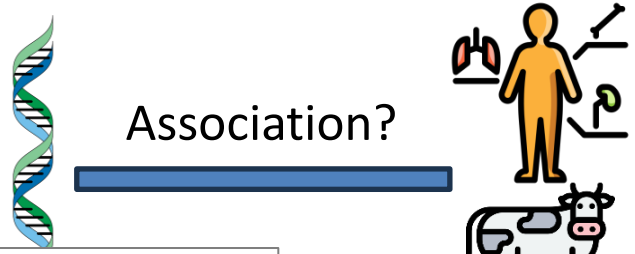
Resource-Efficient Approach for Large-Scale Genome-Wide Association Studies

Mulya Agung

Institut Teknologi Sains Bandung

Resource-Efficient Approach for Large-Scale Genome-Wide Association Studies

- The discovery of associations between genetic variations and traits
- More than 90,000 GWAS have been performed as of 2024 (Harris et al., 2025)



Association?

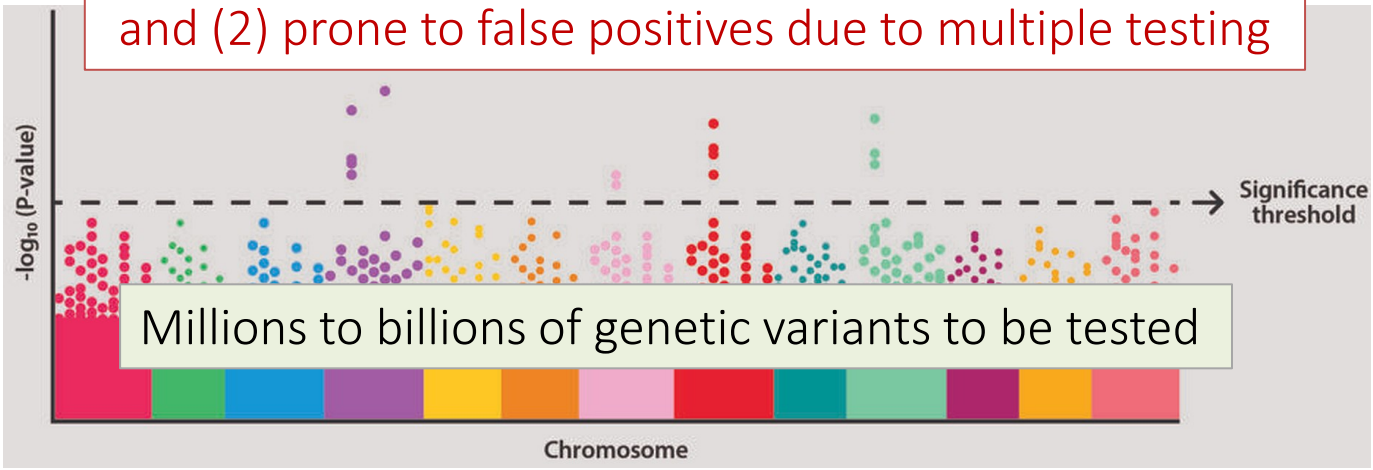
Genotype

Variation in the genome between individuals

Phenotype

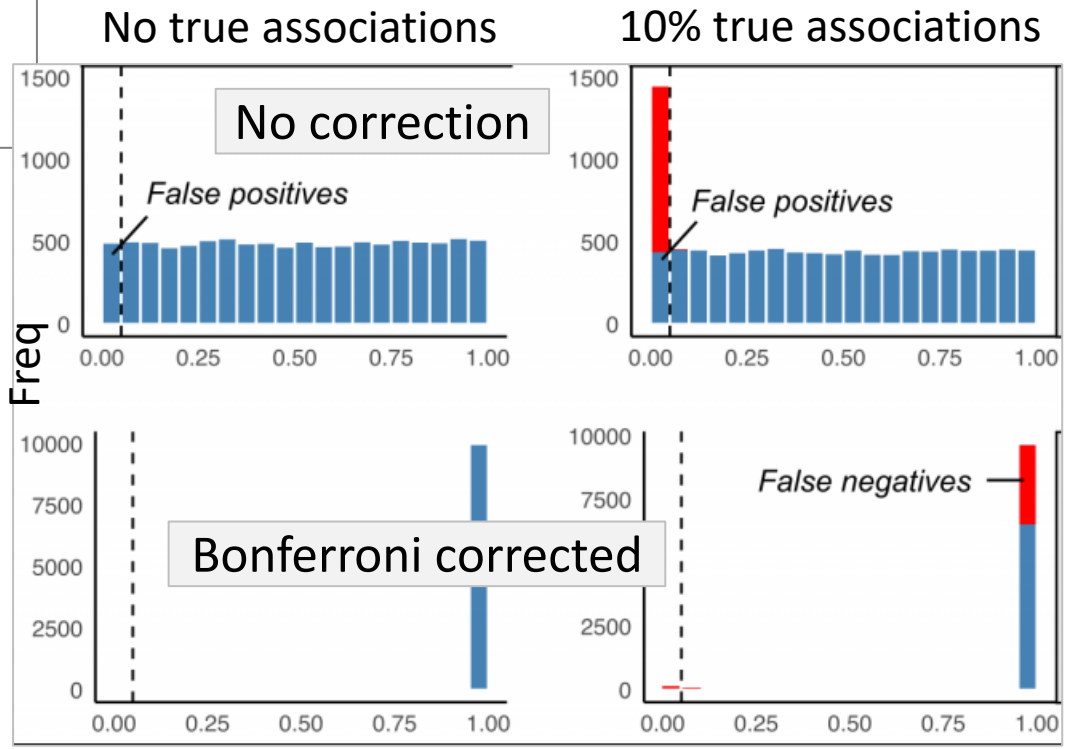
- Measurable traits (height, BMI)
- Diseases (diabetes, cancer)
- Yield, resistance (milk, fruit, crops)

