Assembling a Cassava Transcriptome using Galaxy on a High Performance Computing Cluster

Aobakwe Matshidiso

Supervisor: Prof Chrissie Rey
Co-Supervisor: Prof Scott Hazelhurst

Next Generation Sequencing

The Good

RNA-Seq, CHiP-Seq, Bar-Seq, CHiP-chip, whole genotyping, miRNA profiling

SOLiD, Roche, Ion Torrent, Illumina, Pacific Biosciences, etc.

De novo sequencing, re-sequencing,

Holistic versus Reductionist

Minimal to no reliance on a priori knowledge

Next Generation Sequencing

~~The Bad~~ **The Challenges**

A plethora of data processing tools

Millions of short reads

Steep Learning Curves

Large memory requirements

Gigabytes of disk space

Aims

- Investigate the efficiency of transcriptome assembly and alignment tools
- Use the above-mentioned assembly and alignment tools to establish a cassava transcriptome

SOLiD Data

- RNA extracted from leaf tissue at three time points: 12, 32, 67 days post inoculation with South African Cassava Mosaic Virus (SACMV)

- From CMD-resistant TME3 and CMD-Susceptible T200 cassava cultivars
- Using SOLiD 4 System from Applied Biosciences

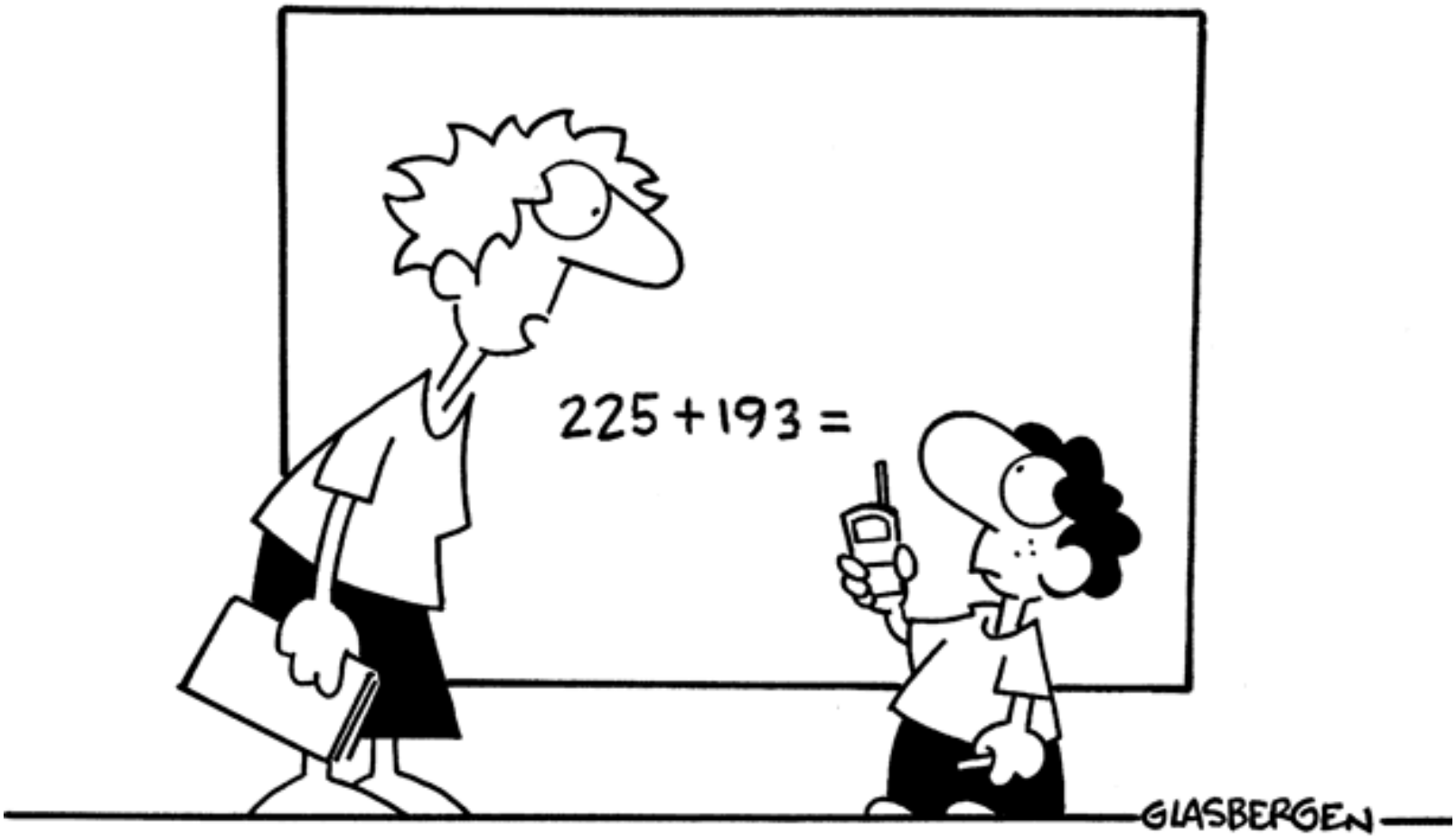
The High Performance Computing Platform

- Ubuntu “Lucid” Virtual Machine
- Eight 3 Ghz Processing Cores, 72 GB RAM
- accessible by SSH over high-speed internet
- On top of the ZA-Wits-Core Cluster:
  100 processor cores
- MPI Parallelization

Tools

- **DE NOVO ASSEMBLERS**
  De Bruijn Graphs: ABySS, Velvet
  Overlap Layout Consensus: Shore, Shorty
  Greedy: SSAKE, SHARCGS

- **REFERENCE-BASED ALIGNMENT**
  BWT: Bowtie, BWA
  Reads Hashing: SHRiMP, PerM, SOCS
  Genome Hashing: PASS, MOSAIK

