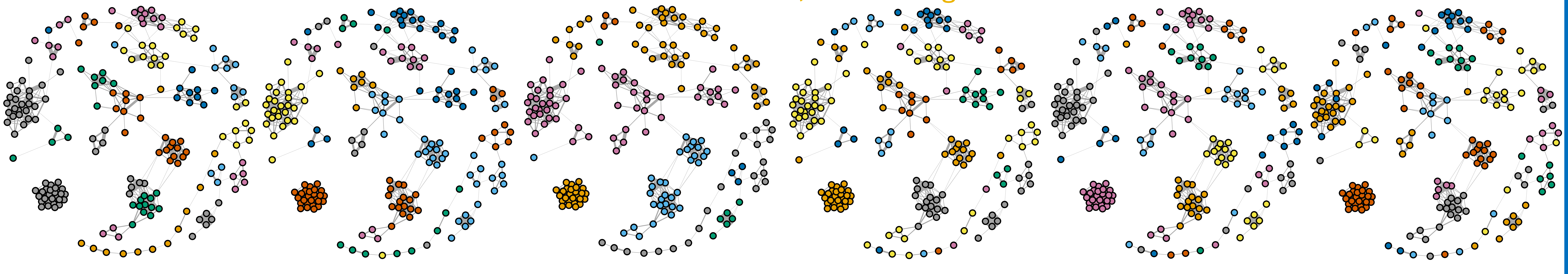
Pitt DBMI

Community Detection for Big Biological Networks with ExoLabel

Aidan Lakshman, Erik S. Wright

Research reported in this work was supported by the National Library of Medicine of the National Institutes of Health under award number **5T15LM007059-35**.

National Library of Medicine



Label Propagation | Fast Label Propagation | Louvain | **ExoLabel** | MCL (I=1.4) | MCL (I=2.0)

Problem

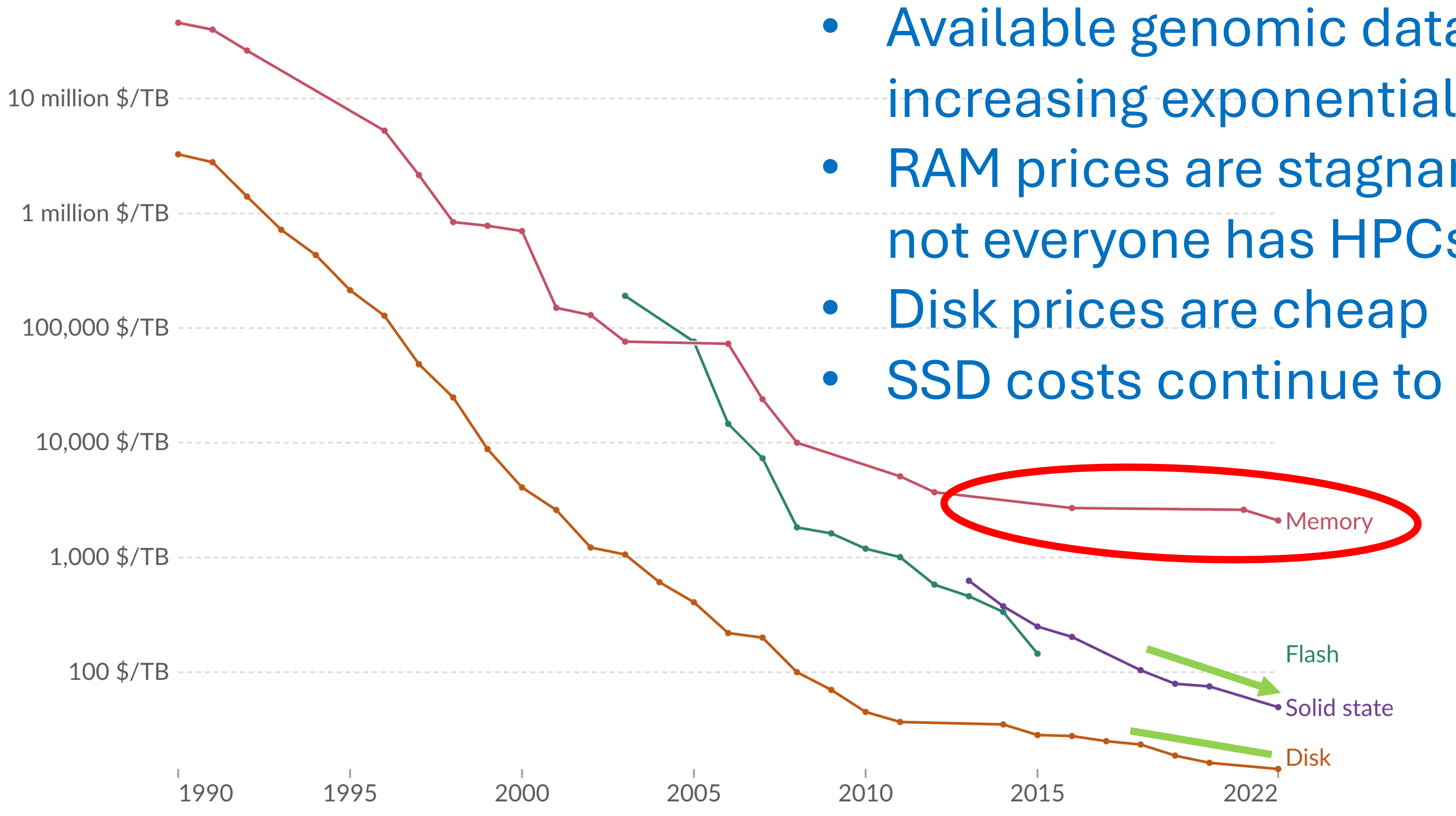
- Comparative genomics depends on **Orthology Detection**
- **Community Detection algorithms** can infer orthology groups from sequence similarity networks
- Current approaches **fail to scale** to modern genomic data

