High-throughput Algorithms for Genome-Wide Association Studies

Diego Fabregat-Traver and Prof. Paolo Bientinesi
In collaboration with Dr. Yurii Aulchenko

AICES, RWTH Aachen
fabregat@aices.rwth-aachen.de

BGRS/SB, June 25th – 29th, 2012
Novosibirsk, Russia

Genome-wide association studies

Aim at...

- Identify association between genetic markers and phenotypes of interest
- Significant association highlights genomic regions involved in the control of a trait

Genome-wide association studies

Aim at...

- Identify association between genetic markers and phenotypes of interest
- Significant association highlights genomic regions involved in the control of a trait

How?

- Variance Components based on linear mixed-models

$$\begin{cases} b &= (X^T M^{-1} X)^{-1} X^T M^{-1} y \\ M &= \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- $X \in R^{n \times p}$, single-nucleotide polymorphism
- $y \in R^n$, phenotype
- $h^2, \sigma^2 \in R$, heritability and residual variance
- $\Phi \in R^{n \times n}$, kinship matrix
- $b \in R^p$, genetic effect

- $n \in [1{,}000, ..., 10{,}000]$
- $p \in [1, ..., 20]$

$$\begin{cases} b_i & = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M & = \sigma^2(h^2\Phi + (1-h^2)I) \end{cases}$$

with $1 \le i \le m$

- $X \in R^{n \times p}$, single-nucleotide polymorphism
- $y \in R^n$, phenotype
- $h^2, \sigma^2 \in R$, heritability and residual variance
- $\Phi \in R^{n \times n}$, kinship matrix
- $b \in R^p$, genetic effect

- $n \in [1{,}000, ..., 10{,}000]$
- $p \in [1, ..., 20]$
- $m \in [10^6, ..., 10^7]$

Scenario 1: Single-trait analysis

$$\begin{cases} b_{ij} = (X_i{}^T M_j{}^{-1} X_i)^{-1} X_i{}^T M_j{}^{-1} y_j & \text{with } 1 \leq i \leq m \\ M_j = \sigma_j^2 (h_j^2 \Phi + (1 - h_j^2) I) & \text{and } 1 \leq j \leq t. \end{cases}$$

- $X \in R^{n \times p}$, single-nucleotide polymorphism
- $y \in R^n$, phenotype
- $h^2, \sigma^2 \in R$, heritability and residual variance
- $\Phi \in R^{n \times n}$, kinship matrix
- $b \in R^p$, genetic effect

- $n \in [1{,}000, ..., 10{,}000]$
- $p \in [1, ..., 20]$
- $m \in [10^6, ..., 10^7]$
- t is 1 or $\approx 10^5$

Scenario 2: Multiple-trait analysis

Scenario 1

- Sample size: $10{,}000$
- # covariates: 2
- # SNPs: $36{,}000{,}000$
- # phenotypes: 1

Scenario 1

- Sample size: $10{,}000$
- # covariates: 2
- # SNPs: $36{,}000{,}000$
- # phenotypes: 1
- Data set: ≈ 3 TB

Tool	Time
EMMAX	40 days
GWFGLS	20 days
FaST-LMM	53 hours

Scenario 1

- Sample size: $10{,}000$
- # covariates: 2
- # SNPs: $36{,}000{,}000$
- # phenotypes: 1
- Data set: ≈ 3 TB

Tool	Time
EMMAX	40 days
GWFGLS	20 days
FaST-LMM	53 hours

Scenario 2

- Sample size: $1{,}000$
- # covariates: 2
- # SNPs: $1{,}000{,}000$
- # phenotypes: $100{,}000$

Scenario 1

- Sample size: $10{,}000$
- # covariates: 2
- # SNPs: $36{,}000{,}000$
- # phenotypes: 1
- Data set: ≈ 3 TB

Tool	Time
EMMAX	40 days
GWFGLS	20 days
FaST-LMM	53 hours

Scenario 2

- Sample size: $1{,}000$
- # covariates: 2
- # SNPs: $1{,}000{,}000$
- # phenotypes: $100{,}000$
- Data set: ≈ 3 TB

Tool	Time
EMMAX	≈ 3 years
FaST-LMM	> 1 year
GWFGLS	≈ 9 months

Scenario 1

- Sample size: $10{,}000$
- # covariates: 2
- # SNPs: $36{,}000{,}000$
- # phenotypes: 1
- Data set: ≈ 3 TB

Tool	Time
EMMAX	40 days
GWFGLS	20 days
FaST-LMM	53 hours
CLAK-CHOL	?