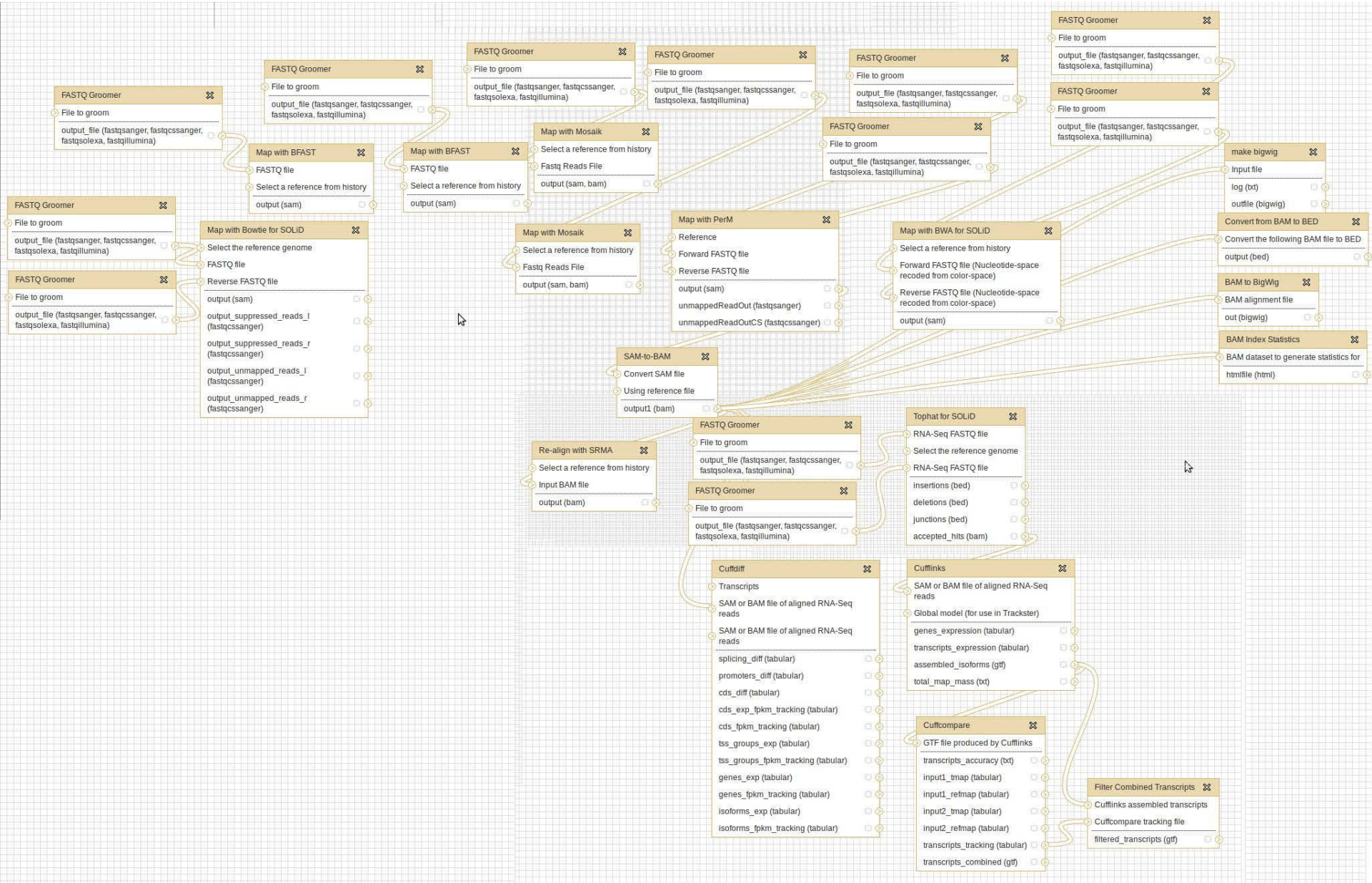


"You have to solve this problem by yourself. You can't call tech support."

Workflow Pipeline: Galaxy

- Workflow planning
- Centralize the Investigation
- Easy to install, Easy to configure [???]
- Large Datasets
- Track Workflow Histories

Workflow Pipeline: Galaxy



Quality Control

Dataset: EA – Resistant, 12 dpi, SACMV negative

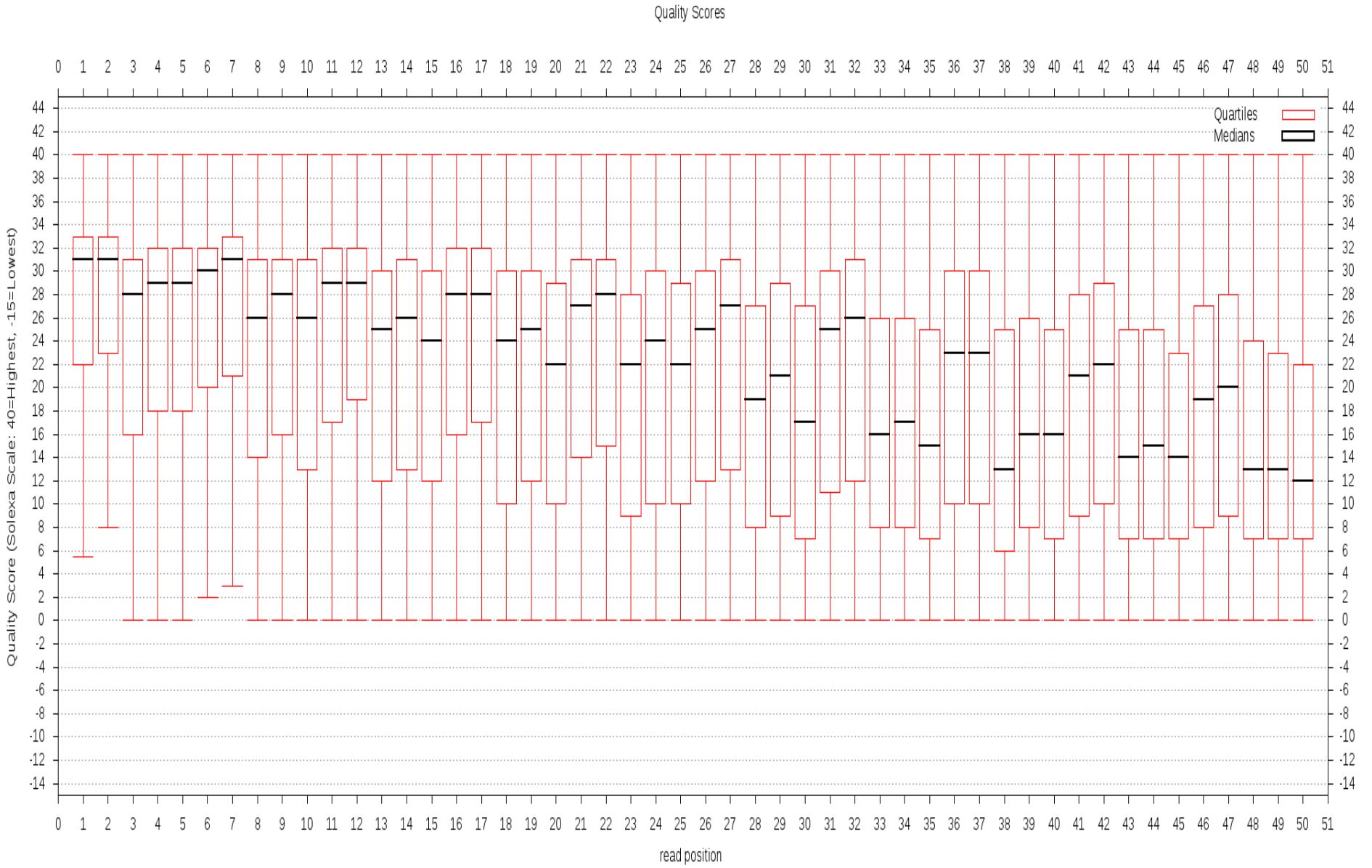
	Raw Forward	Raw Reverse
# of Reads	58, 133, 361	58, 133, 361
% GC	48%	50%
Sequence Length	50	35