Association testing is (1) computationally expensive and (2) prone to false positives due to multiple testing



Millions to billions of genetic variants to be tested

Multiple testing correction is needed to maintain false positives

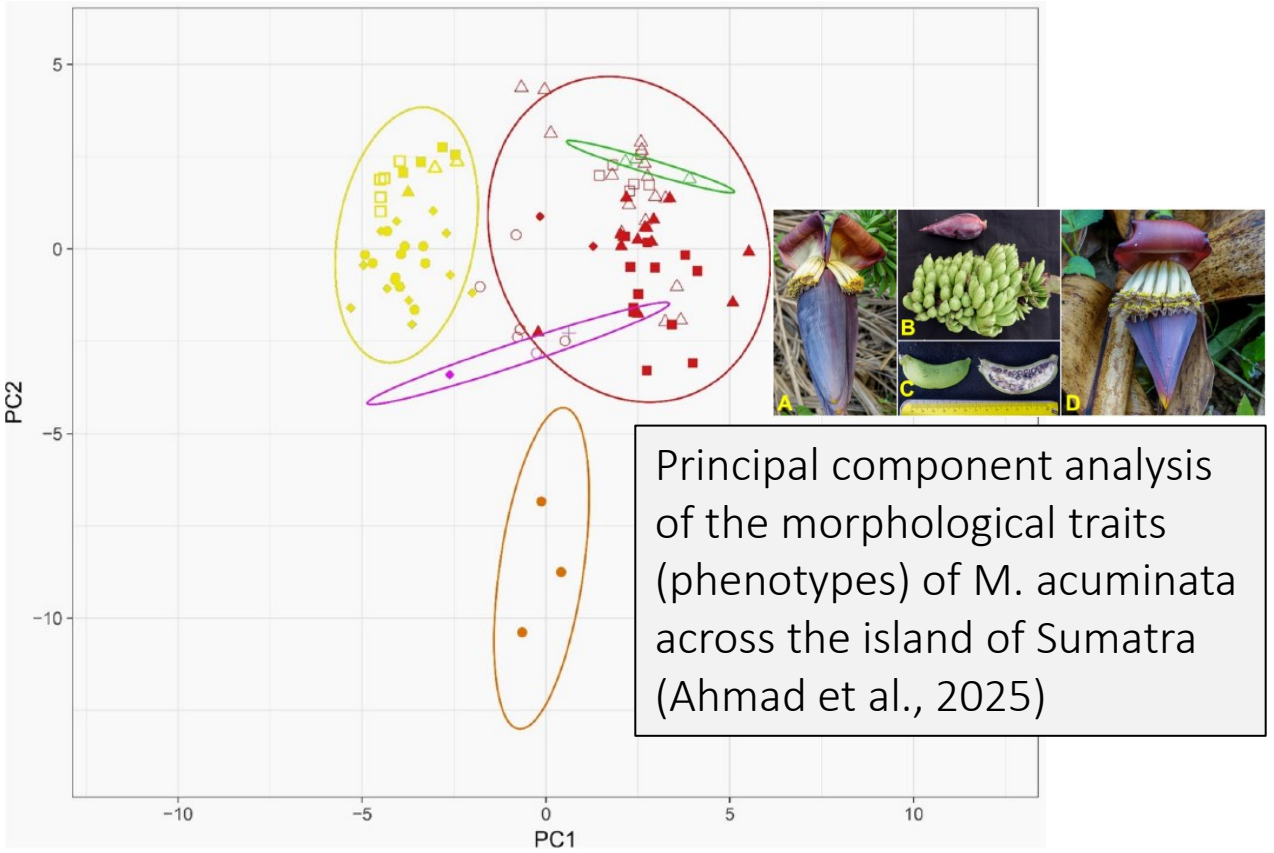


Resource-Efficient Approach for Large-Scale Genome-Wide Association Studies

- Bananas are an essential fruit worldwide due to their rich nutrients. Southeast Asia is a major part of the origin of bananas (*Musa acuminata*)
- **Genetic exploration of wild ancestors of banana cultivars** is important for conservation and breeding

- Banana producers look for specific morphology, fruit quality, yield, and agronomic features
- **The decrease in sequencing cost allows whole-genome sequencing in bananas** (tens of millions of variants reported)
- **High genetic diversity on morphological traits across wild varieties**

