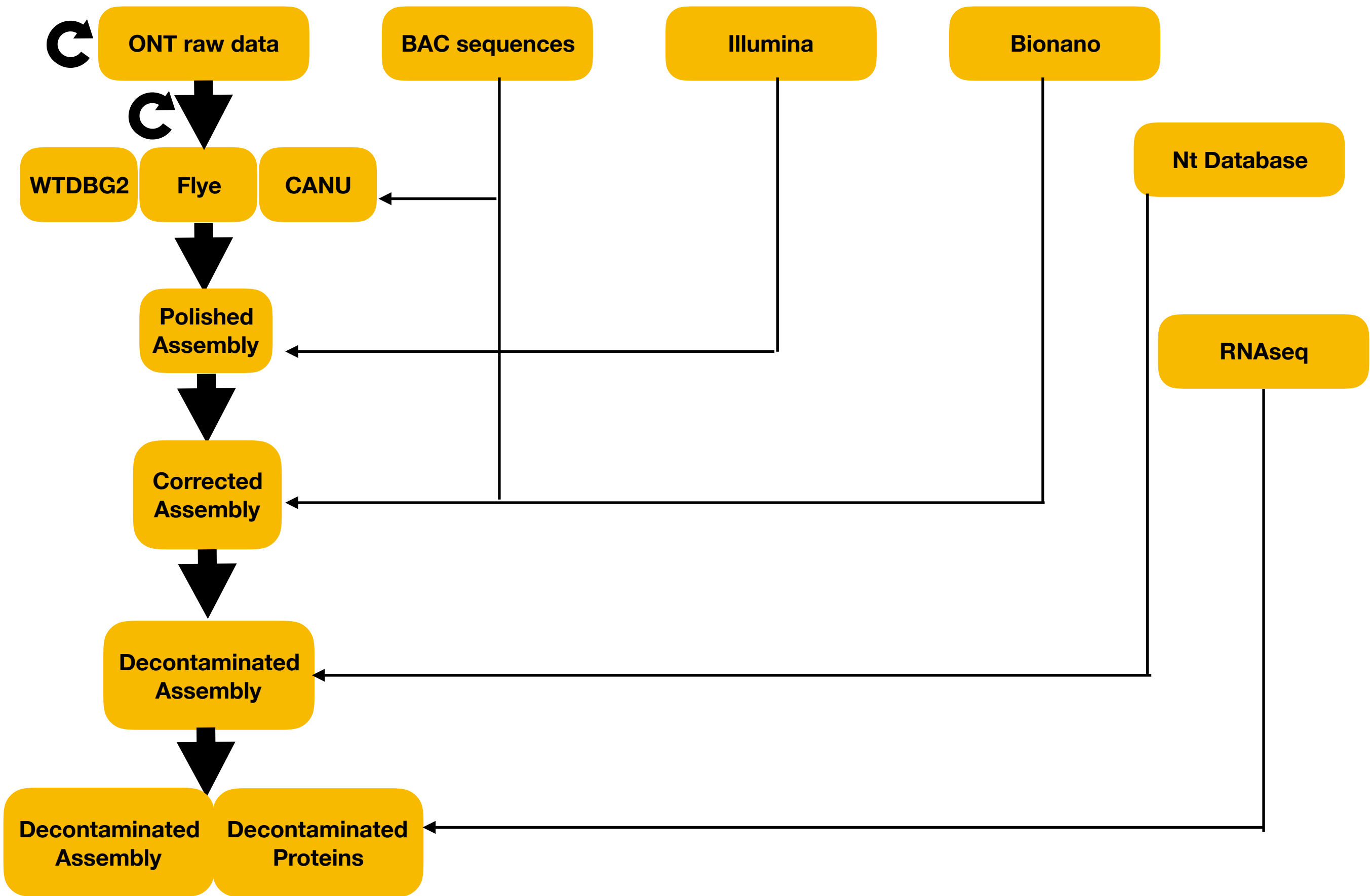


New Cassava project

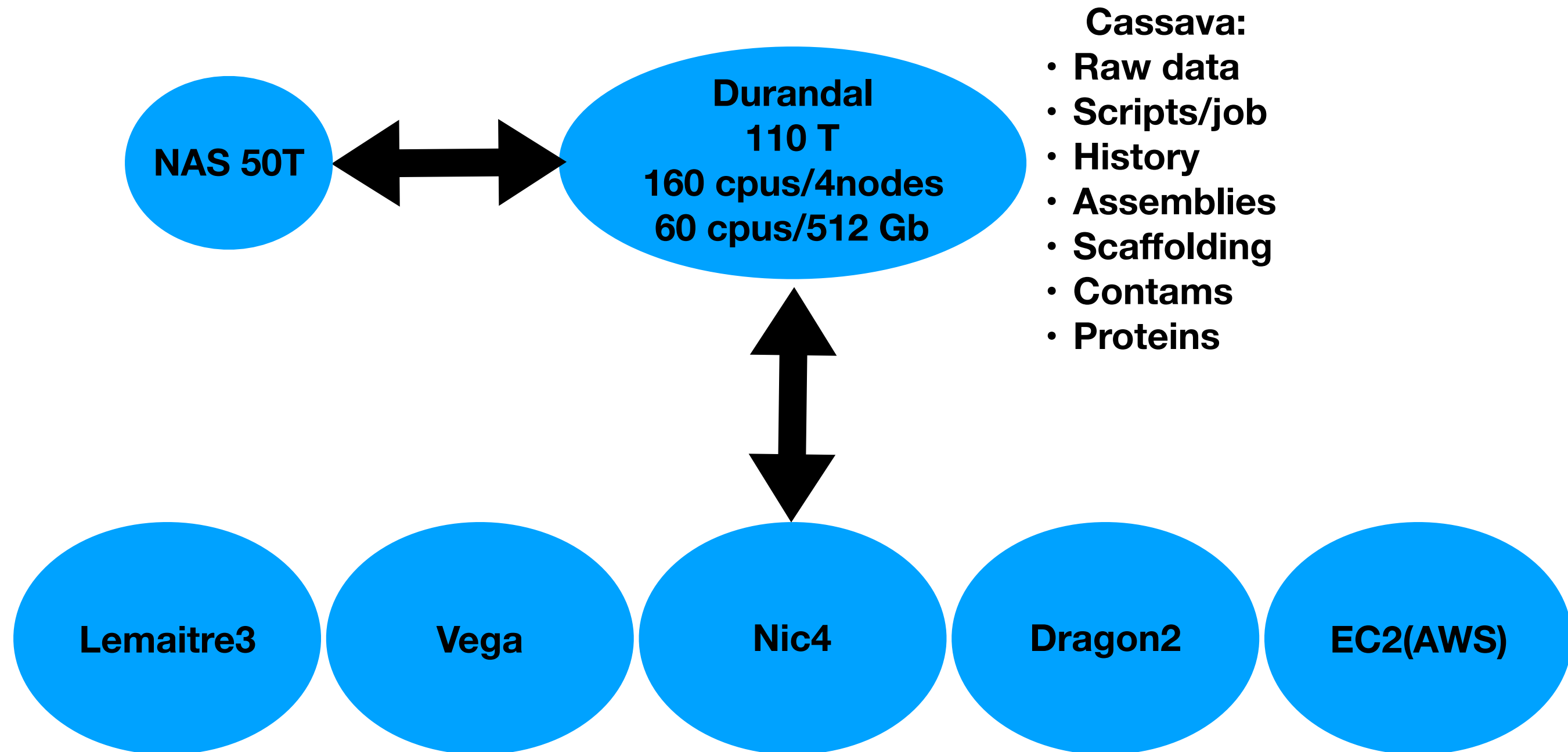
Data	Fasta
Program	MAKER
Time	30 days
Storage	500 Gb
RAM	256 Gb
CPU	160
Computer	HPC

