

Data Handling Pipelines

Mario Caccamo
Norwich, UK



The Genome Analysis Centre (TGAC)

A new facility to provide critical mass and excellence in genomics specialised in **animal, microbial** and **plant** research:

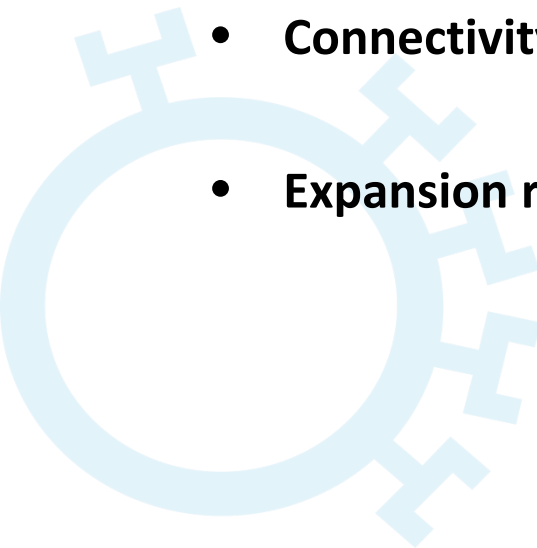
- high throughput sequencing
- new technology platforms
- **bioinformatics**
- impact through innovation and enterprise

TGAC will complement work being carried out at The Wellcome Trust Sanger Institute and MRC / NERC Centres.

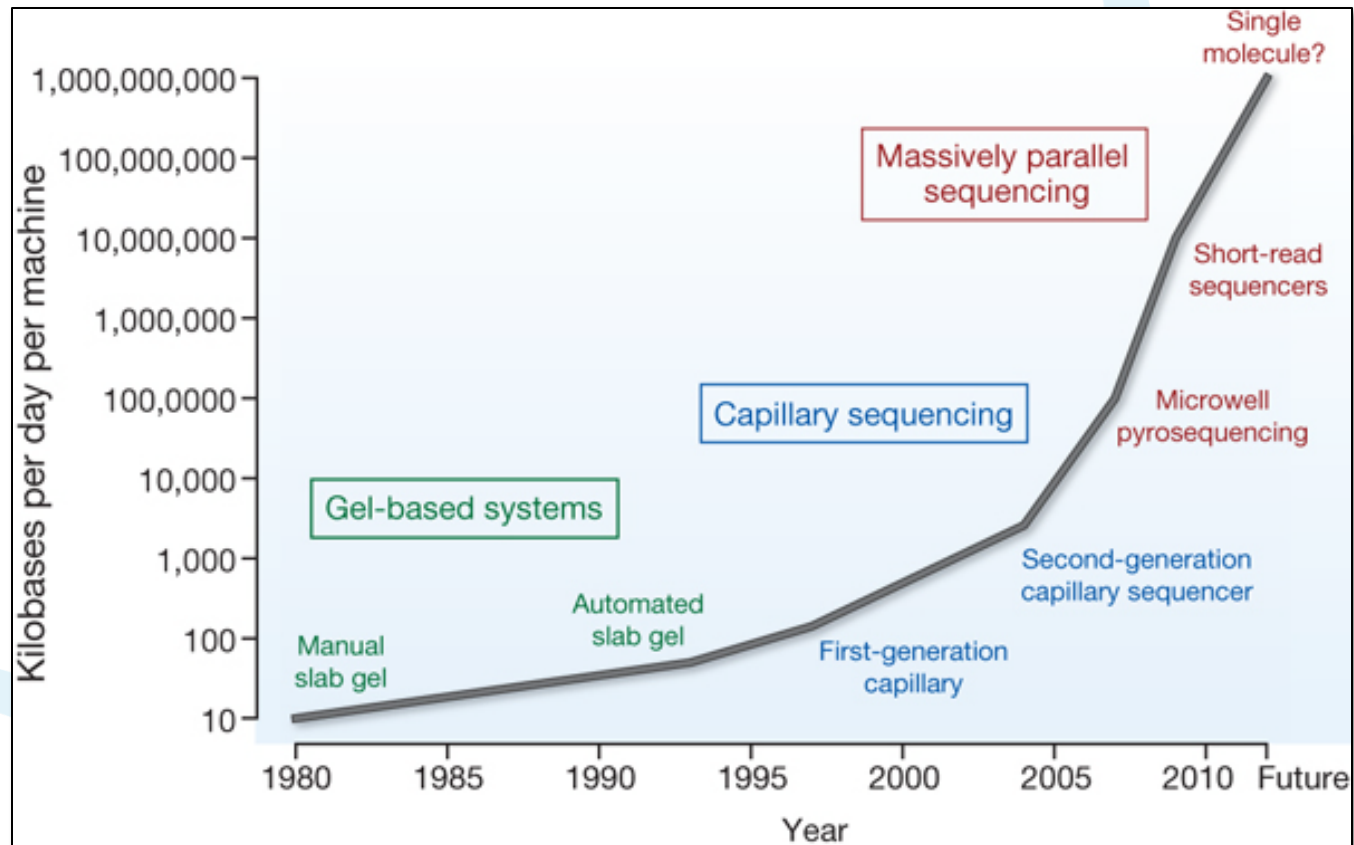


Considerations

- **Space**
- **Electric Consumption**
- **Storage / HPC cluster**
- **Staff**
- **Connectivity**
- **Expansion rate**



Sequencing Technologies



Michael R. Stratton, Peter J. Campbell & P. Andrew Futreal
Nature **458**, 719-724(9 April 2009)

Heavy Weights



HiSeq 2000

75-120 bases reads

~200 Gb / expt

~8 days



AB SOLiD 4 System

50 bases reads

~100 Gb / expt (300Gb 4hq)