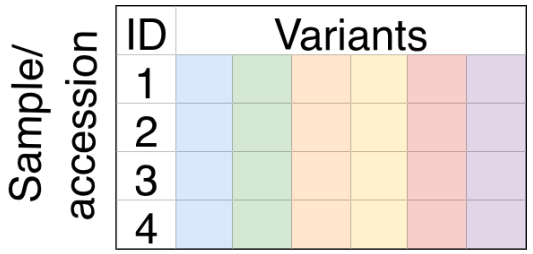
The large number of variants in whole-genome sequencing opens possibilities for discoveries, but also poses the GWAS challenges

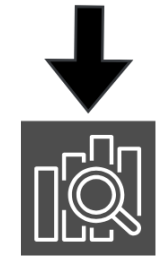


Principal component analysis of the morphological traits (phenotypes) of M. acuminata across the island of Sumatra (Ahmad et al., 2025)

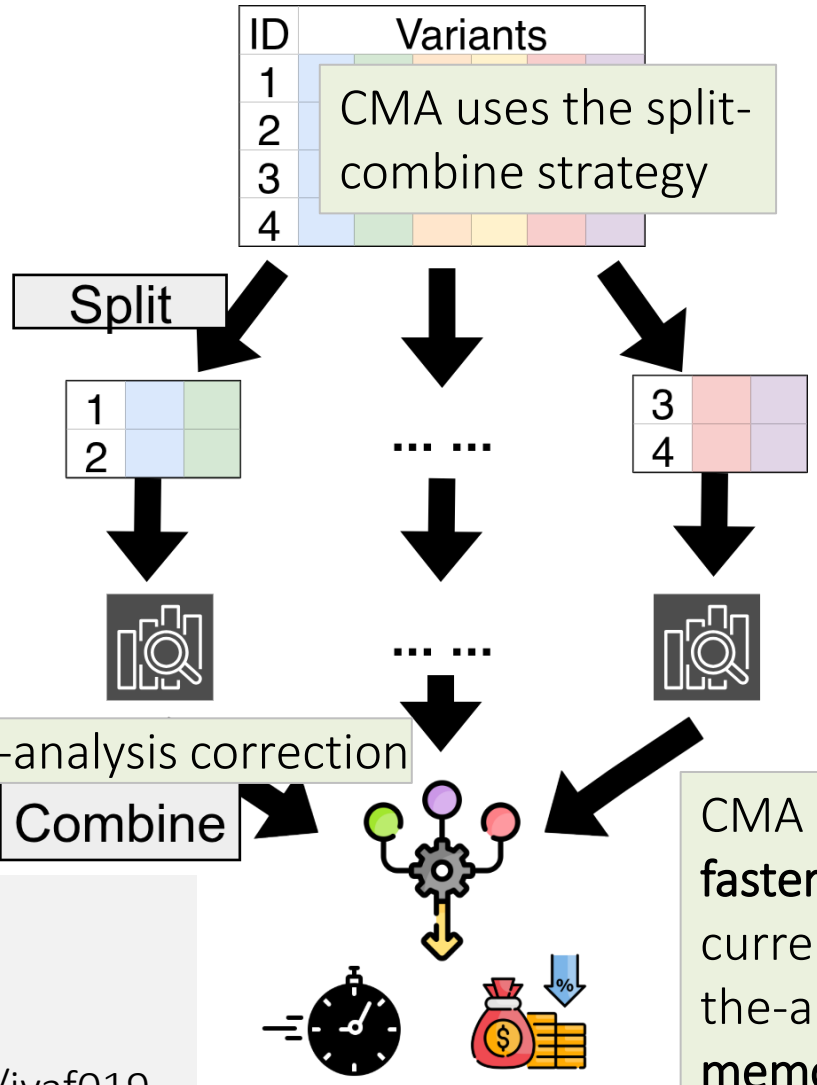
Proposed Method: CMA: Divide-And-Conquer GWAS