Overrepresented Sequences

Sequence	Count	Percentage
CAAAACGACTCTCGGCAACGGATAT CTCGGCTCTCGCATCGATGAAGAA	884 830	2%
AACGACTCTCGGCAACGGATATCTC GGCTCTCGCATCGATGAAGAACGC	339 864	1%

Quality Control

EA RAW Forward Reads



Alignment Statistics

Bowtie Alignment: EA Raw with Default Settings

No. CPU Cores	1	2	4	6	8
User time	31, 261s	31, 179s	31, 869s	33, 322s	33, 530s
System Time	54.33 s	48s	52.48s	53.21s	62.66s
% of CPU	99	199%	399%	598%	791%
Wall Clock Time	8:43:58	4:21:20	2:13:16	1:32:56	1:10:41
Reads Processed	58, 133, 361	58, 133, 361	58, 133, 361	58, 133, 361	58, 133, 361
Reads with Alignment	193	193	193	193	193
Reads with Failed Alignments	58, 133, 168	58, 133, 168	58, 133, 168	58, 133, 168	58, 133, 168

The Comparative Study:

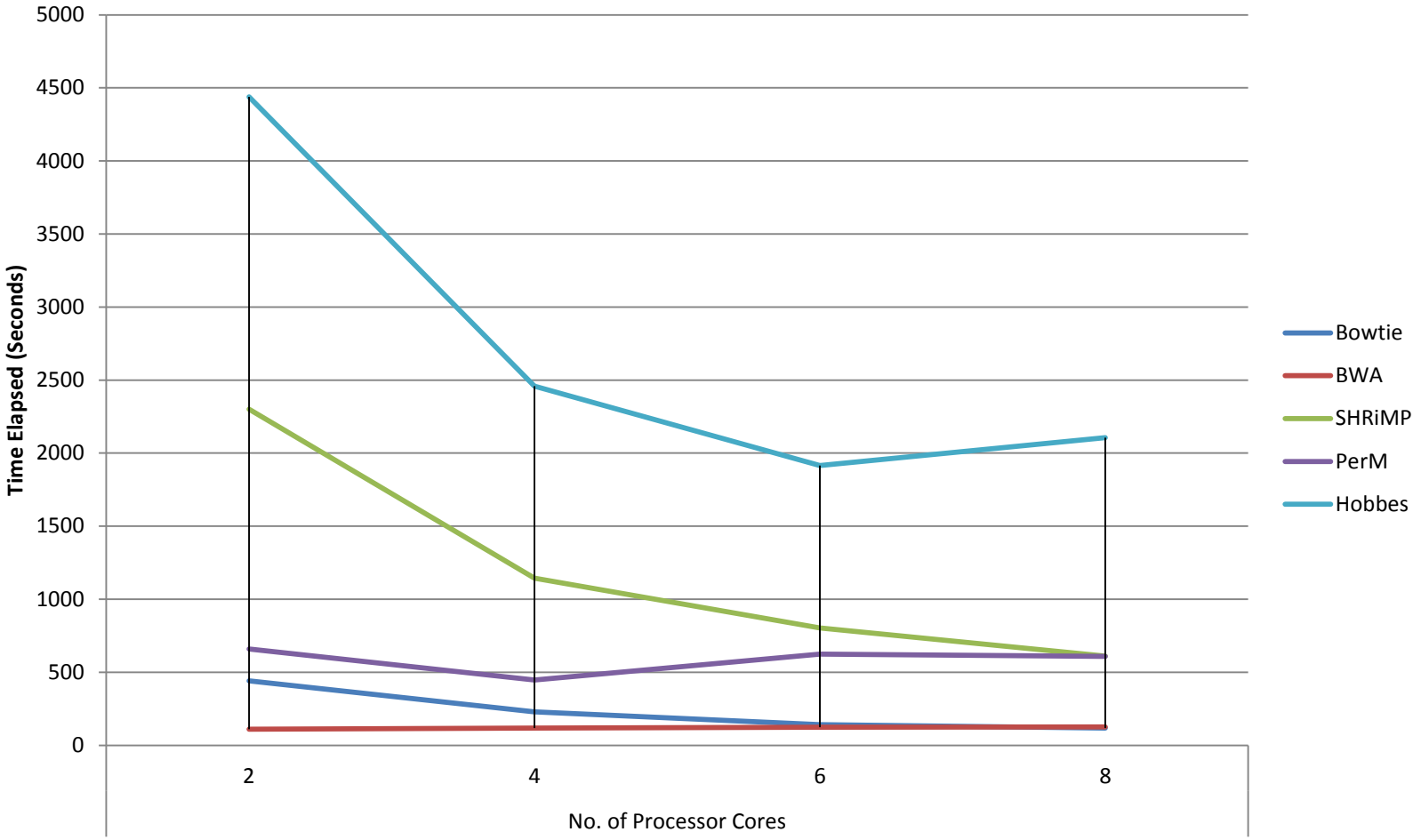
DBG Assemblers

	N	Min	Med	Mean	Max	N50	N50 length
Velvet19	743, 122	37	41	44	256	310, 736	42
Velvet25	41, 927	49	45	69	436	14, 497	74
ABySS 19	19, 763, 844	19	19	20	382	8, 928, 450	19
ABySS 25	6, 333, 943	25	26	28	378	2, 791, 286	26

The Comparative Study:

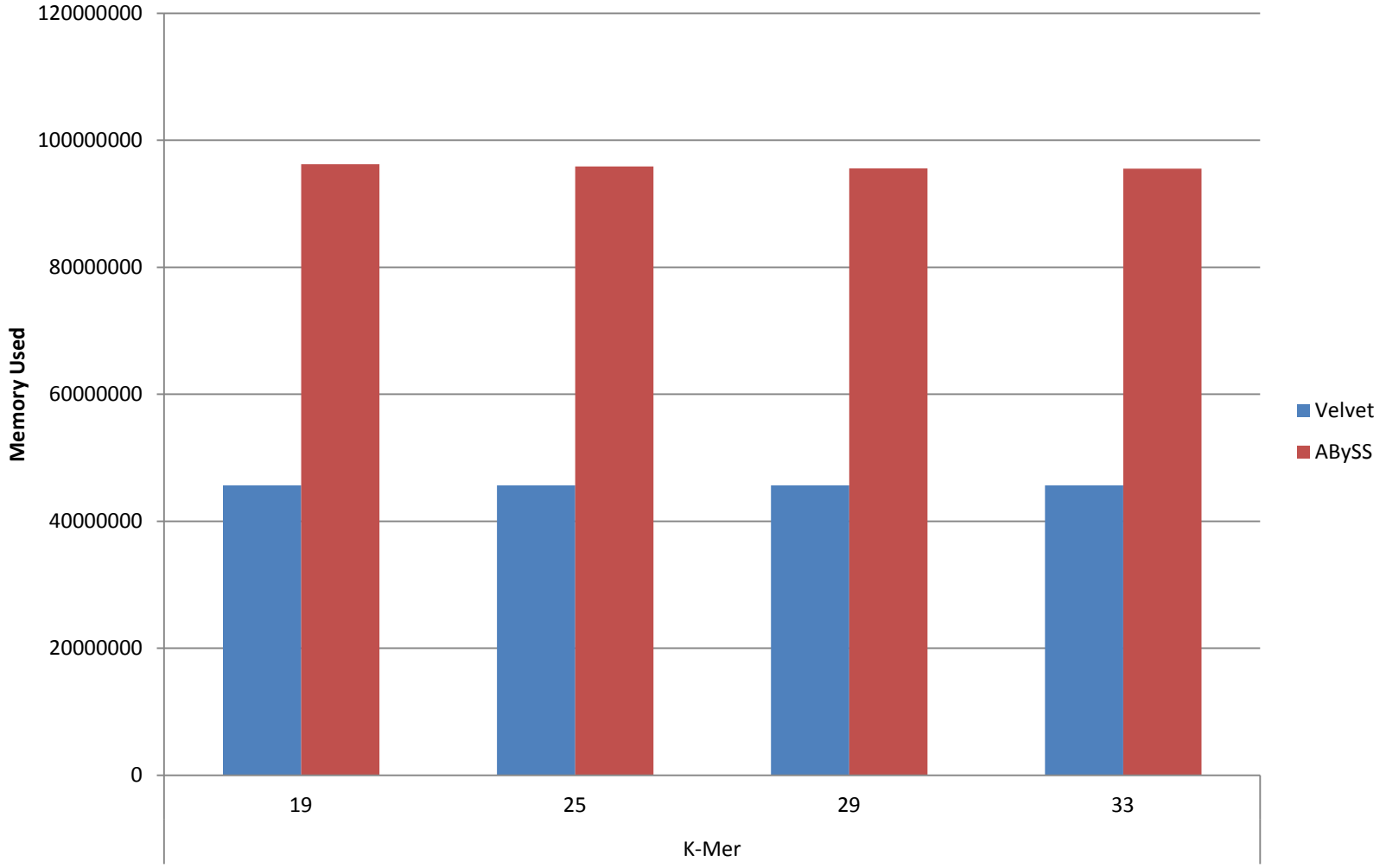
Alignment Tools

Time Lapsed per processor cores



The Comparative Study:

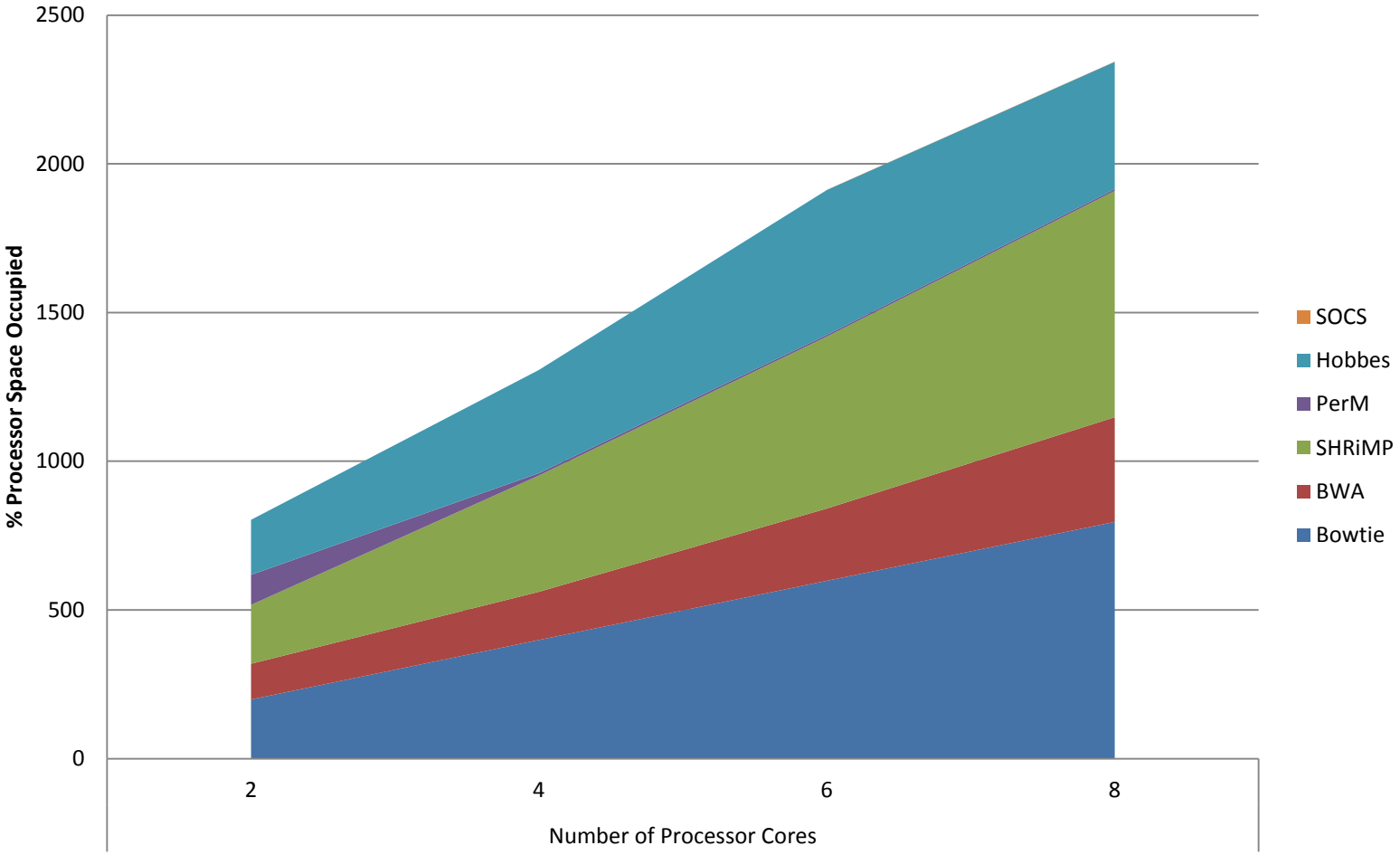
De Bruijn Assemblers



The Comparative Study:

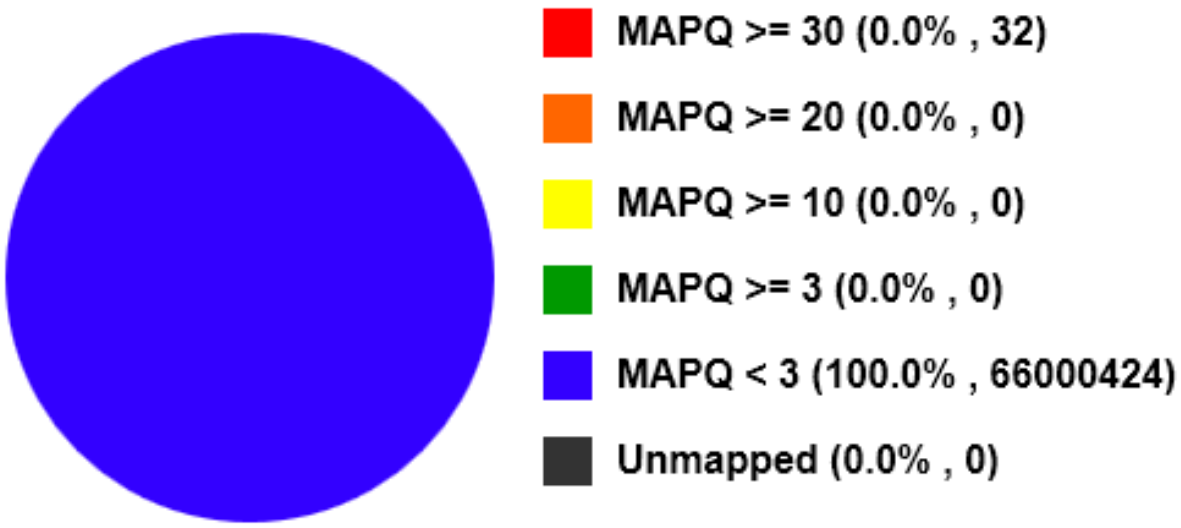
Aligners

% of CPU Used per Processor Core



Alignment

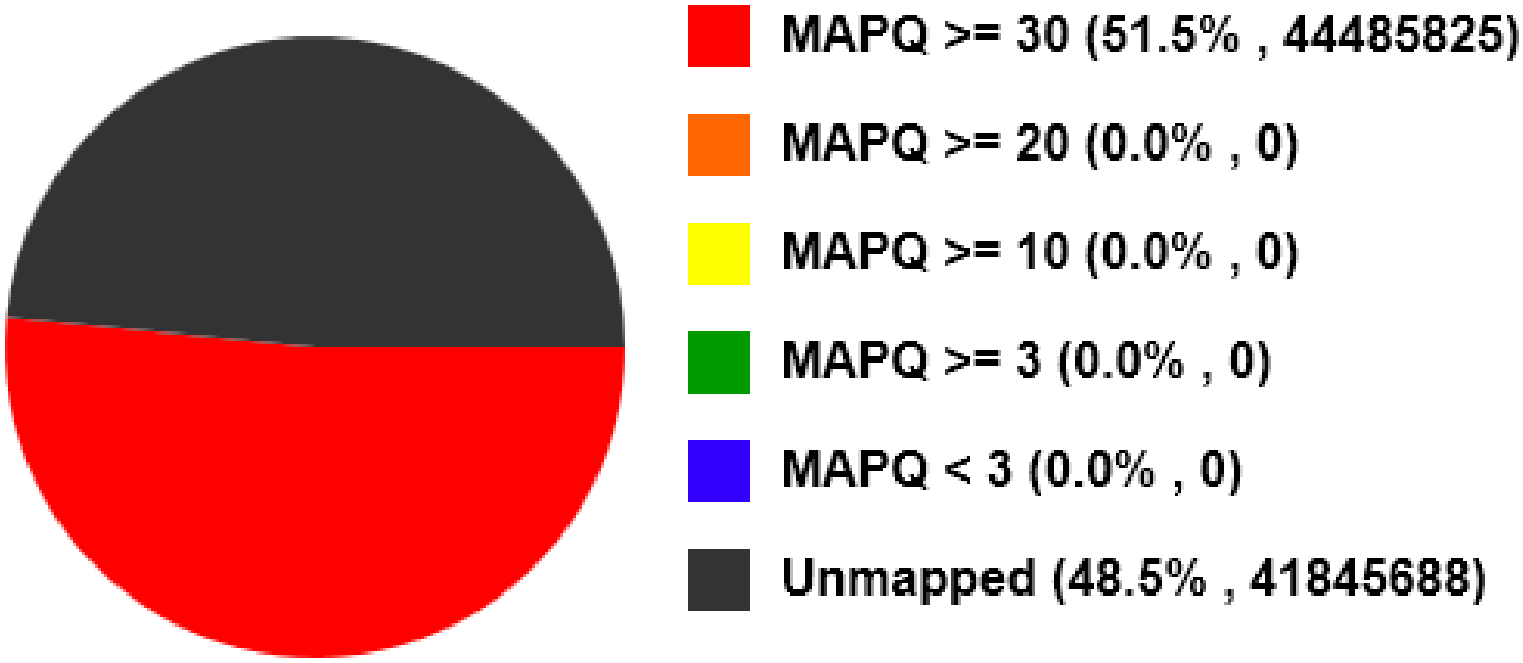
Mapping Quality – Bowtie, EA Enhanced, Default Settings



Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences. The percentage and number of alignments in each category is given in brackets.