~10 days



Light Weights



SOLiD PI System



GS Junior

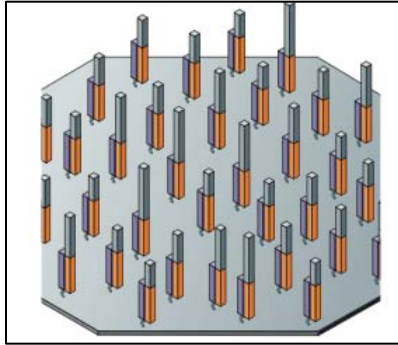


Ion Torrent

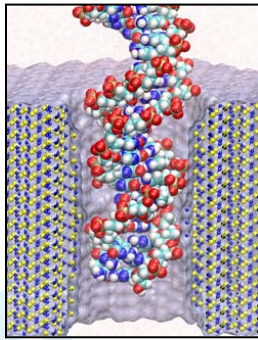


iSCAN

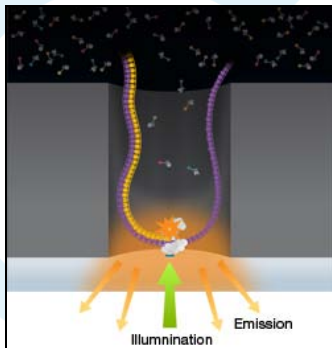
Single-Molecule Sequencing



Single molecule: primer immobilized

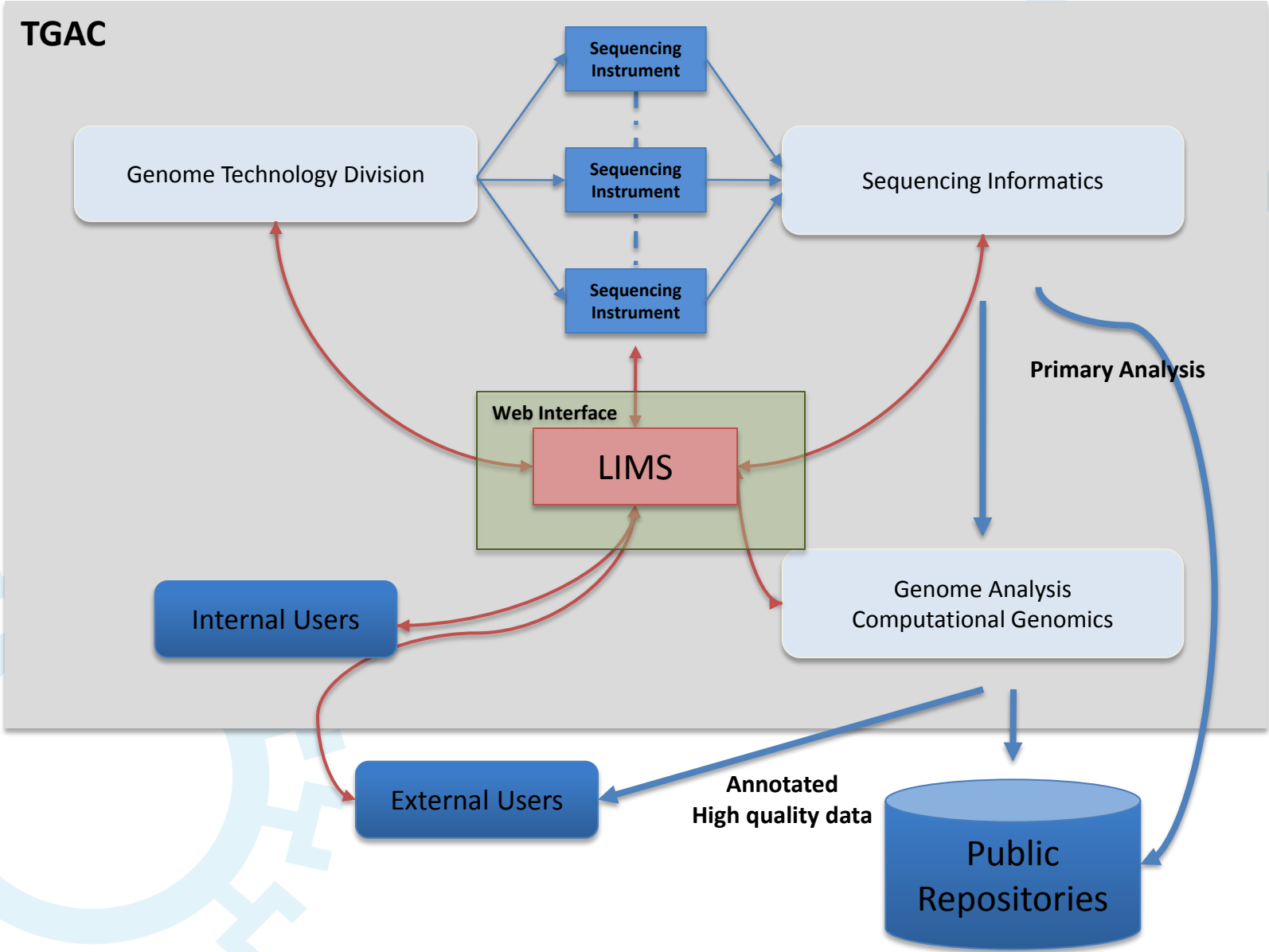


DNA sequence detection as molecules pass through a nm - sized pore

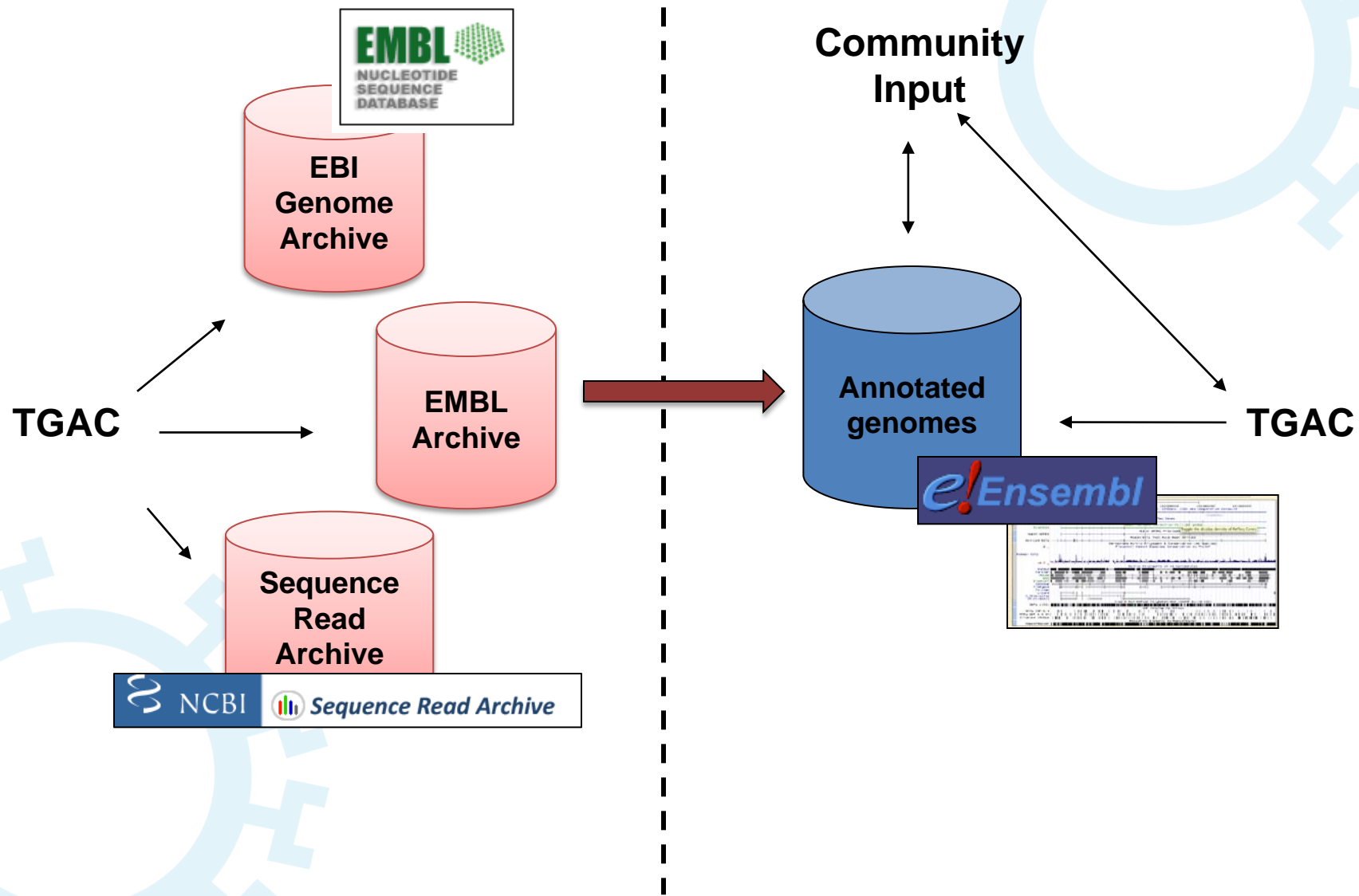


Single molecule sequencing by enzyme tethered in 20 nm hole

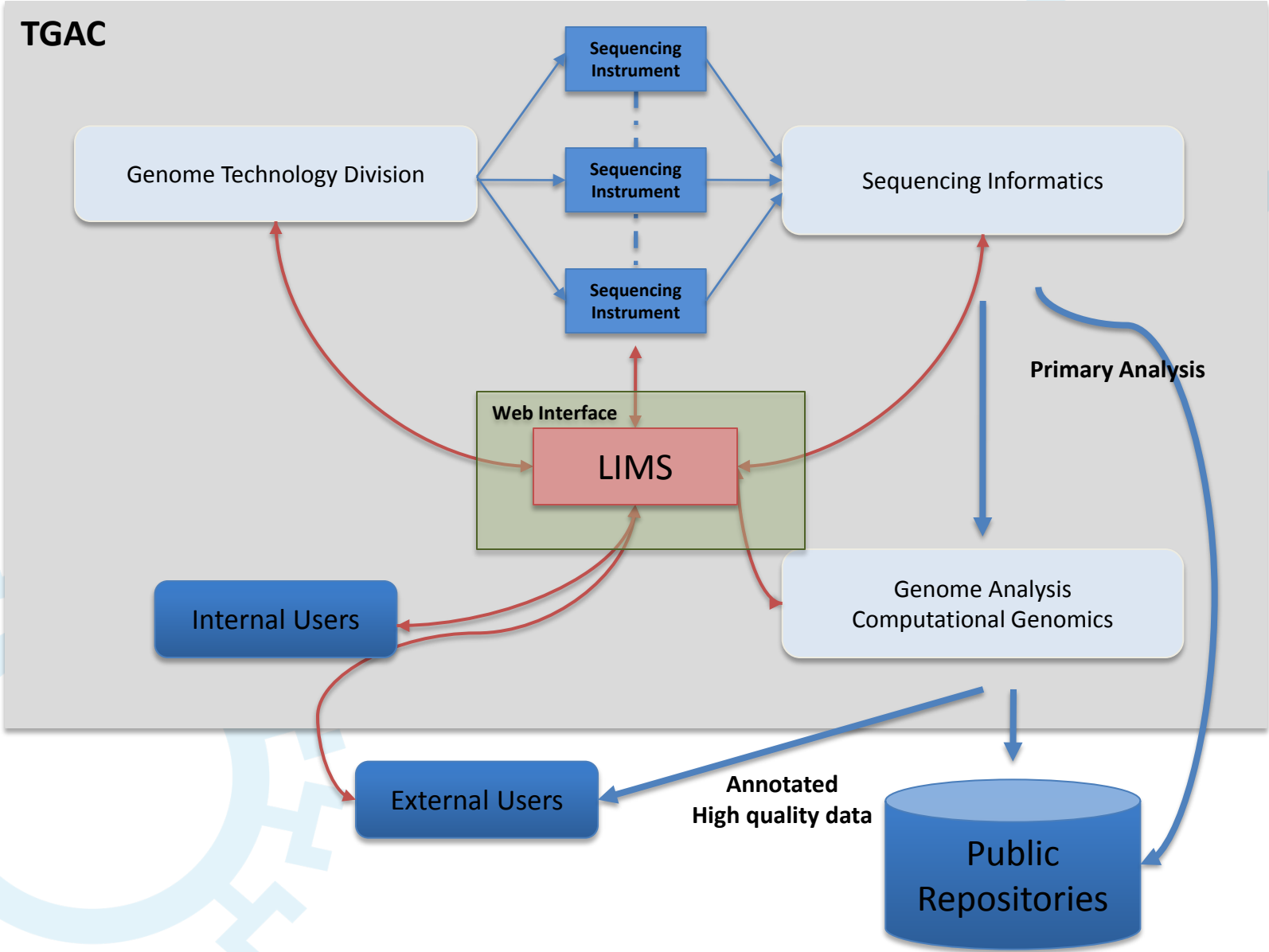