Protein prediction

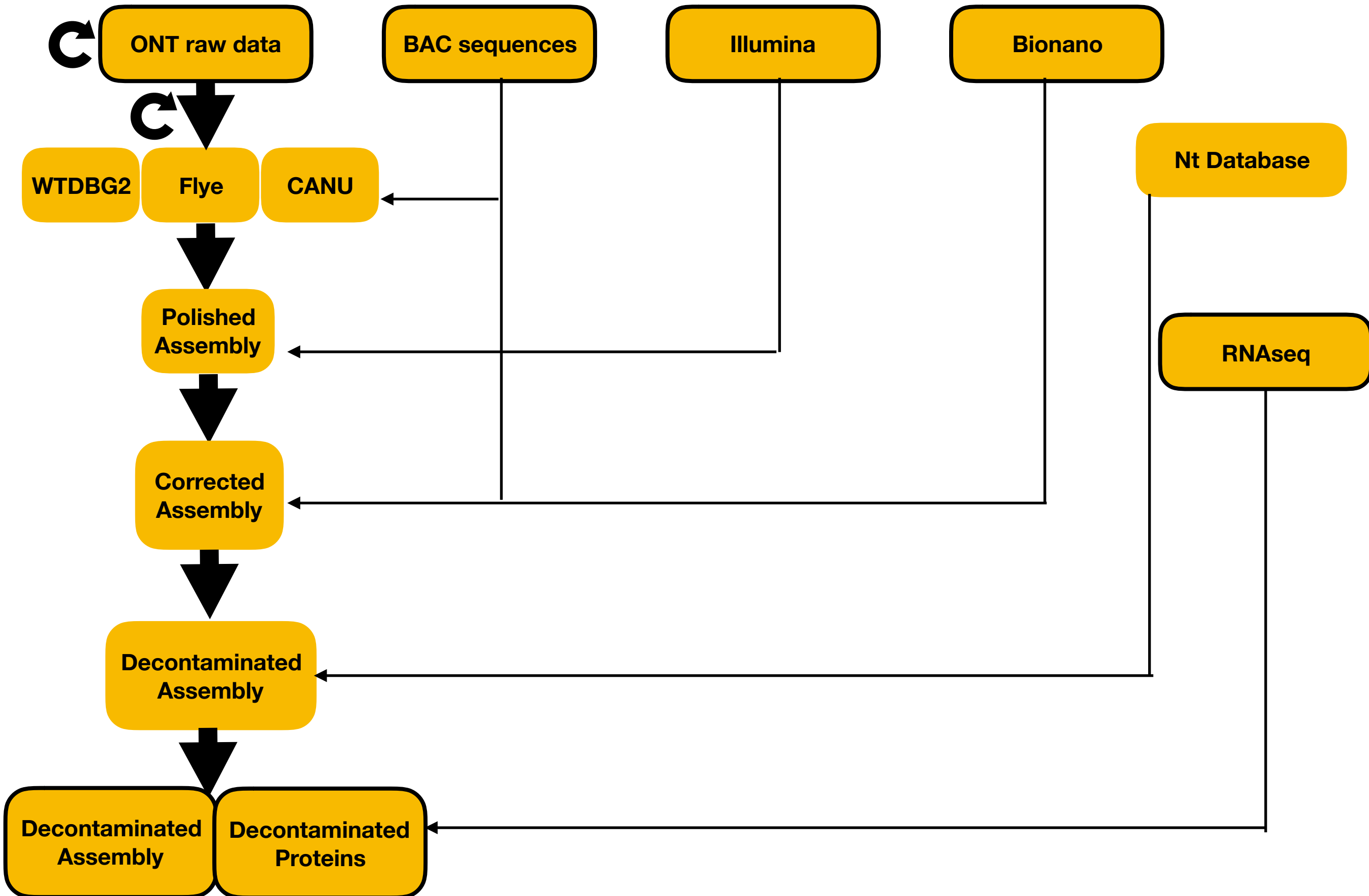