Traditional GWAS uses whole samples and variants

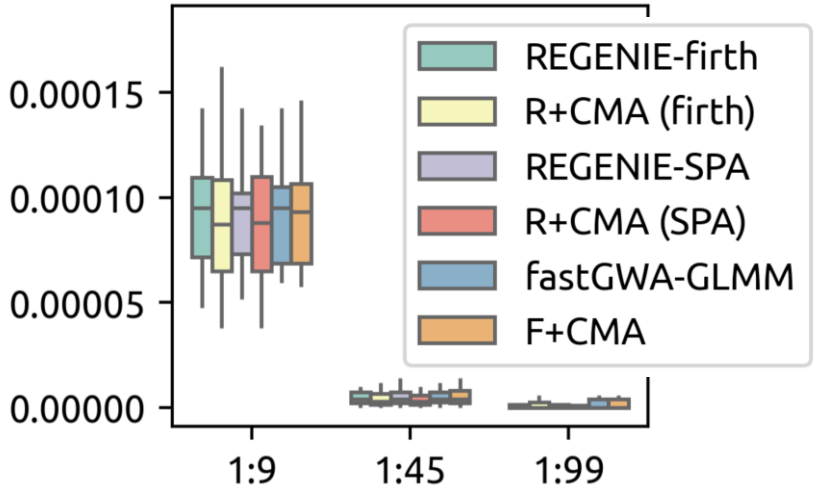


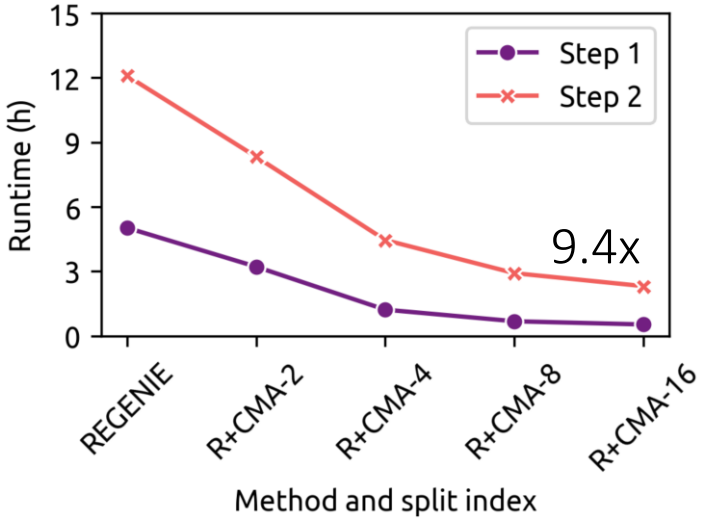


GWAS



Type 1 error rates on human datasets





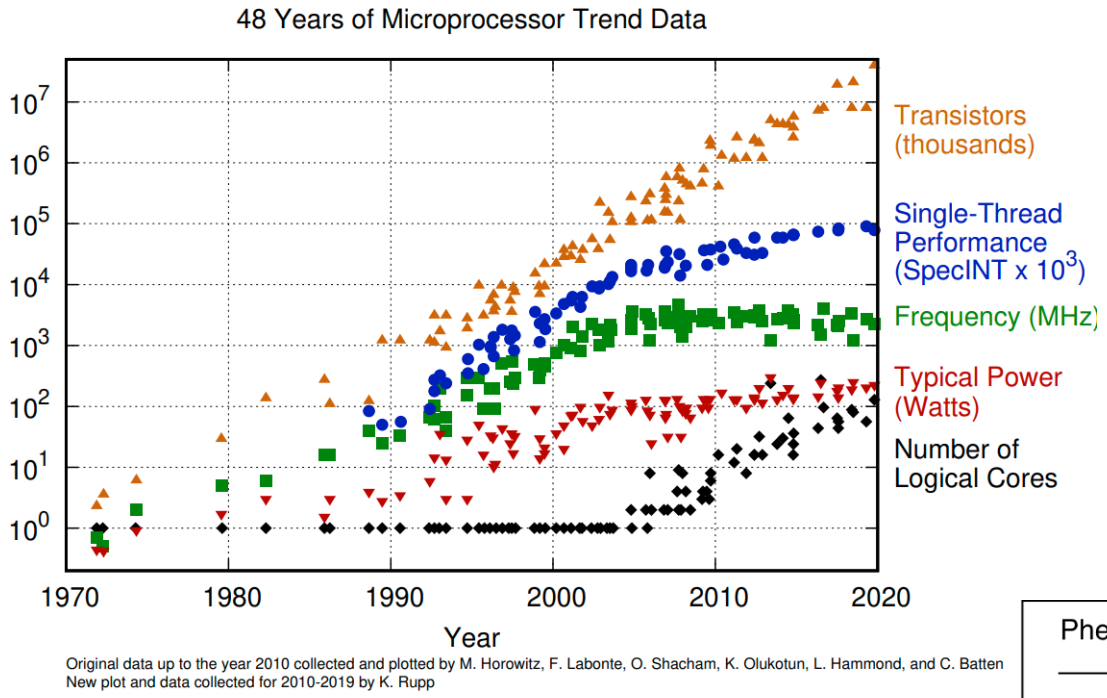
CMA is **much faster** than the current state-of-the-art, with **less memory usage**

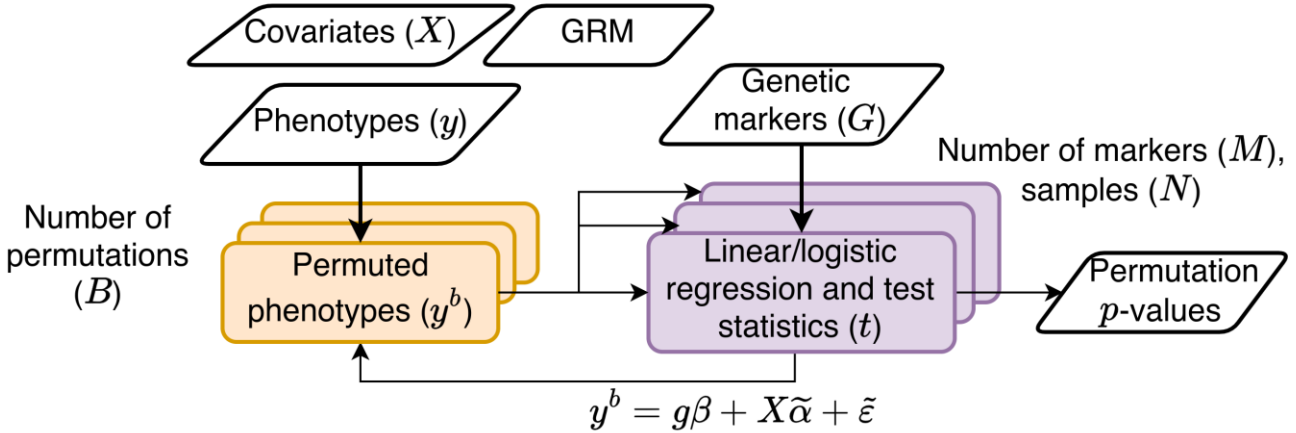
CMA is available on https://git.ecdf.ed.ac.uk/cma
Publication on GENETICS: https://doi.org/10.1093/genetics/iyaf019