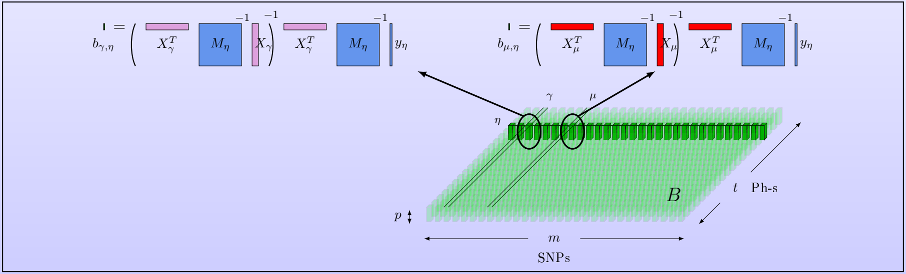
Scenario 2

- Sample size: $1{,}000$
- # covariates: 2
- # SNPs: $1{,}000{,}000$
- # phenotypes: $100{,}000$
- Data set: ≈ 3 TB

Tool	Time
EMMAX	≈ 3 years
FaST-LMM	> 1 year
GWFGLS	≈ 9 months
CLAK-EIG	?

Can we do better? Yes, HOW?

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$



Algorithms - CLAK-CHOL

RWTH AACHEN UNIVERSITY

Single phenotype analysis ($t = 1$)

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- Typically, based on eig(Φ): $O(n^3)$

AICES

RWTH AACHEN UNIVERSITY

Single phenotype analysis ($t = 1$)

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- Typically, based on eig(Φ): $O(n^3)$
- CLAK-CHOL based on chol(M): $O(n^3)$

AICES

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- Typically, based on eig(Φ): $O(n^3)$
- CLAK-CHOL based on chol(M): $O(n^3)$

$O(n^3) = O(n^3)$?

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- Typically, based on eig(Φ): $O(n^3)$
- CLAK-CHOL based on chol(M): $O(n^3)$

$O(n^3) = O(n^3)$?

	Chol	Eig
# operations	$\frac{1}{3}n^3$	$\frac{10}{3}n^3$
Efficient?	+	−
Scalable?	+	−

Algorithms - CLAK-CHOL

Single phenotype analysis ($t = 1$)

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

- Typically, based on eig(Φ): $O(n^3)$
- CLAK-CHOL based on chol(M): $O(n^3)$

$O(n^3) = O(n^3)$?

	Chol	Eig
# operations	$\frac{1}{3}n^3$	$\frac{10}{3}n^3$
Efficient?	+	−
Scalable?	+	−

Asymptotical cost is only part of the story

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

Traditional

$ZWZ^T = \Phi$

$X_i' := Z^T X_i$ $\quad m \times (2n^2)$

AICES

Algorithms - CLAK-Chol

Single phenotype analysis ($t = 1$)

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

Traditional

$ZWZ^T = \Phi$

$X_i' := Z^T X_i \qquad m \times (2n^2)$

CLAK-Chol

$LL^T = M$

$X_i' := L^{-1} X_i \qquad m \times (n^2)$

$$\begin{cases} b_i = (X_i^T M^{-1} X_i)^{-1} X_i^T M^{-1} y \\ M = \sigma^2 (h^2 \Phi + (1 - h^2) I) \end{cases}$$

Traditional

$ZWZ^T = \Phi$

$X_i' := Z^T X_i$ $\qquad m \times (2n^2)$

CLAK-CHOL

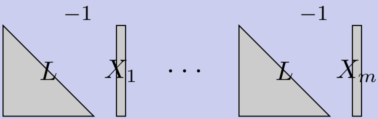
$LL^T = M$

$X_i' := L^{-1} X_i$ $\qquad m \times (n^2)$

The constant makes a big difference

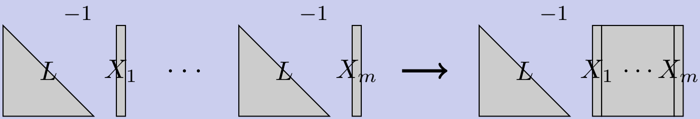
Algorithms - CLAK-CHOL

Single phenotype analysis ($t = 1$)

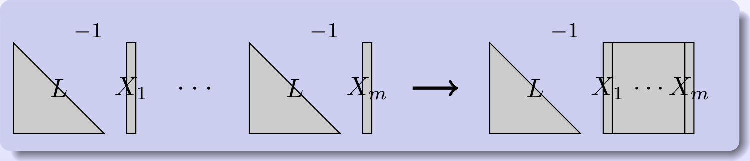


Algorithms - CLAK-CHOL

Single phenotype analysis ($t = 1$)