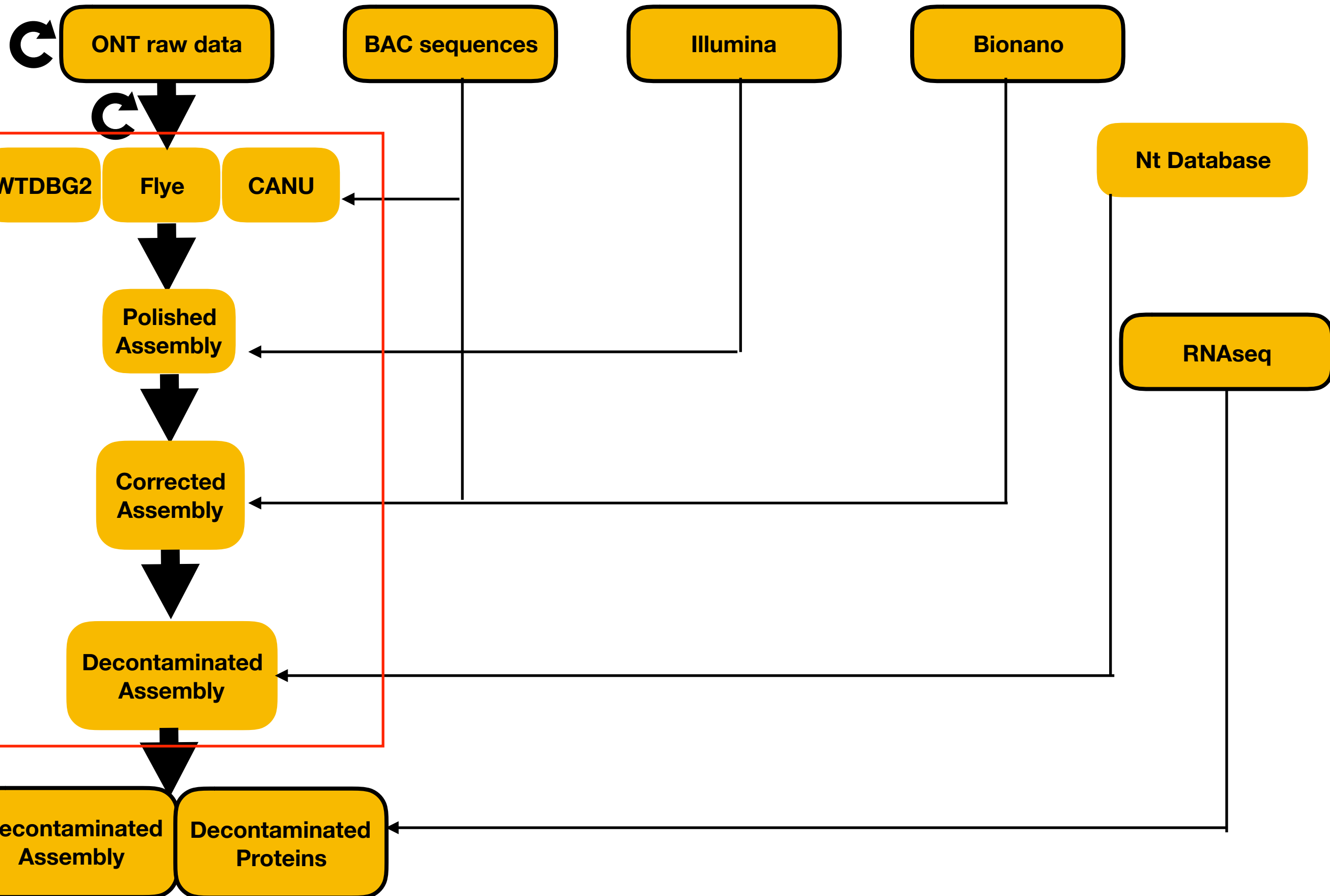
Organisation



Project end