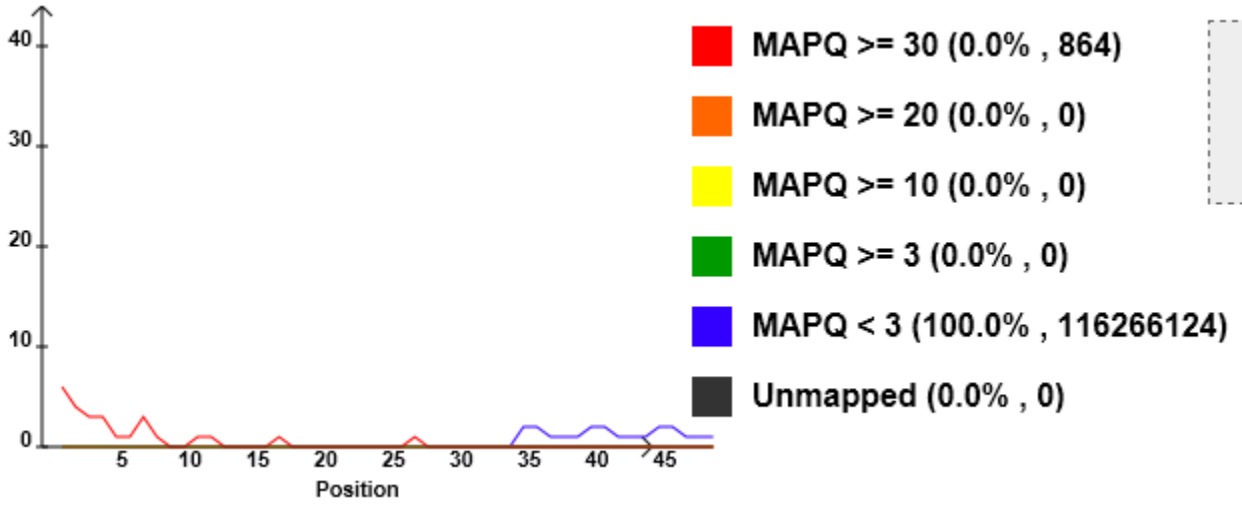
Alignment

Mapping Quality – Bowtie EA Enhanced, Strict/ Enhanced Settings



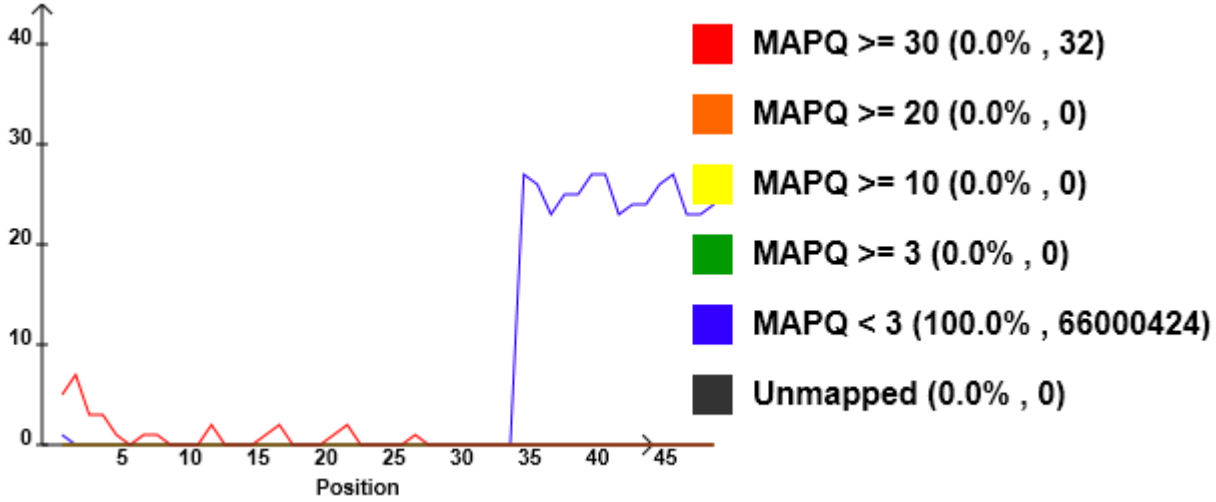
Alignment

Mapping Quality – Bowtie Enhanced Data, Default Settings, Normal Reference



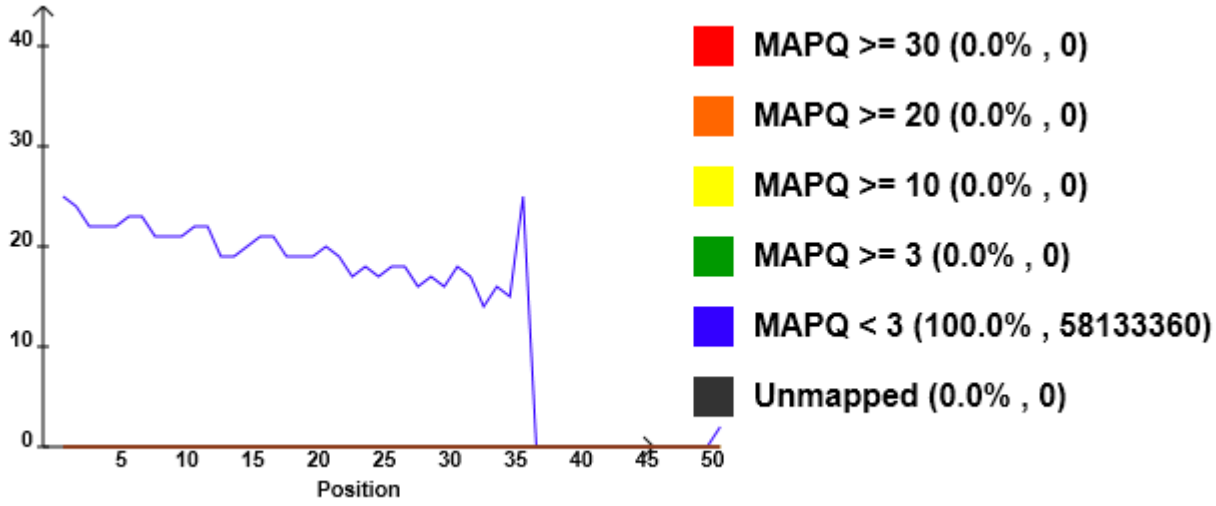
Mean base quality of reads with low and high mapping quality.