Proposed Method: GWA-X: GWAS permutation testing on GPUs

Permutation testing is the gold standard for multiple testing correction to reduce false positive rates

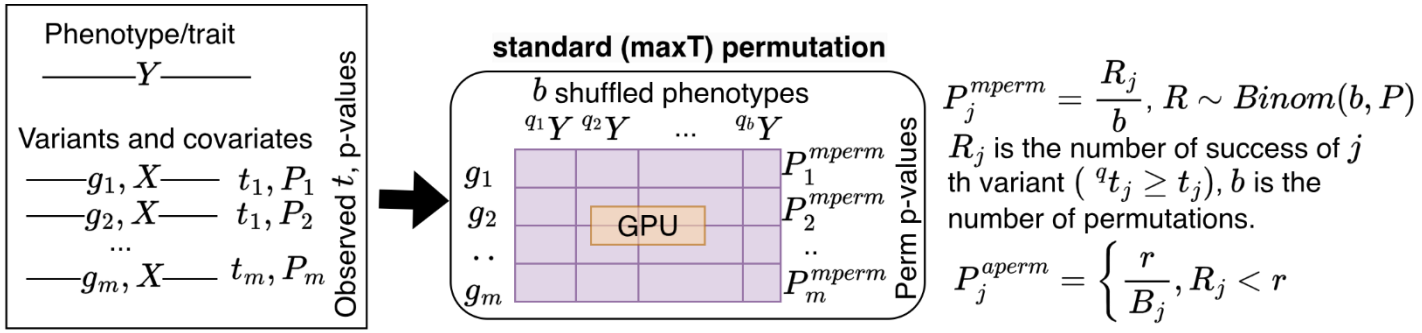
GWAS permutation testing is often impractical due to the large genetic data





The **number of regressions** increases from (M) to $(M \cdot B)$

The complexity of matrix multiplication and inverse: $O(N^3 \cdot M \cdot B)$



$$P_j^{mperm} = \frac{R_j}{b}, R \sim Binom(b, P)$$

R_j is the number of success of j th variant ($^{q}t_j \geq t_j$), b is the number of permutations.

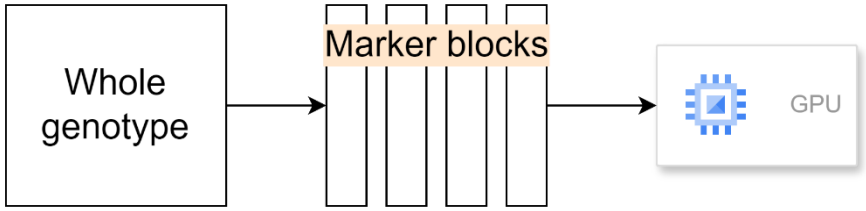
$$P_j^{aperm} = \begin{cases} \frac{r}{B_j}, R_j < r \end{cases}$$

GWA-X is available on https://git.ecdf.ed.ac.uk/magung/gwa-x

Publication on bioRxiv: https://doi.org/10.1101/2024.09.15.613119

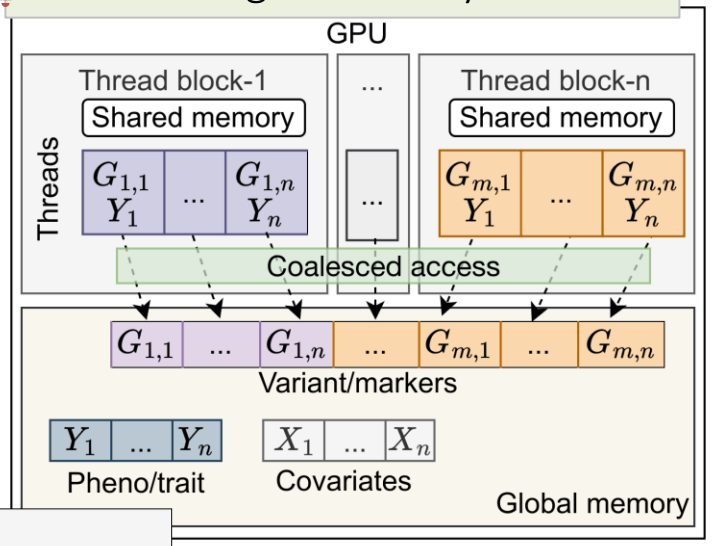
Proposed Method: GPU Optimization and Performance