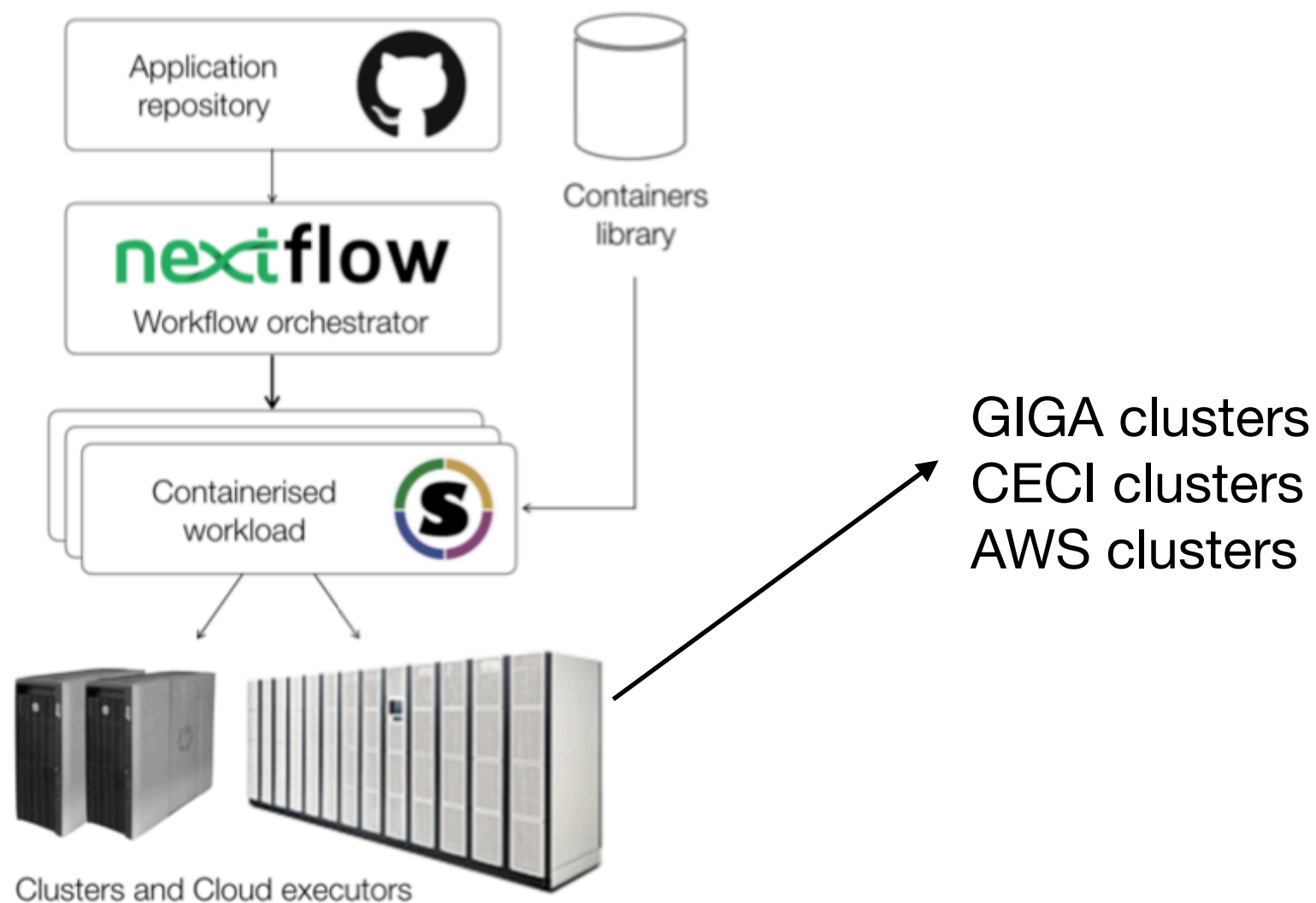


Project end



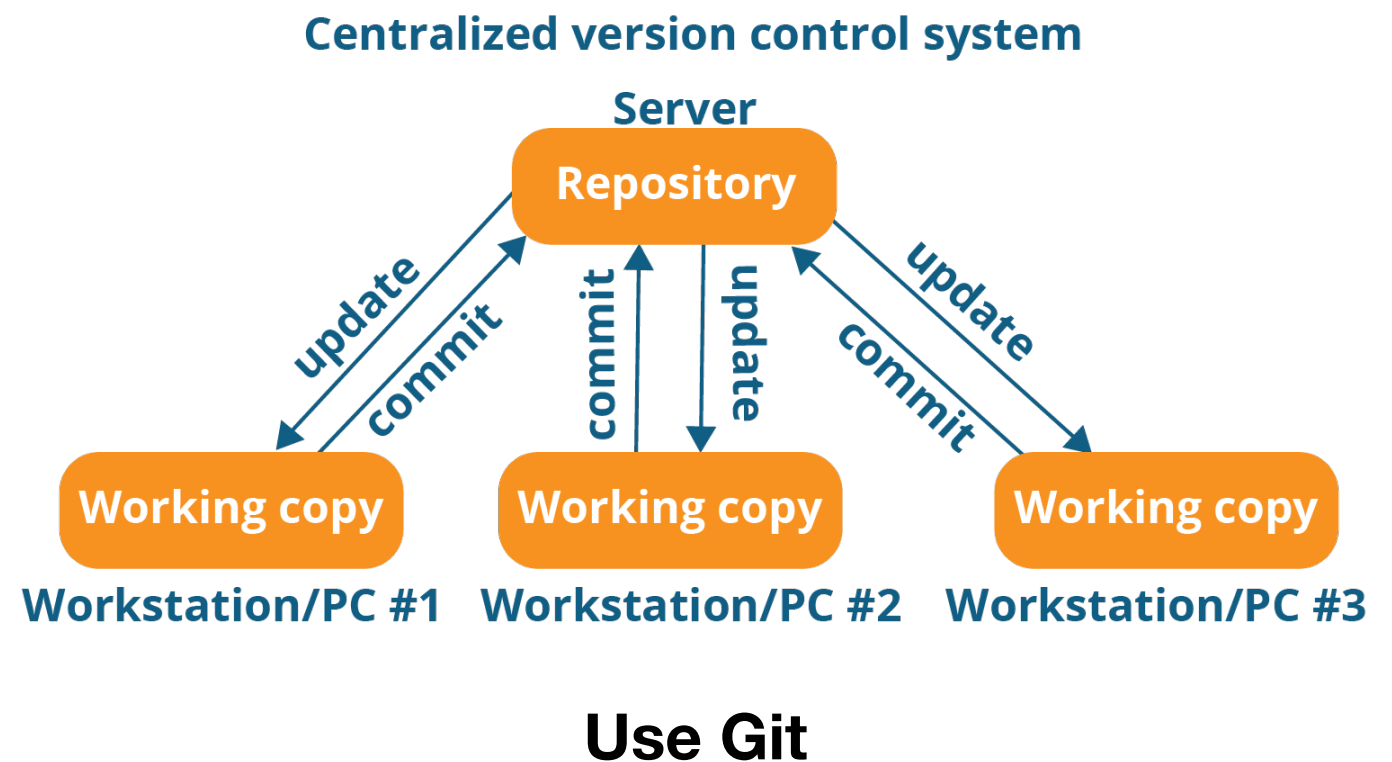