TGAC



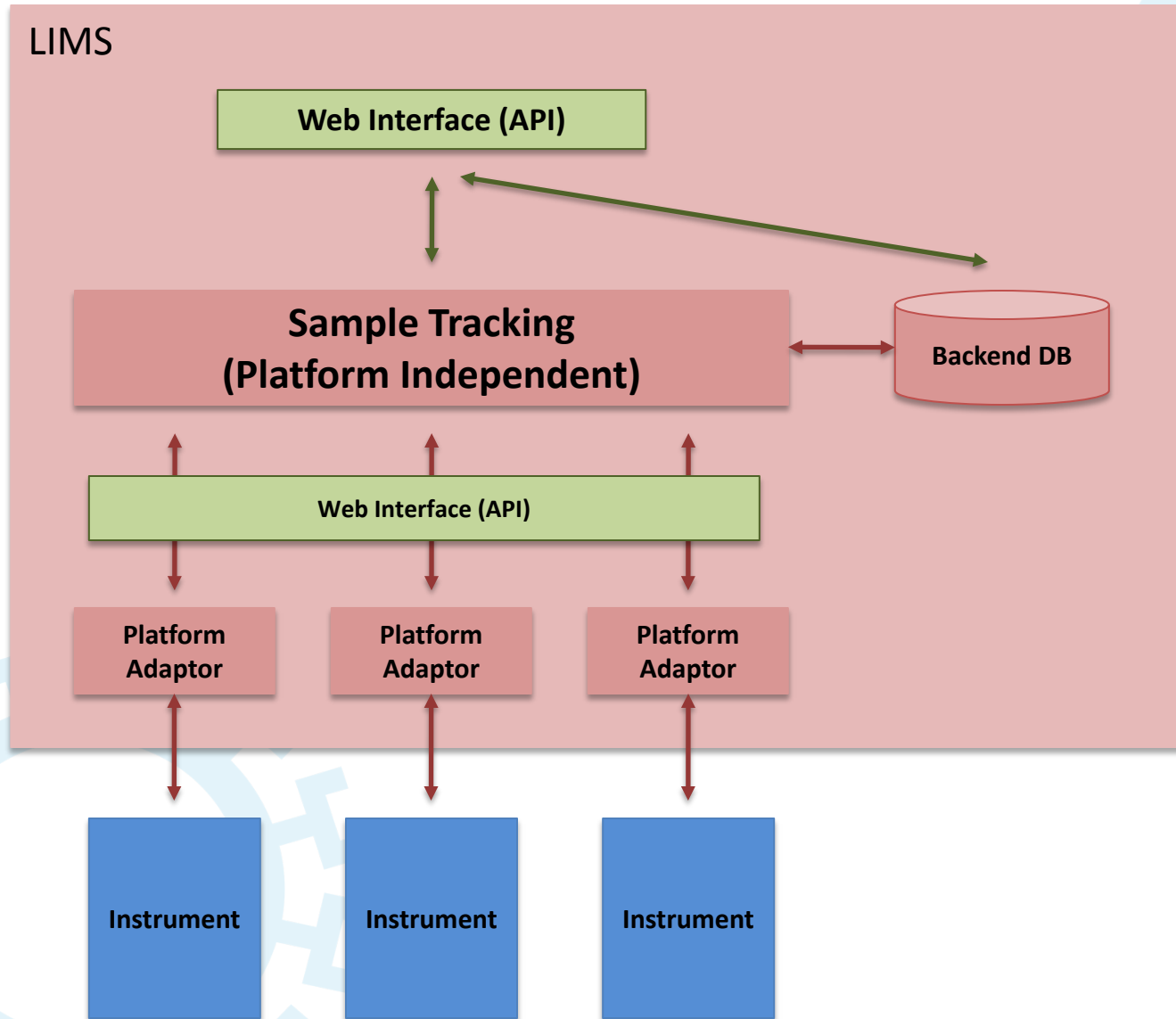
Repositories



TGAC

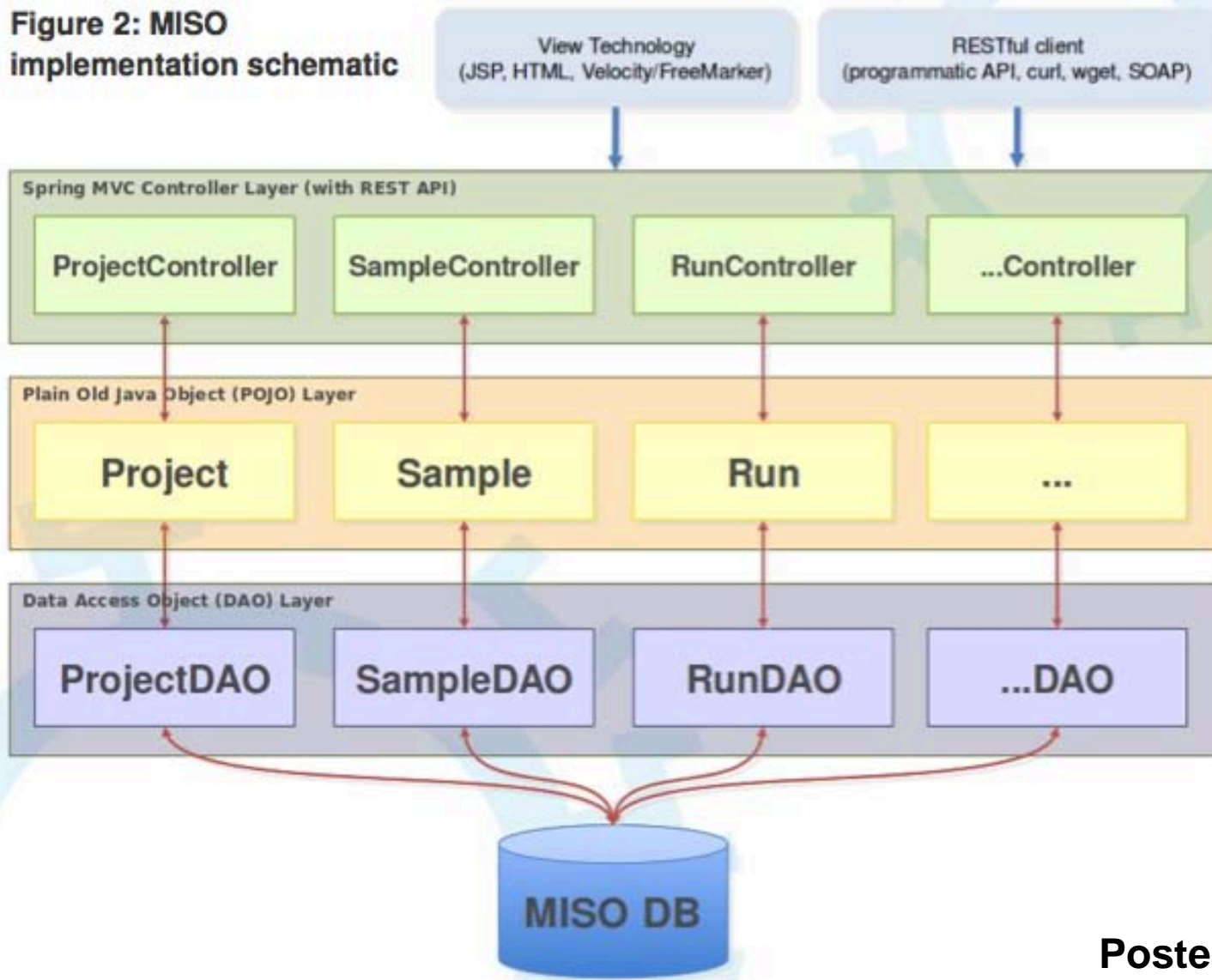


LIMS Architecture



MISO

Figure 2: MISO implementation schematic



Search for...

Edit Study

A study contains information about the sequencing project. Studies can contain any number of sequencing experiments and analysis.

[Get Study Submission Data](#)

Study ID: 1

Project ID: 1

Name: STU1

Accession: [Help](#)

Description: [Help](#)

Study Type:

Permissions *Inherited from project*

Experiment ID	Experiment Name
1	EXP1 <input type="button" value="i"/>

- Other
- RNASeq
- Population Genomics
- Cancer Genomics
- Gene Regulation Study
- Forensic or Paleo-genomics
- Synthetic Genomics
- Epigenetics
- Resequencing
- Transcriptome Analysis
- Metagenomics
- Whole Genome Sequencing

Experiments

[Add new Experiment](#)

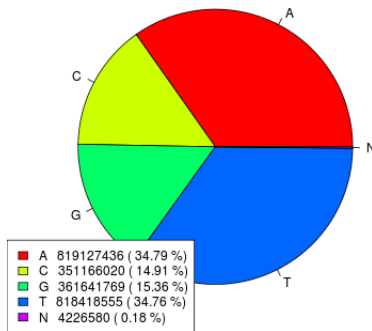
Experiment ID	Experiment Name		Edit	List Sample	List Run
1	EXP1	<input type="button" value="v"/>	Edit	List Sample	List Run

