

SCALABLE GENOMIC DATA MANAGEMENT SYSTEM

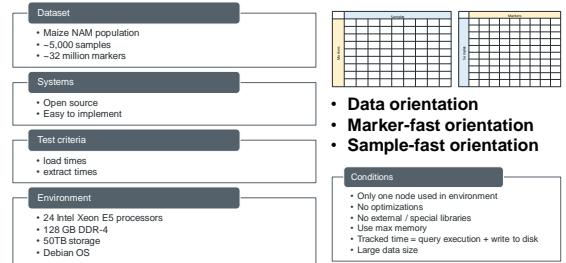
GOBI

January 14, 2019

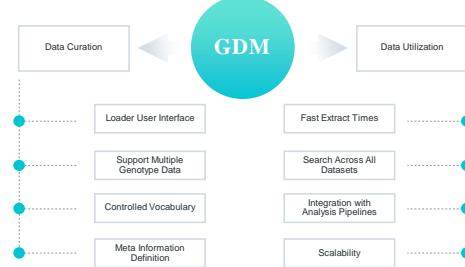
CREATED BY Yaw Nt-Addae

http://gobi-project.org

Benchmarking

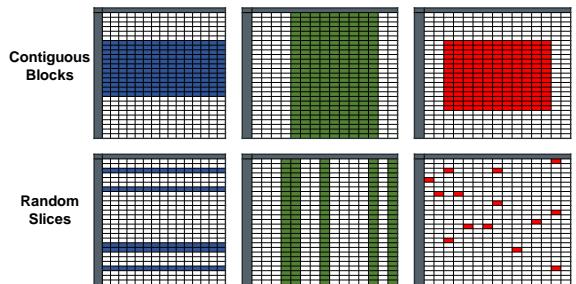


Genomic Data Manager



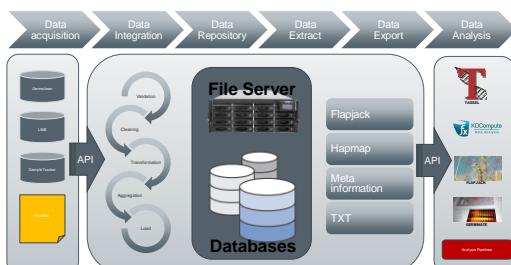
4

Benchmarking – Extracting data



8

GDM Workflow



5

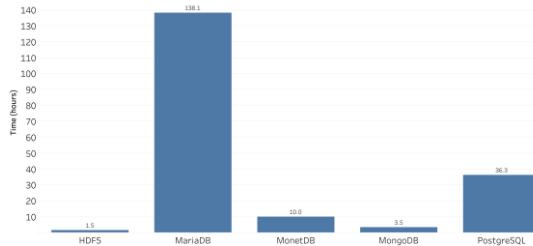
Benchmarking – Database Systems



9

Benchmarking Results – Loading Times

10



Conclusion

13



- ▶ Loading times for MariaDB and PostgreSQL may not be suitable for large datasets
- ▶ MonetDB has a number of column limitation and so was not able to create Sample x Marker orientation
- ▶ Significant effect of data storage orientation on extract times
- ▶ Significant effect of contiguous vs random data extract times
- ▶ Spark does better on random (non-contiguous) extracts
- ▶ Overall, HDF5 showed the most consistency in extract times

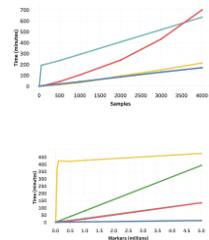
Benchmarking Results – Extract Times

11

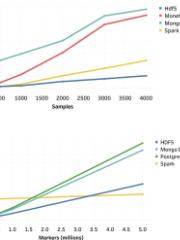
Acknowledgements

14

Contiguous Blocks



Random Slices



Cornell University

- Dave Matthews
- Elizabeth Jones
- Kelly Robbins
- Yaw Nti-Addae

University of Montpellier

- Pierre Larmande

CIRAD

- Adrien Pétel

CIRAD, UMR Intertryp

- Guilhem Sempére

Biodiversity International

- Valentin Guignon

CIMMYT - Mexico

- Victor Ulat

University of Minnesota

- Jon Renner

Benchmarking Results

12