Solution

- We introduce a new algorithm, **ExoLabel**
- **ExoLabel** performs Label Propagation using **disk space to minimize memory consumption**
- **ExoLabel** can identify communities in a network with billions of nodes using **only 100MB of RAM**
- **Exolabel** matches state-of-the-art community detection methods in accuracy and outperforms in runtime & memory

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.



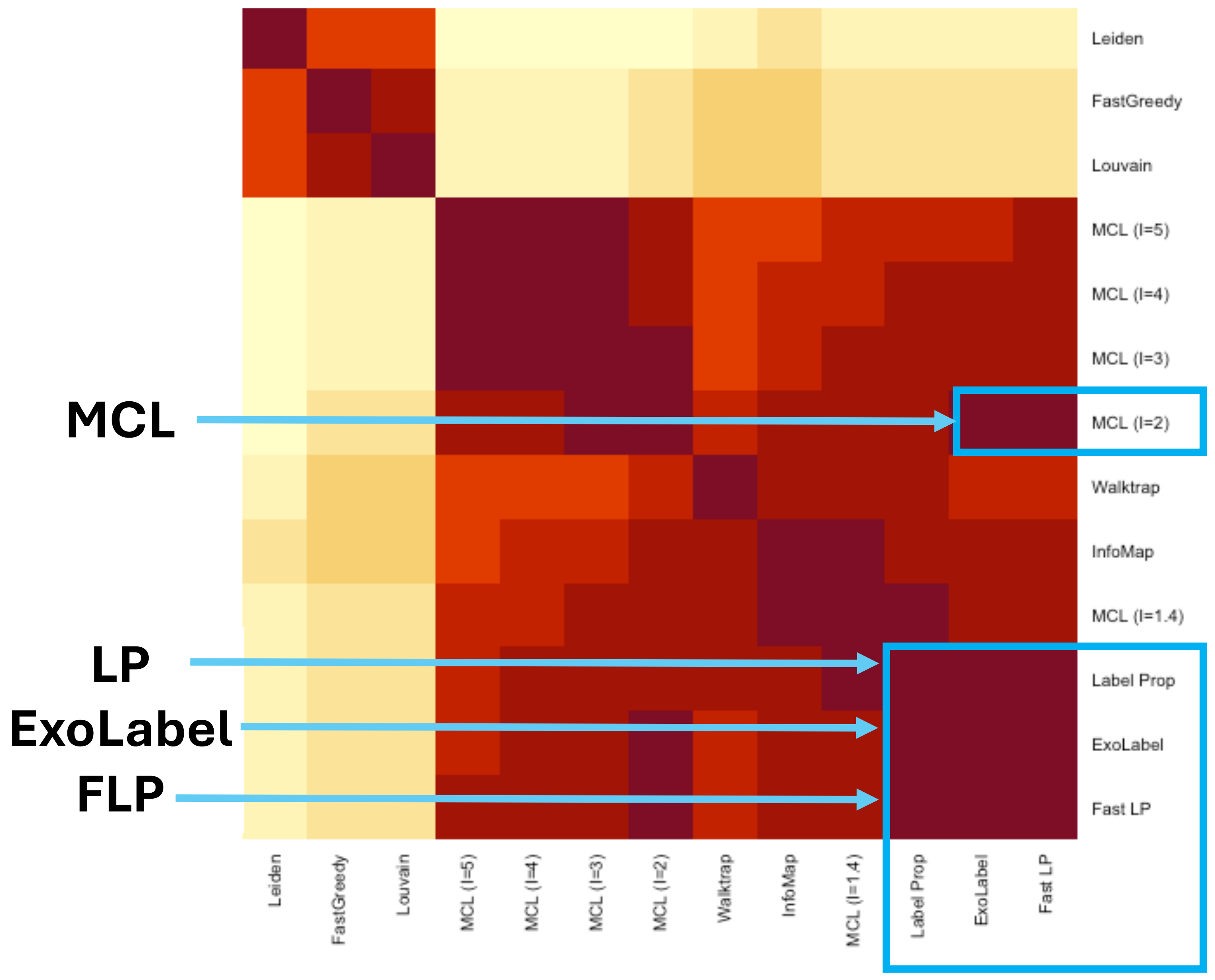
Data source: John C. McCallum (2022) OurWorldInData.org/technological-change | CC BY
Note: For each year, the time series shows the cheapest historical price recorded until that year.