Run Stats

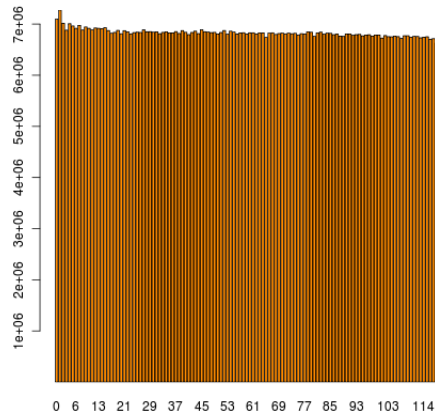


Clostridium botulinum

Nucleotide composition

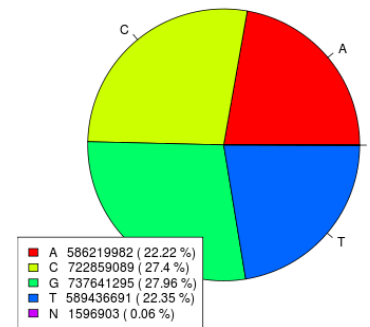


Count in position for nucleotide A

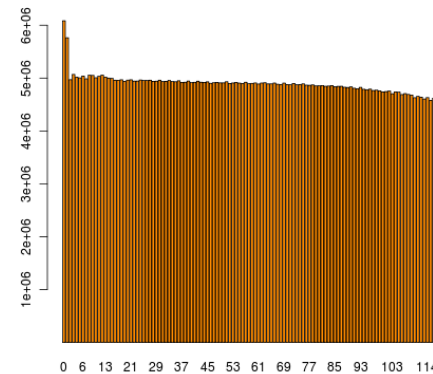


Ralstonia faecalis

Nucleotide composition

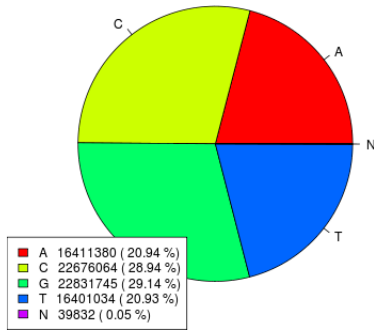


Count in position for nucleotide A

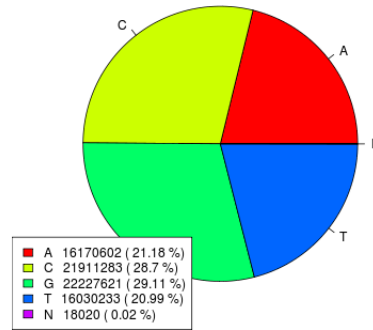


Run Stats

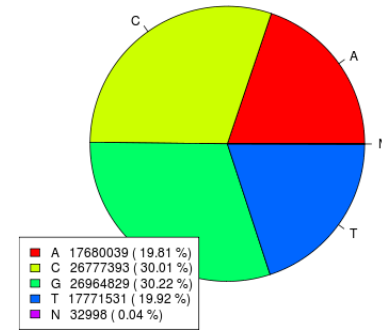
Nucleotide composition