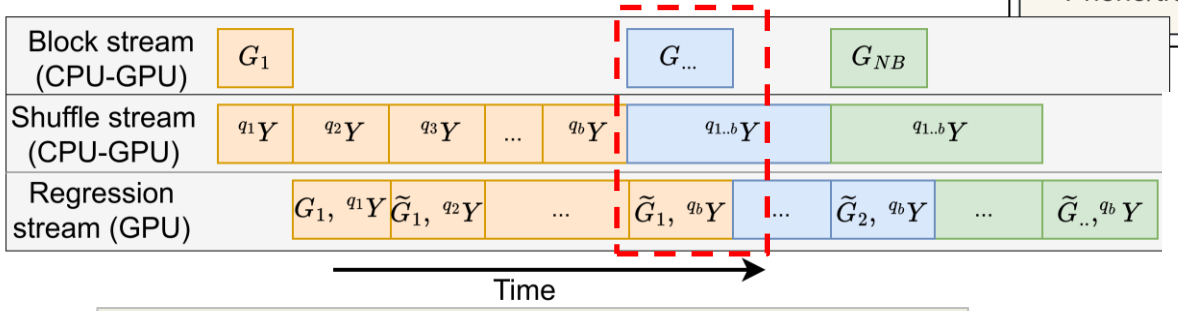
Genotype (G) with m markers/variants, n samples/accessions



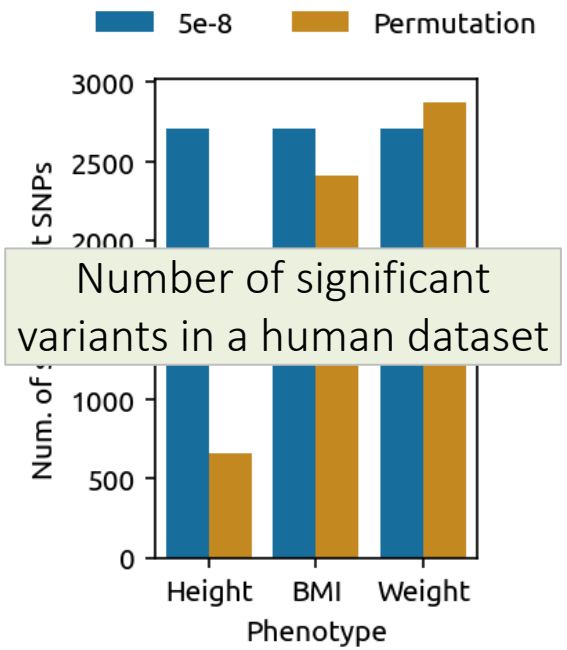
Genotype partitioning to handle a large genotype

Batched regressions by markers

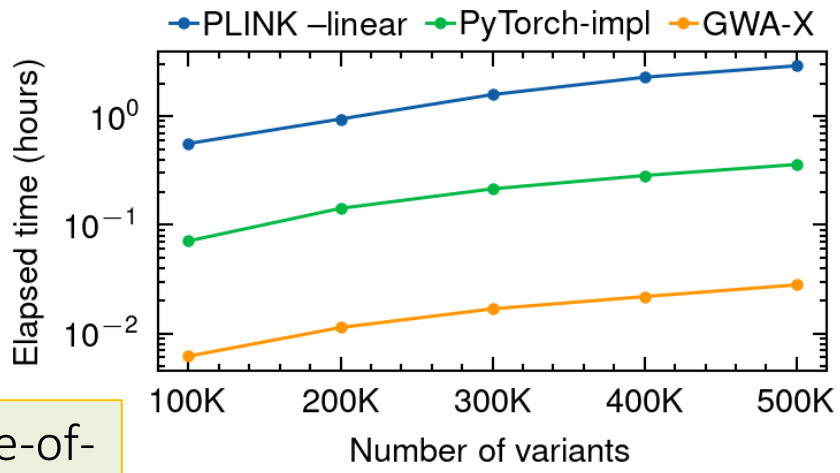


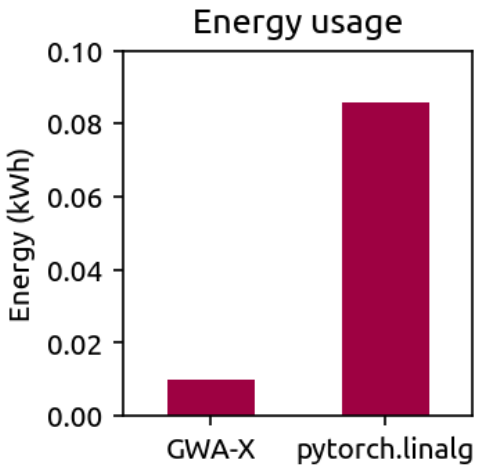


Overlapping data transfers with kernel computation to reduce data transfer delays