- Available genomic data is increasing exponentially
- RAM prices are stagnant, and not everyone has HPCs
- Disk prices are cheap
- SSD costs continue to plummet

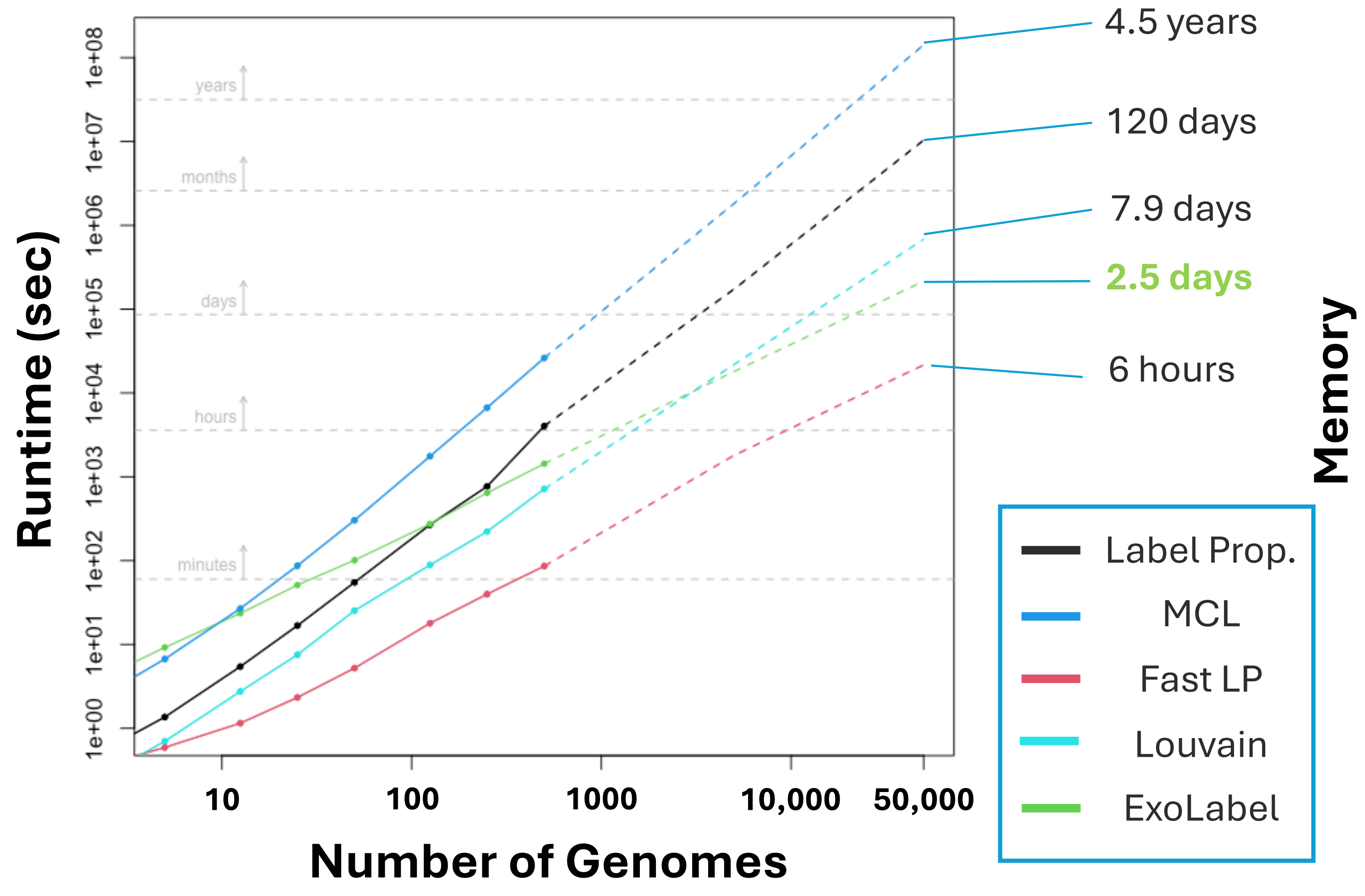
Details

- ExoLabel has log-linear runtime scaling in terms of number of nodes and edges
- Memory scaling is constant
- Disk consumption is linear with respect to number of nodes and edges
- Performance is comparable to low-inflation MCL or Label Propagation