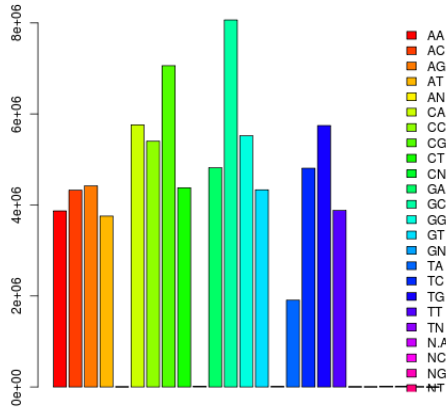
Nucleotide composition



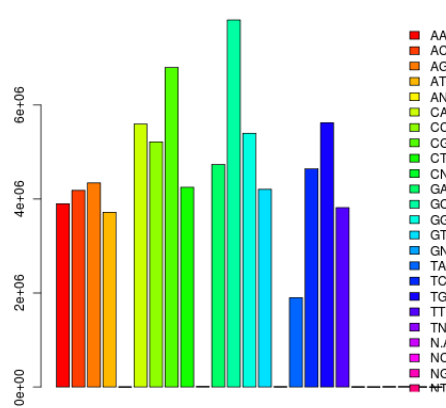
Nucleotide composition



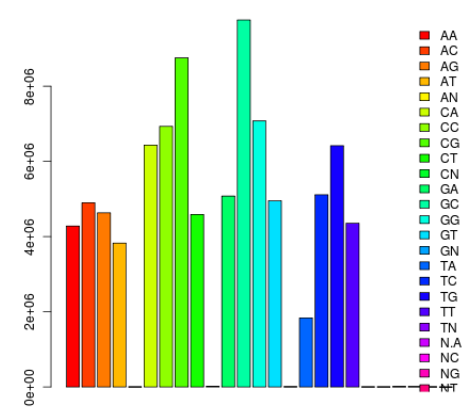
Dinucleotide composition



Dinucleotide composition

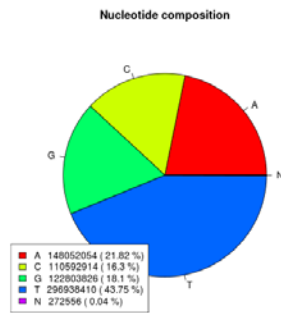


Dinucleotide composition

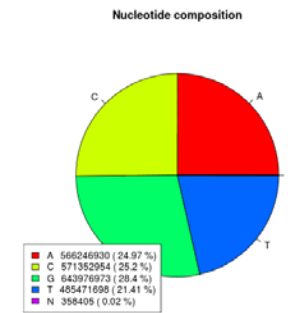


Run Stats

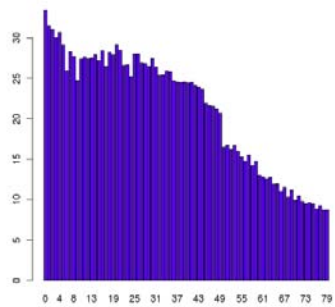
Normalised



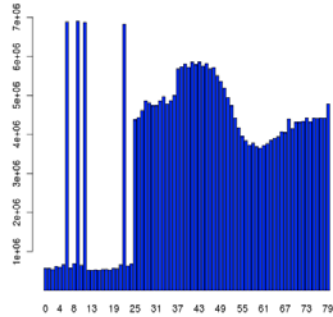
Unnormalised



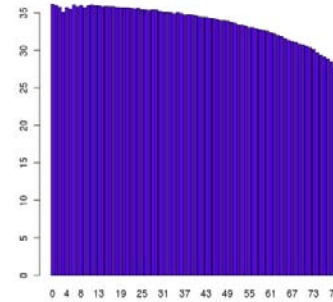