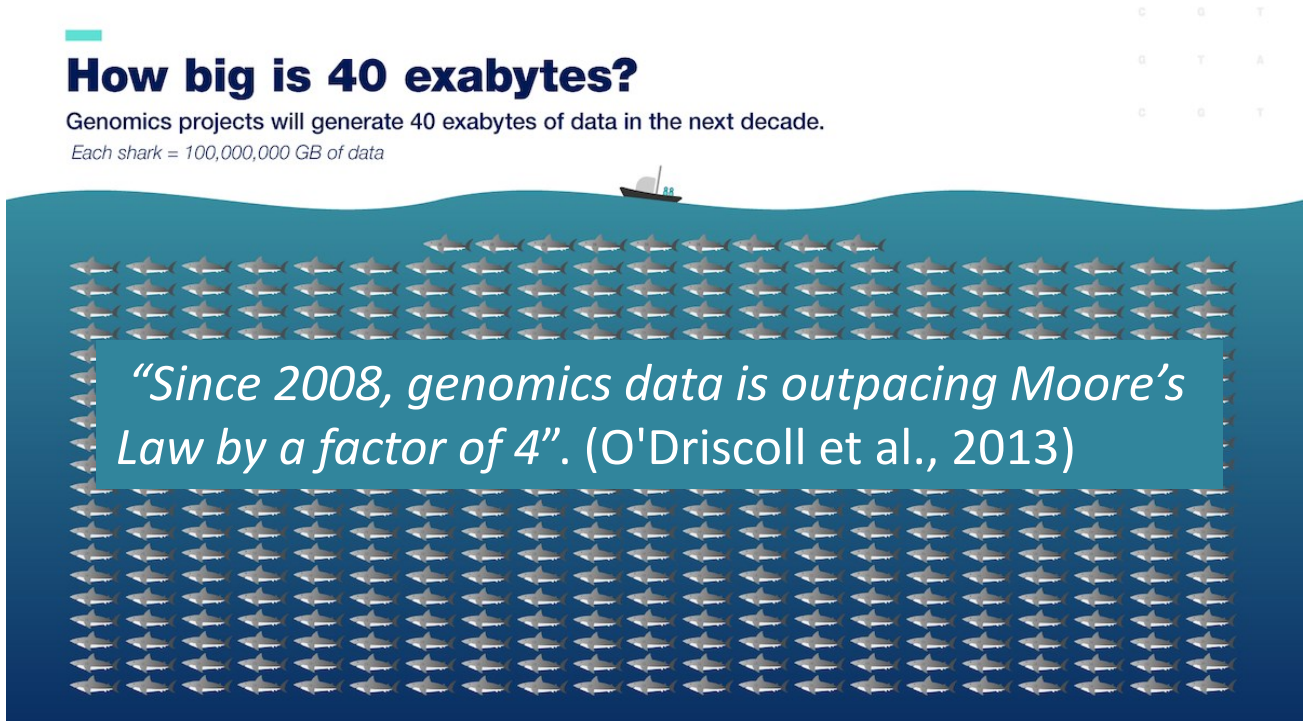
Number of significant variants in a human dataset



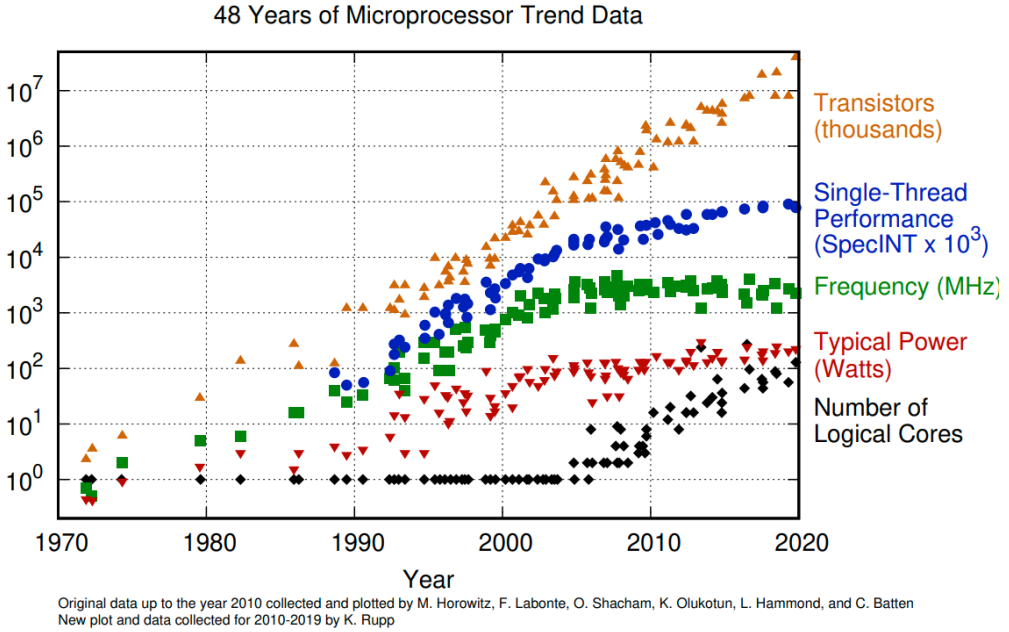


GWA-X is 100x times faster than the current state-of-the-art on CPUs (88.4% less energy usage)

Impact: Efficient and Scalable Approach to Current and Future GWAS



Computer architecture shifts to using accelerators, e.g., GPUs



- The growth of the genomics data size scale requires a resource-efficient method scalable to recent computer architectures
- We anticipate that further advances in developing efficient software in GWAS will adopt our approach to mitigate the challenges associated with large data size

Impact: Democratize the Use of ICT Systems for Genomic Research

"Through the use of cloud computing, data commons can support large-scale data, but this also creates sustainability challenges, due to the cost of large-scale storage and compute". (Grossman, 2019)

Databases

1982–present

- Data repository
- Researchers download data

Data commons

2014–2024

- Supports large datasets and data-intensive computing with cloud computing
- Workspaces
- Common data models
- Core data services
- Data and commons governance
- Harmonized data
- Data sharing
- Reproducible research