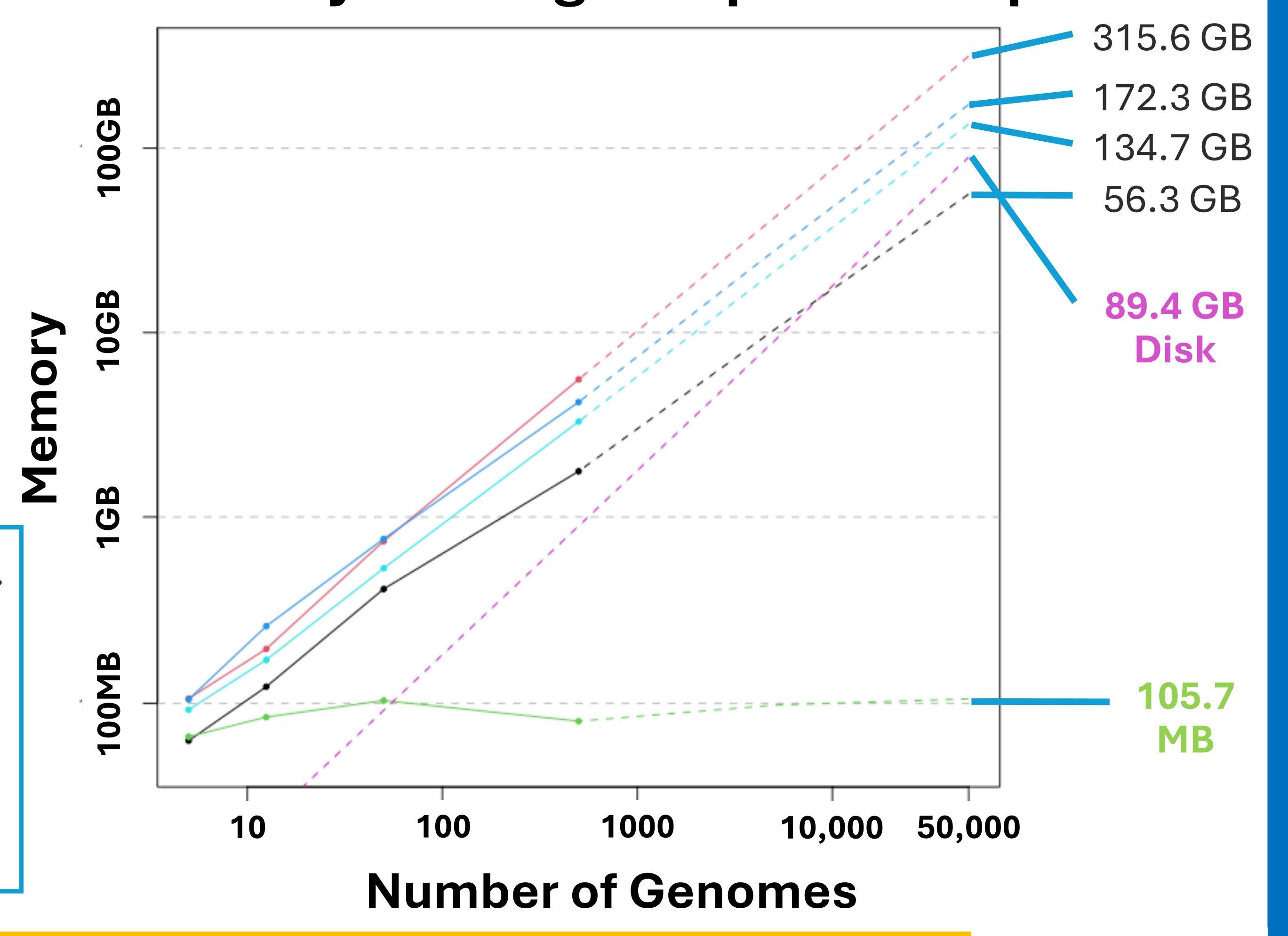
Community Similarity on 50 Prokaryotic Genomes



Runtime Scaling on Sparse Graphs



Memory Scaling on Sparse Graphs



WRIGHT LABORATORY

Our Lab:
www.WrightLabScience.com
www2.DECIPHER.codes

DECIPHER SynExtend

ExoLabel is available in the SynExtend package for R

Contact me!
AHL27@pitt.edu
www.AHL27.com