Qualities in position for nucleotide T



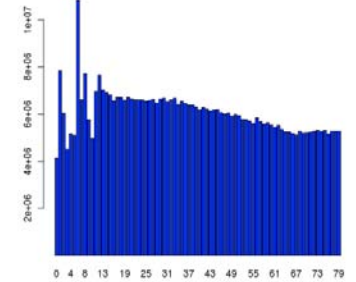
Count in position for nucleotide T



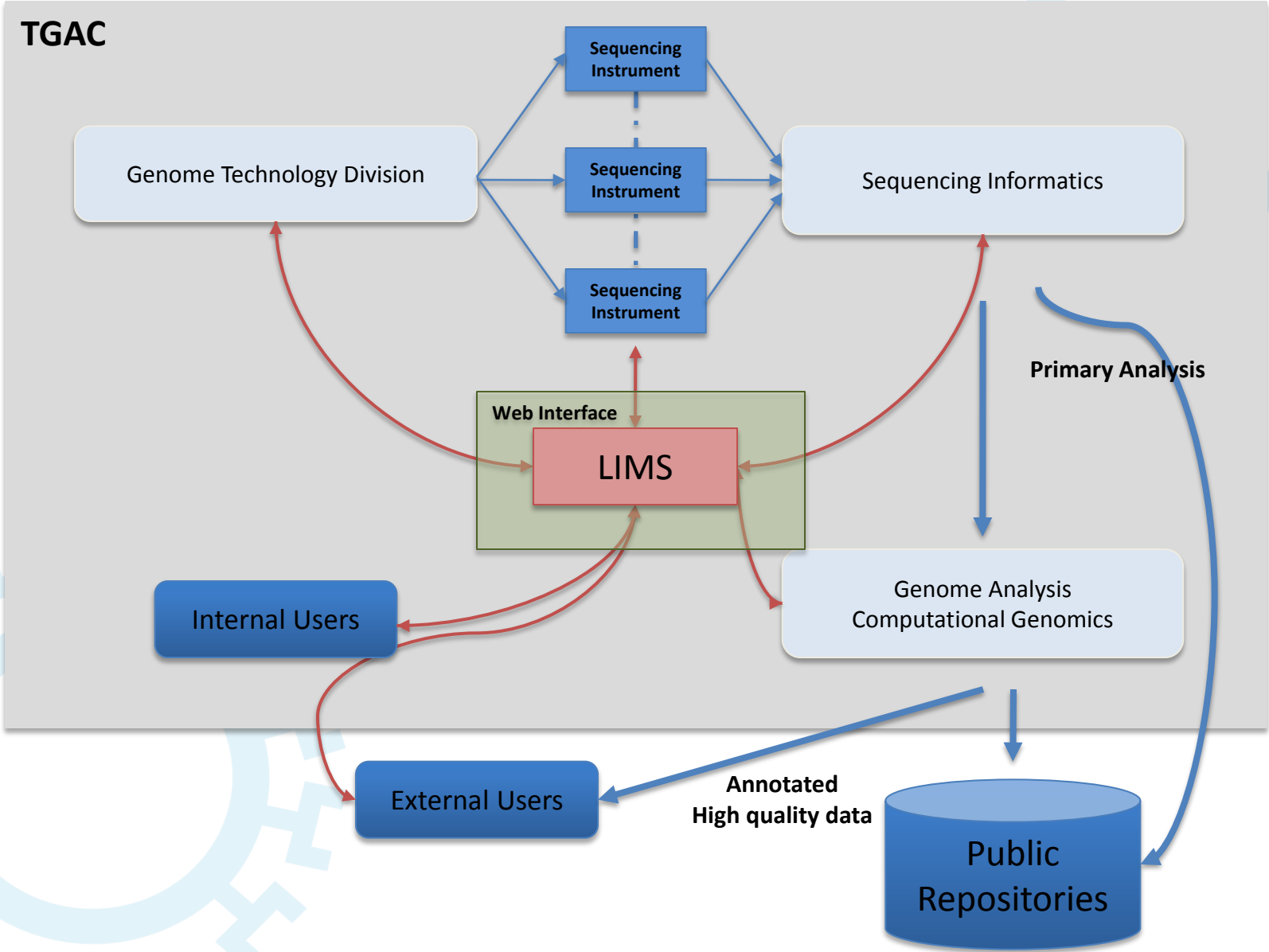
Qualities in position for nucleotide T

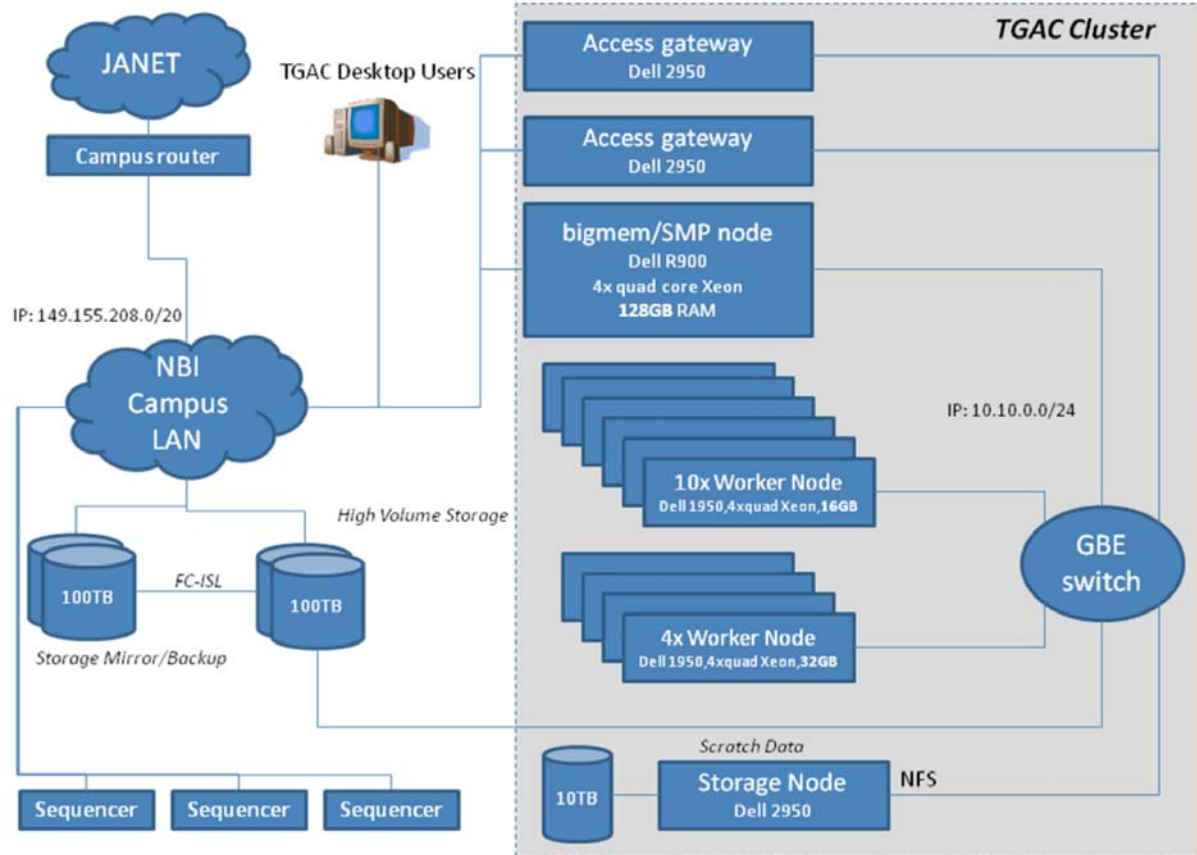


Count in position for nucleotide T

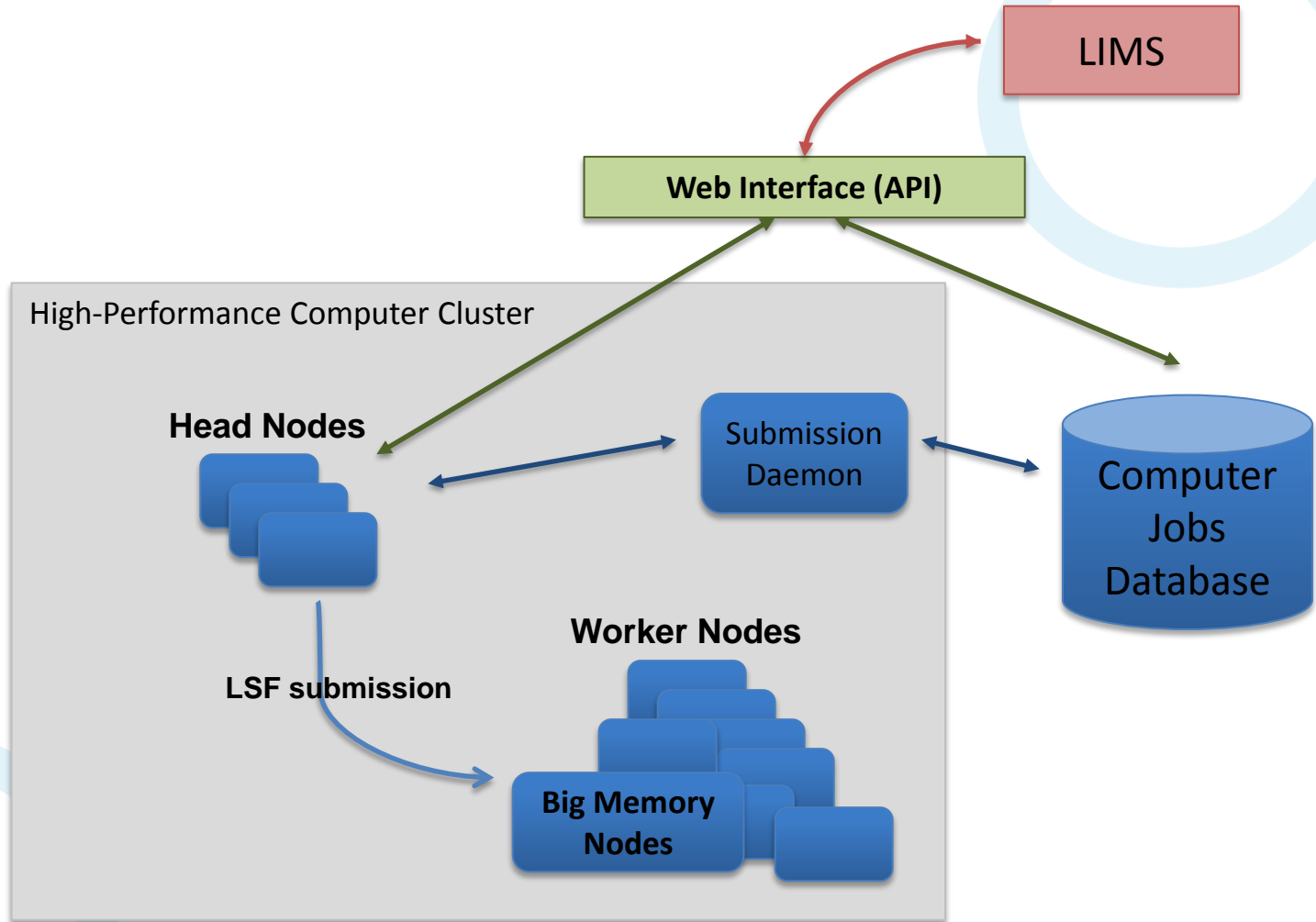


TGAC

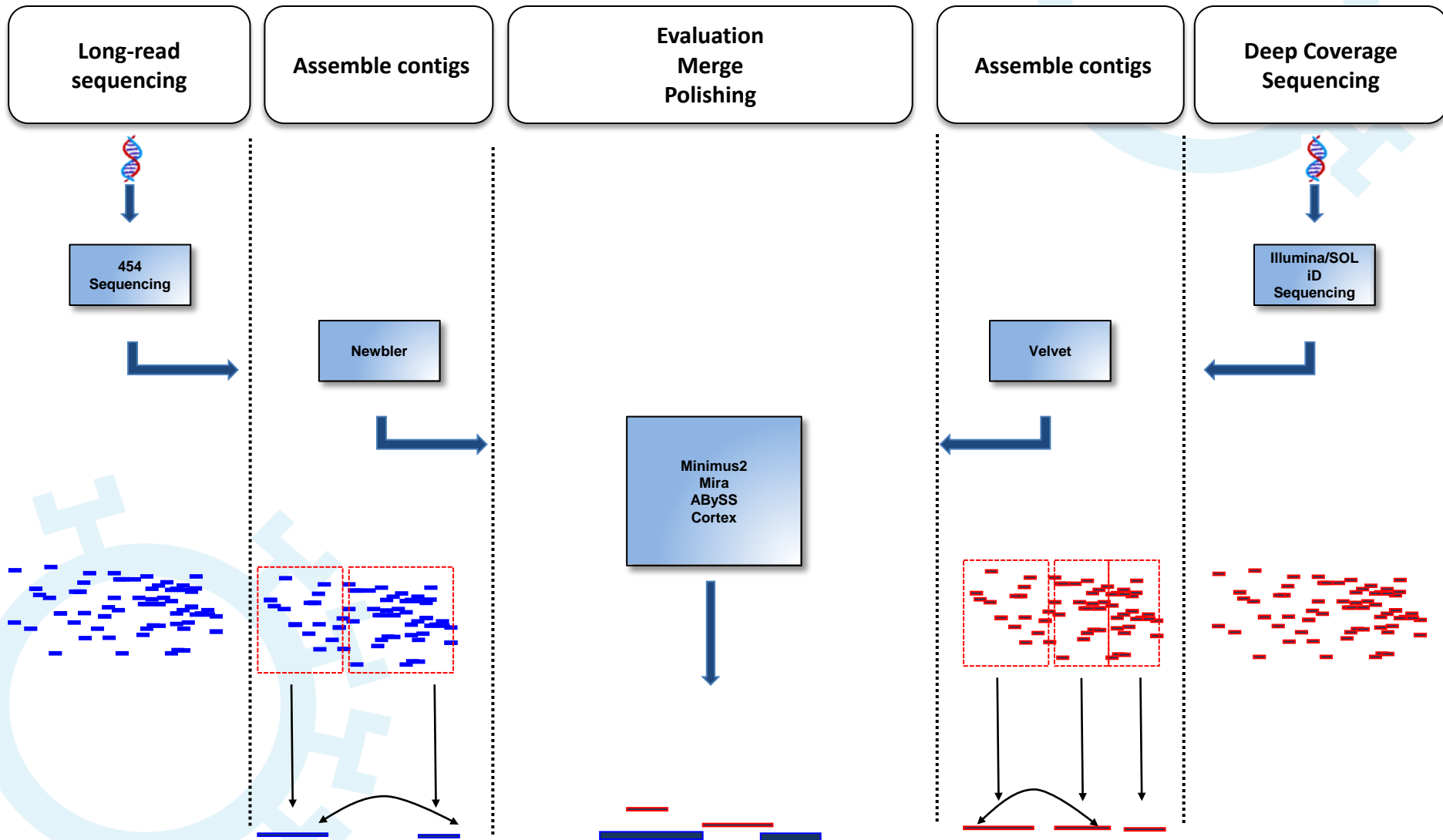




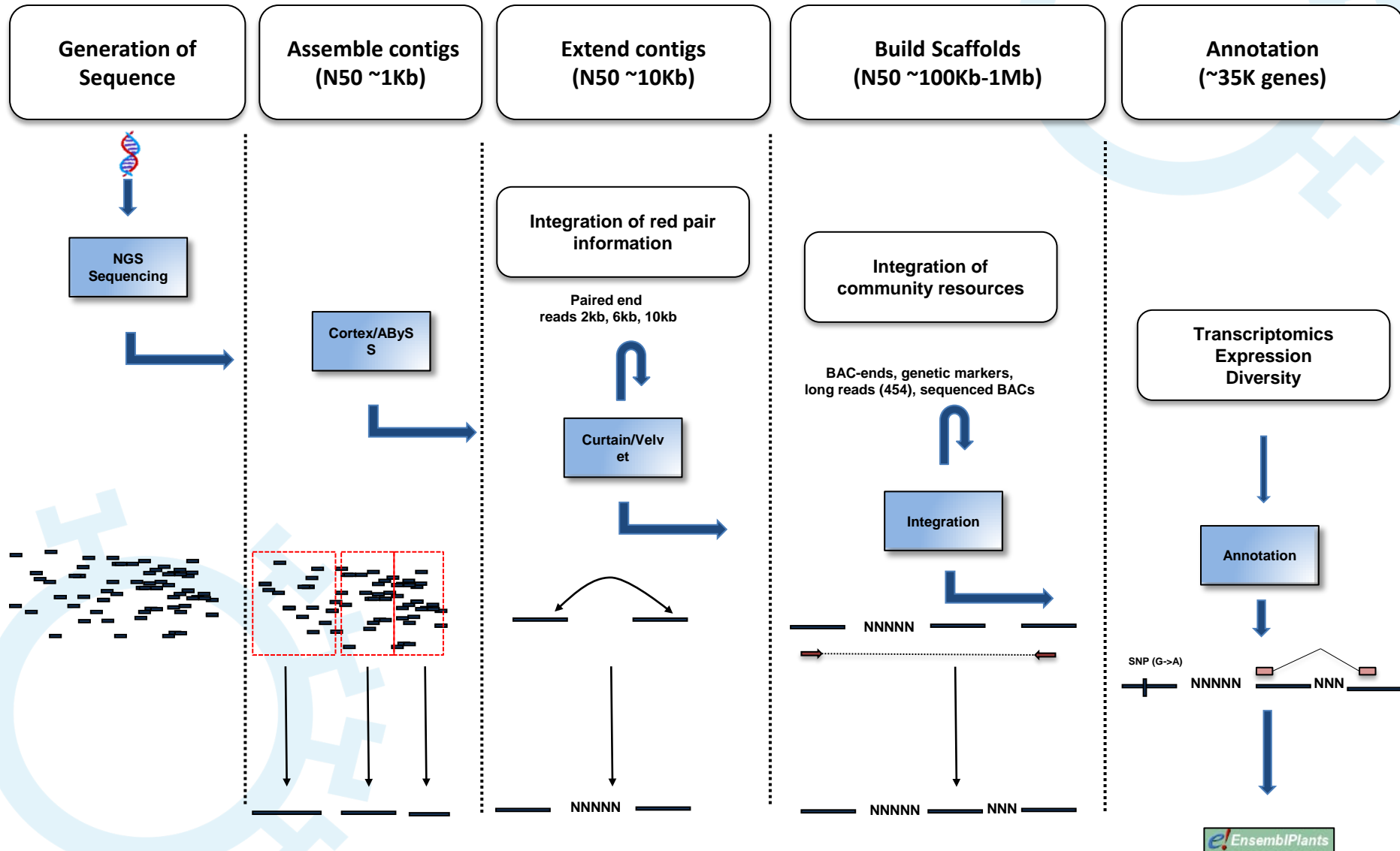
Computer Jobs: Scheduling and Management