Bowtie Enhanced Data, Enhanced Settings, Normal Reference

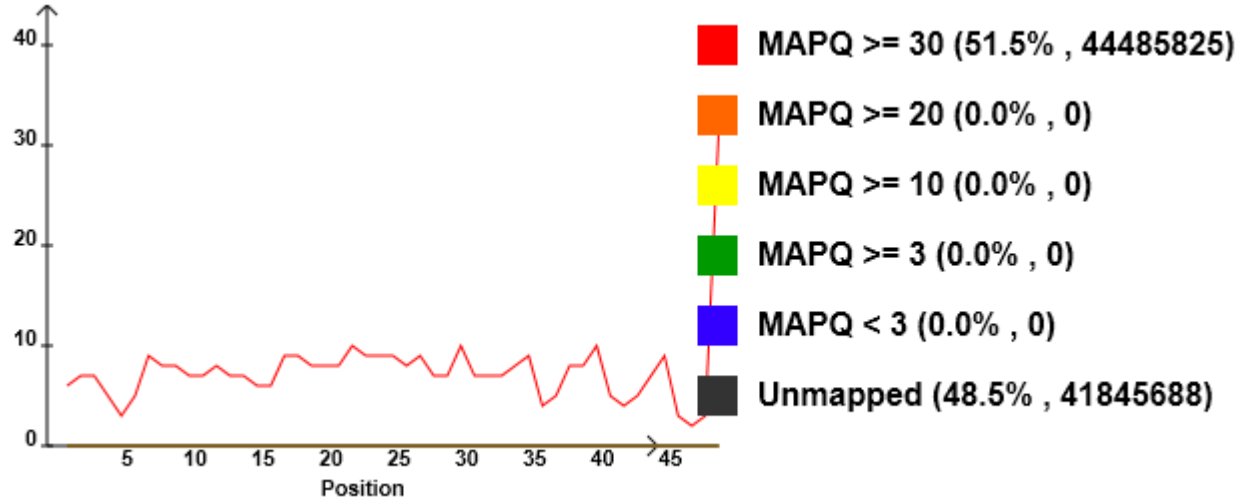


Alignment

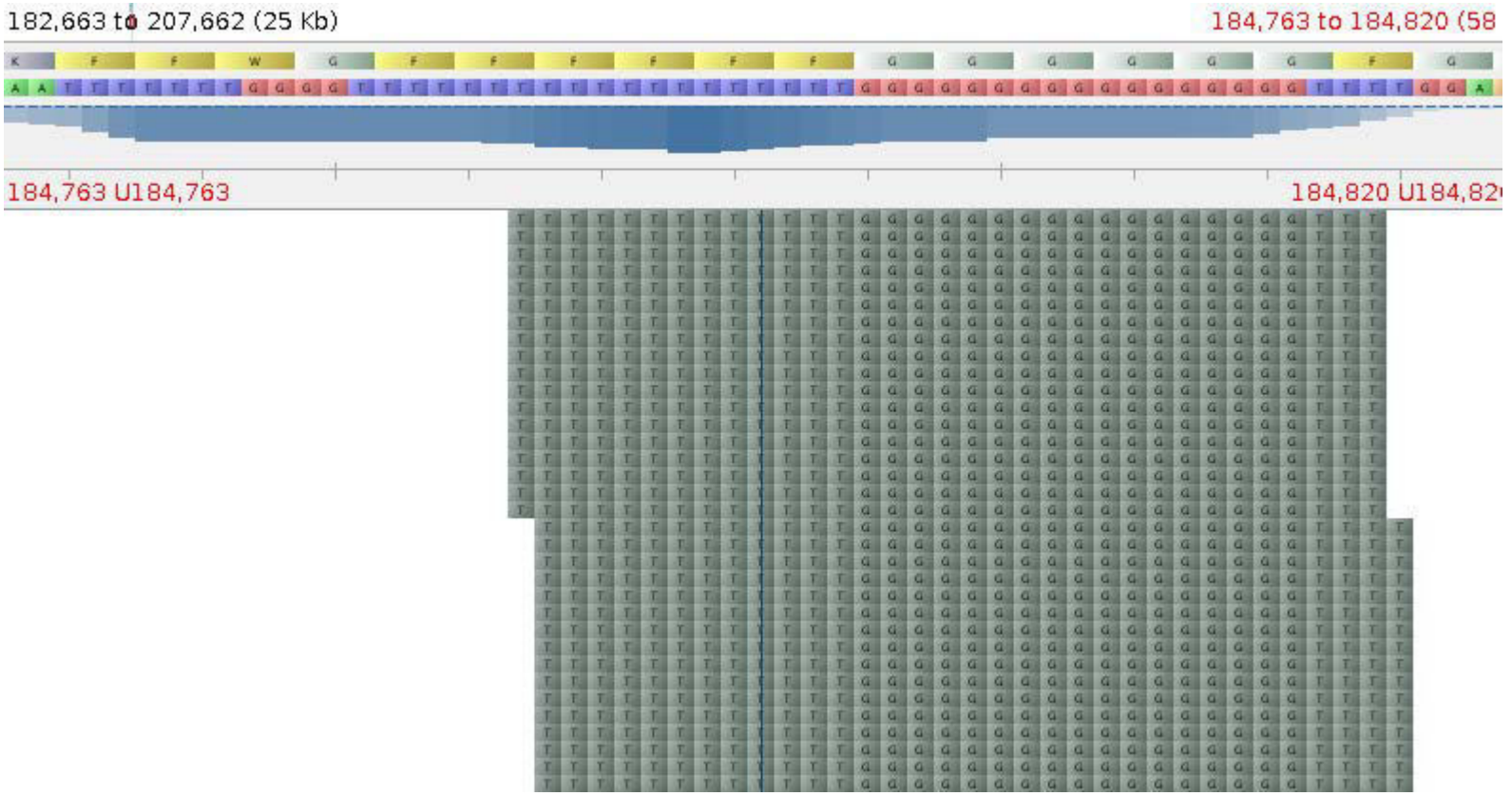
Mapping Quality – BWA Enhanced Data, Default Settings, Softmasked Reference