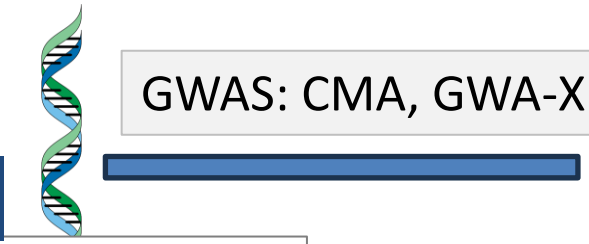
2010–2020

- Supports large datasets and data-intensive computing with cloud computing
- Researchers can analyze data with their own virtual workspaces and applications and collaborate with other researchers with collaborative workspaces (data do not have to be downloaded to be analyzed)

Trends in Genetics

Resource-efficient GWAS will promote the sustainability goals in ICT and benefit the research community, especially researchers with limited resources and budgets

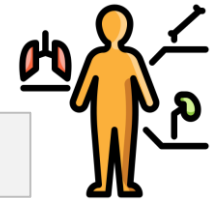
Impact: Fostering Collaboration for Better Society

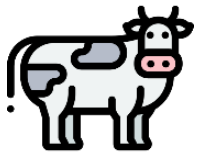


GWAS: CMA, GWA-X

Genotype

Variation in the genome between individuals





Phenotype

- Measurable traits (height, blood pressure)
- Diseases (diabetes, cancer)
- Yield, resistance (milk, fruit, crops)

- We anticipate collaboration between researchers from multiple disciplines conducting **genomic research in humans, animals, and plants**
- Our GWAS project on wild bananas is one example of the collaboration
  - **Identifying the genetic variants controlling agronomic traits** in plants is important for developing new varieties with desirable traits, e.g., higher yield and quality

Research output: GWAS Pipeline for Bananas

Public dataset (genotype-phenotype)
- Trait focus: *Fusarium* resistance, morphology, fruit quality, and yield

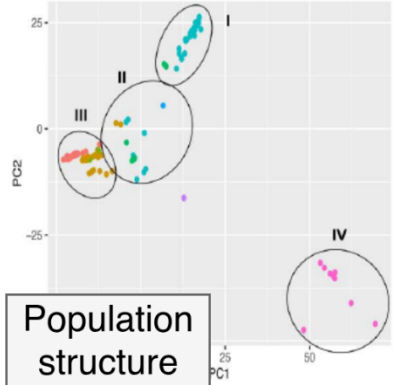
Collab: BRIN's Botany lab

1. Data collection (*Musa acu.*)

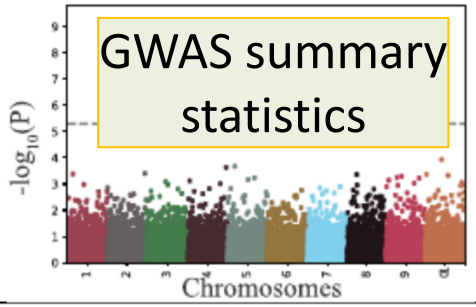
2. Genotyping

3. Phenotyping

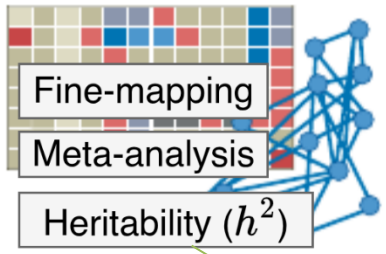
4. Quality control



5. Association testing (CMA, GWA-X)



e. Post-GWAS analysis



Genetic basis of *M. acuminata*

Linkage and kinship reference across varieties

Category	Trait
Morphology	Pseudostem height
	Single leaf width
	Time from flowering to harvest
Fruit quality	pH
	Total soluble solids
	Titratable acidity
Yield	Hands weight
	Number of hands in a bunch
	Fruit length
	Pulp percentage
	Pulp dry weight percentage
	Peel thickness

Conclusion

- GWAS is a powerful tool for unravelling genetic variants associated with phenotypes or traits of interest, such as morphological traits, diseases, and agronomic traits
- The emerging scale of genomic data size opens possibilities for discoveries, but also poses a big data challenge
- We present a resource-efficient approach, consisting of CMA and GWA-X, to mitigate the GWAS challenge associated with the large-scale data size, including
  - A divide-and-conquer method for partitioning a large GWAS workload into workloads of smaller subsets to be run on smaller-scale computers
  - A GWAS permutation testing approach on GPUs
- One of our ongoing projects is to adopt our approach for GWAS on wild bananas (*M. acuminata*), an essential fruit in the ASEAN region, aimed at investigating genetic variants associated with disease resistance and better yield
- We anticipate that further development of GWAS tools will be based on our approach for the large datasets of humans, plants, and animals

Collaborators

Thank you!

CMA and GWA-X are open source
You are welcome to contribute!

CMA

GWA-X

Thanks to our collaborators:

CMA & GWA-X

- Tenesa group (University of Edinburgh, UK)
- Dr. Hyojung Paik (KISTI, South Korea)

Genomic diversity of Indonesian wild bananas

- Fajaruddin Ahmad (Research Center of Applied Botany, BRIN)