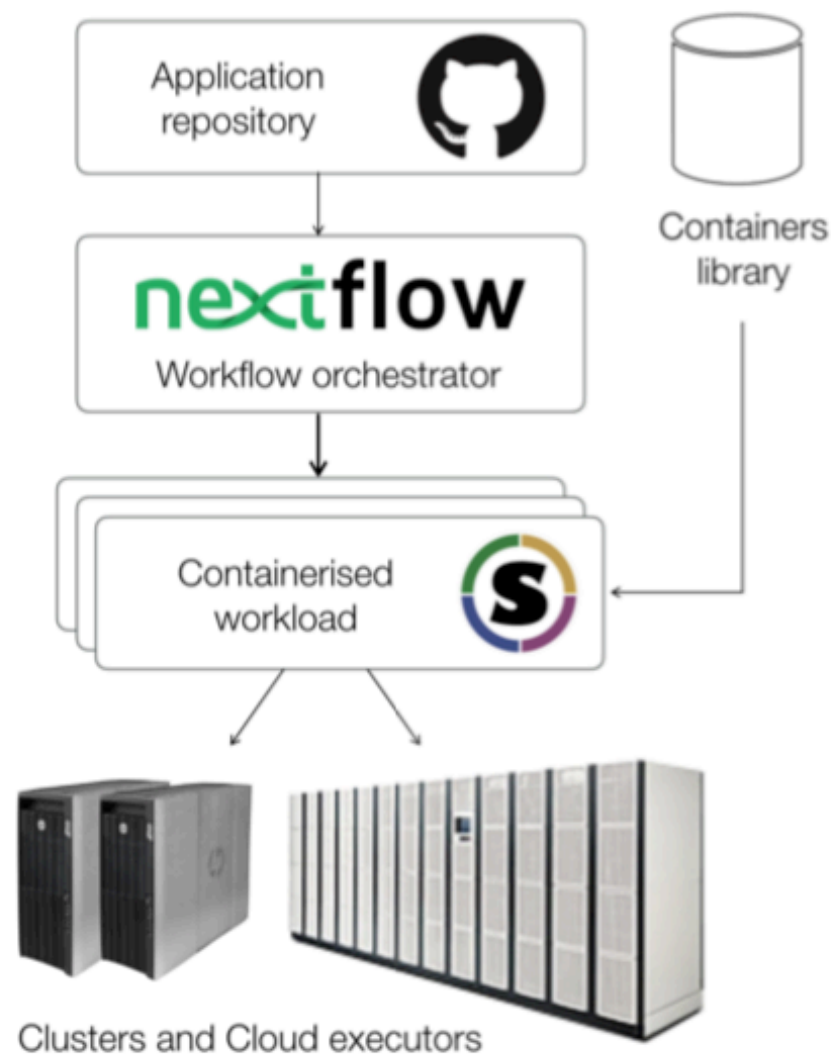
Data management