Bowtie Enhanced Data, Enhanced Settings, Normal Reference



The Transcriptome

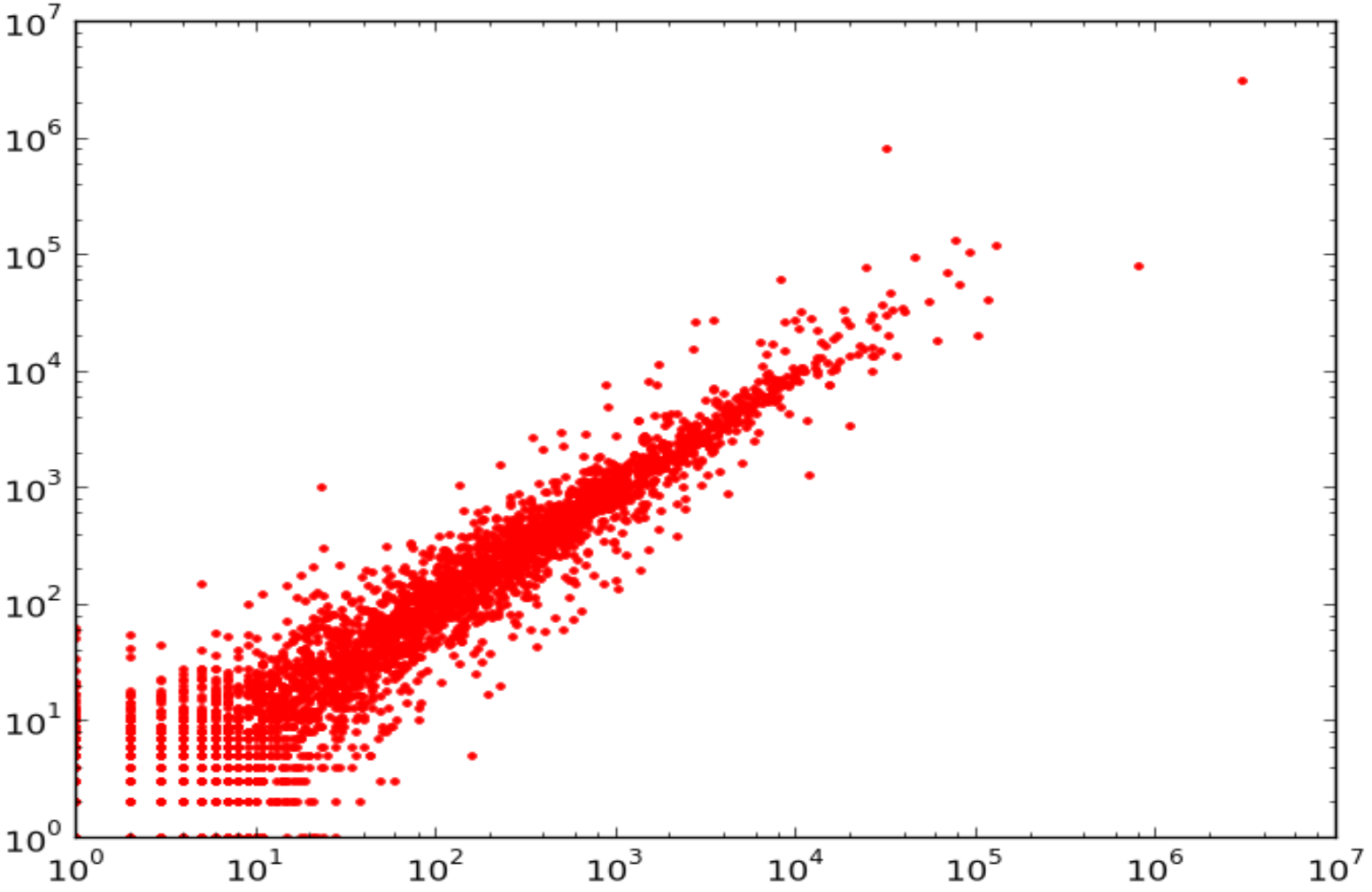


Visualization of reads mapped to a scaffold
in Tablet 2.6

Quantile Normalization

Gene Counts Across Datasets

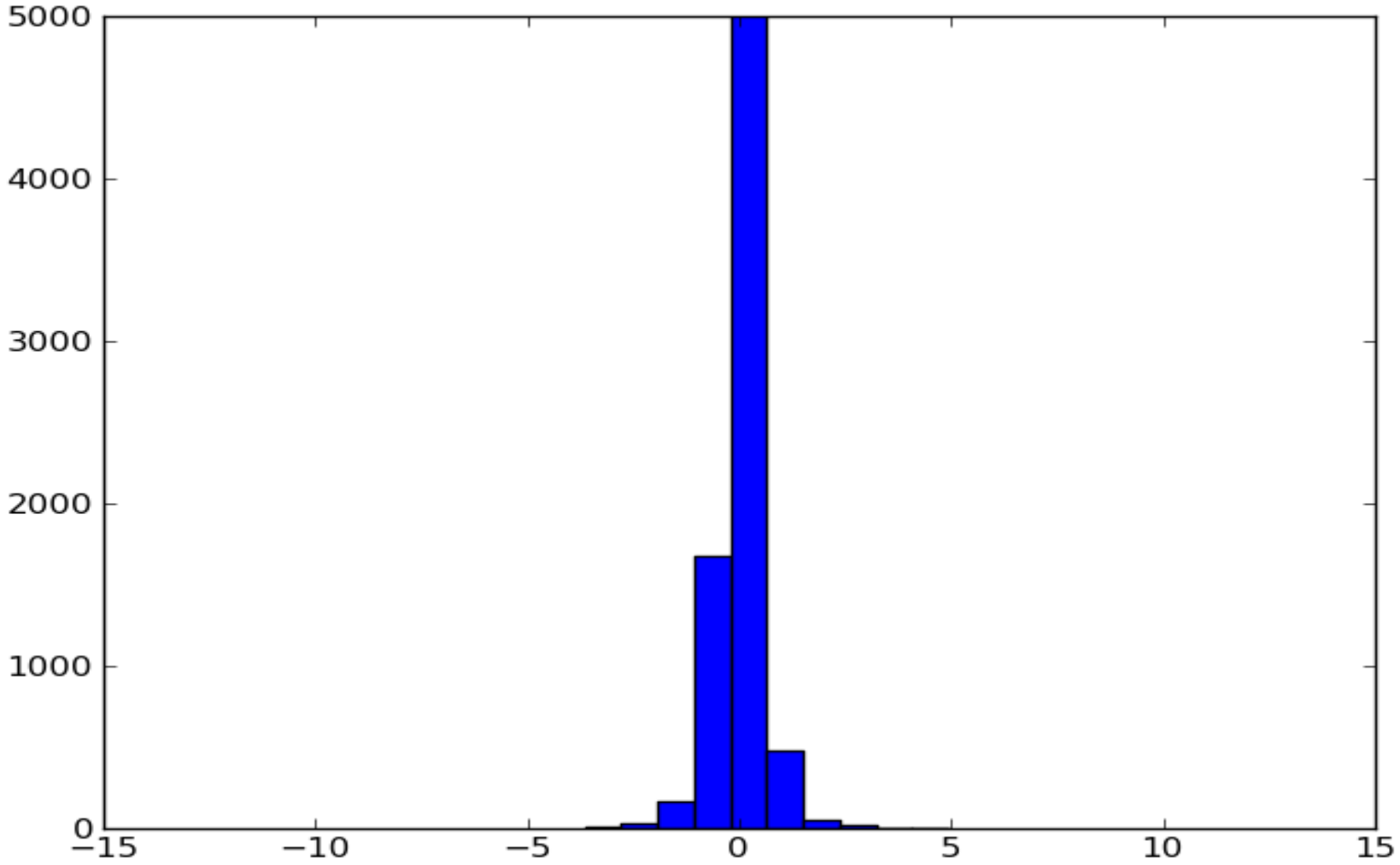
Normal State versus Disease Challenged (Resistant vs Susceptible, 67 dpi)



Naïve Expression Analysis

Distribution of Log2 Ratios of Gene Expression Counts

Normal State versus Disease Challenged (Resistant vs Susceptible, 67 dpi)



In Progress

- Gaining new knowledge relating to host-pathogen interactions, in the cassava-SACMV relationship
- Proceeding to build a cassava disease-challenged transcriptome
- Annotation of the cassava genome in progress

Conclusion

- Quality Control is Important,
  Pre-processing is critical
- Galaxy can enhance and facilitate bioinformatics analyses
- It’s important to optimise your software to obtain quality results
- A High Performance system will greatly enhance your studies

Funding

GAUTENG PROVINCE
AGRICULTURE AND RURAL DEVELOPMENT
REPUBLIC OF SOUTH AFRICA