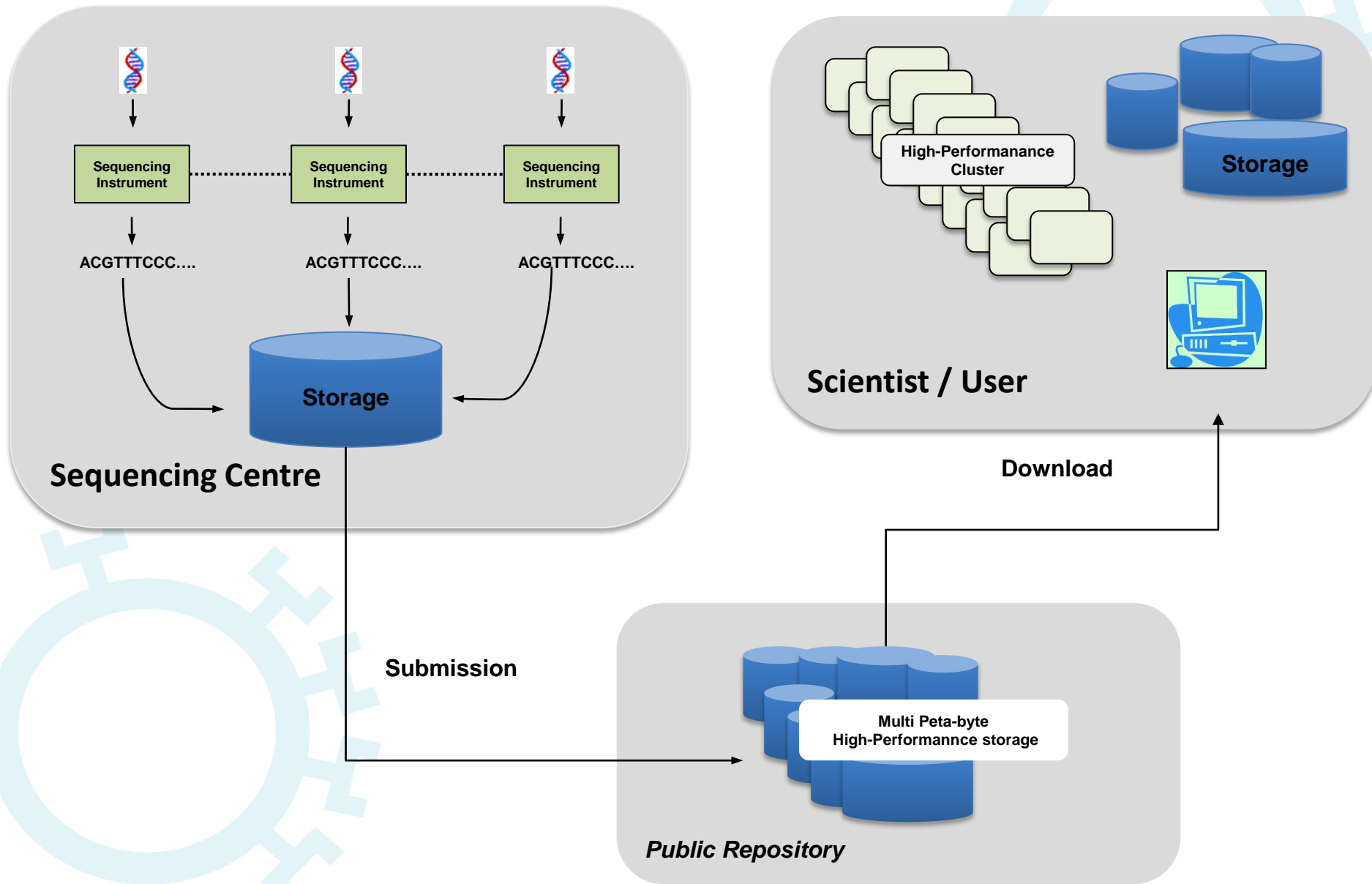
De Novo Pipeline – small genomes



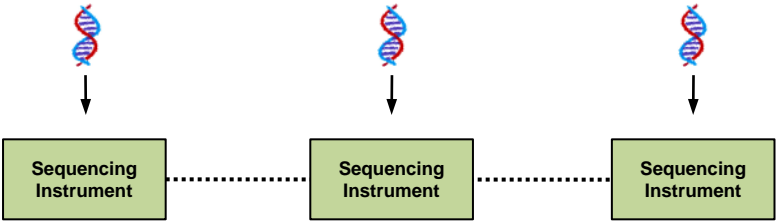
De Novo Pipelines – large genomes



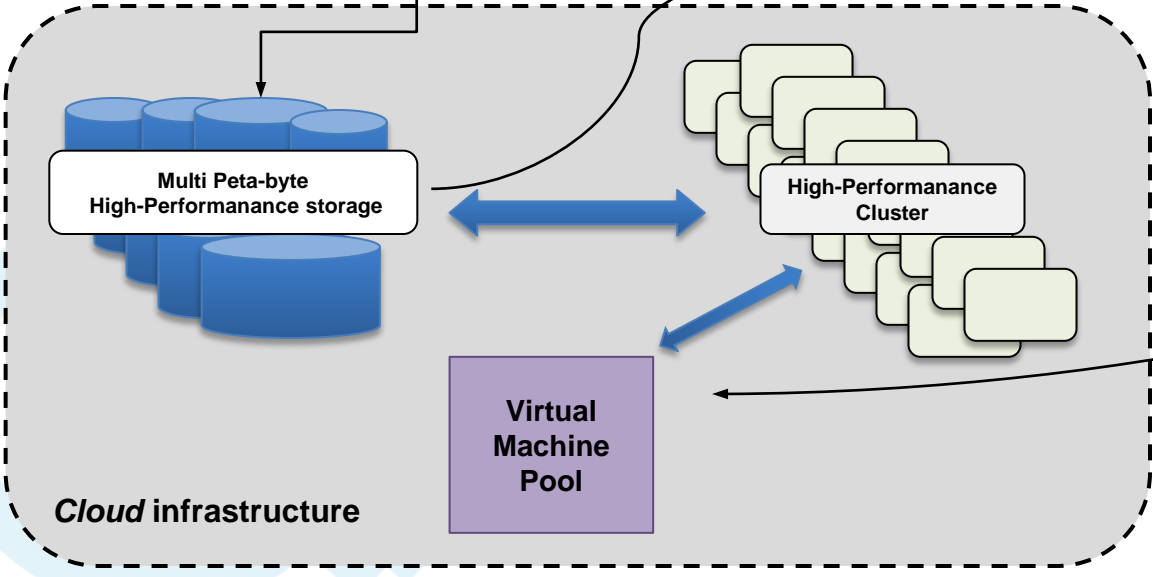
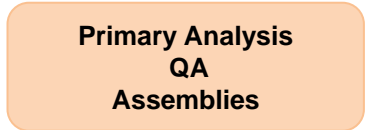
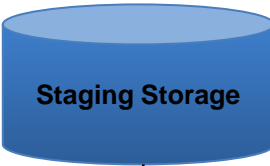
Dissemination of the Data



Sequencing Centre



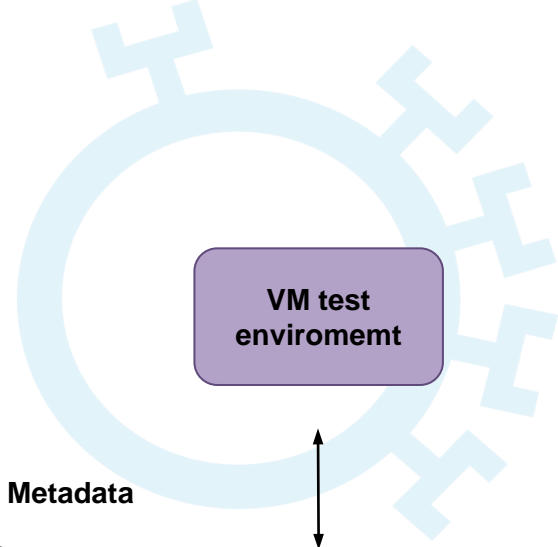
ACGTTTCCC.... ACGTTTCCC.... ACGTTTCCC....



Metadata

Analysis Output

Analysis Submission



Acknowledgements

- **TGAC:** *Rob Davey*, Xingdong Bian and Nizar Drou.
- **TOC:** Chris Bridson, Paul Fretter and Bob Findlay.
- **Eagle Genomics:** Richard Holland.



Thanks!