- Reproducibility - Portability - Collaborations



Data management

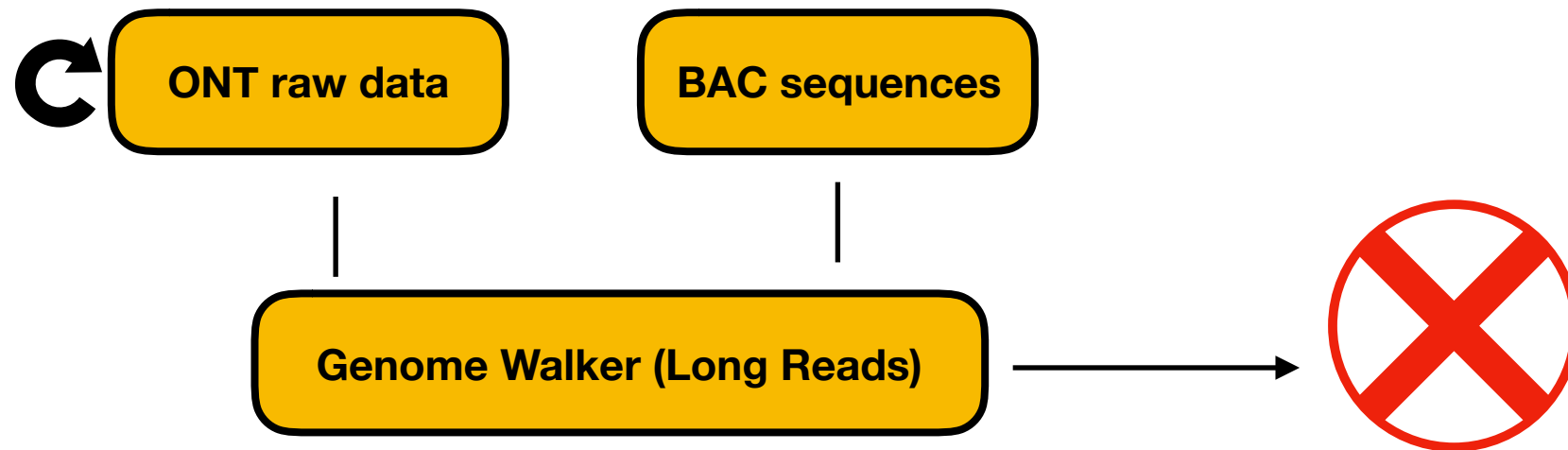
- Reproducibility - Portability - Collaborations



<https://github.com/nextflow-io/awesome-nextflow>

Best practice

- **Reproducibility**
- **Code versioning**
- **Identify repetitive tasks**



Reference

Reference



1. Reads mapping

Reference



2. Alignment creation (Minimap 2)

Reference

C



Used for

3. Consensus - fasta reader