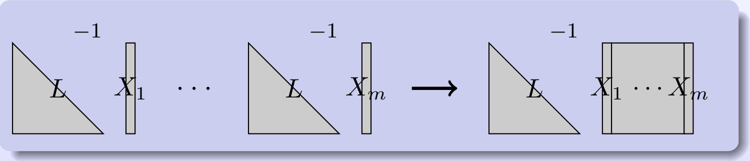
Single phenotype analysis ($t = 1$)



- Many TRSVs vs one single large TRSM

Algorithms - CLAK-CHOL

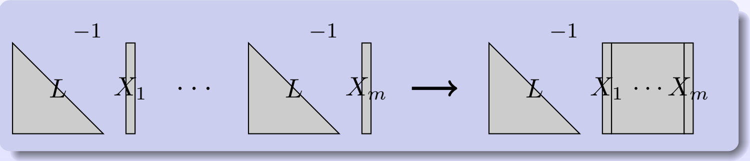
Single phenotype analysis ($t = 1$)



- Many TRSVs vs one single large TRSM
- Same amount of computation

RWTH AACHEN UNIVERSITY

Single phenotype analysis ($t = 1$)



- Many TRSVs vs one single large TRSM
- Same amount of computation
- Different efficiency

Operation	Efficiency	Scalability
One TRSM	90%	+
m TRSVs	15%	–

To take away

Yes, asymptotical cost is important, but...

- Careful with the **constants** ($\frac{1}{3}n^3$ vs $\frac{10}{3}n^3$, $2n^2$ vs n^2)
- The **efficiency** of the operations plays an important role
- The **scalability** of the operations is also important

Out-of-core algorithms

Problem

- Data does not fit in RAM (terabytes of data)
- Loading data from disk is slow $\rightarrow$ processor stalls

Out-of-core algorithms

Problem

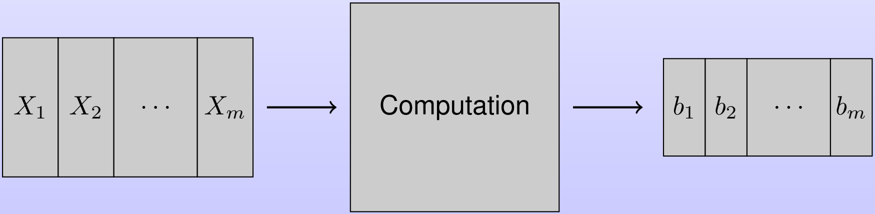
- Data does not fit in RAM (terabytes of data)
- Loading data from disk is slow $\rightarrow$ processor stalls

Approach

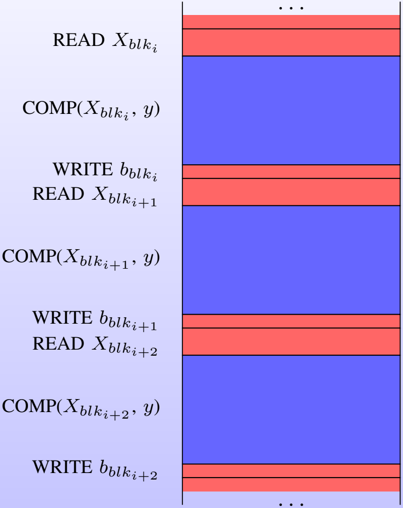
- Overlapping vs Non-overlapping
- **Goal**: hide the overhead due to data transfers

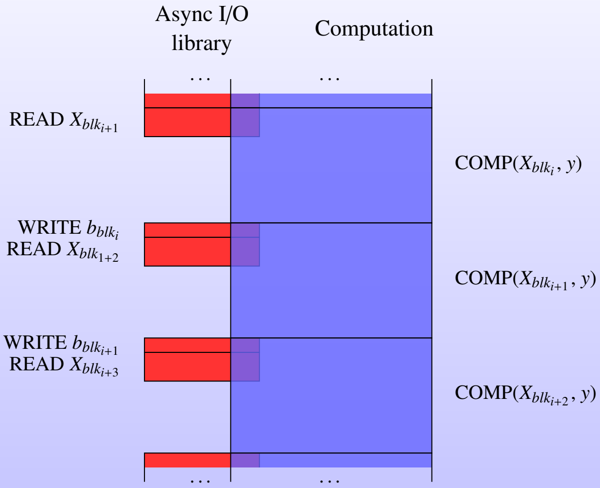
We regard the problem as:

- an input stream of X's (SNPs)
- an output stream of b's (the computed effects)



Non-overlapping





Async I/O library
Computation
READ $X_{blk_{i+1}}$
WRITE b_{blk_i}
READ $X_{blk_{1+2}}$
WRITE $b_{blk_{i+1}}$
READ $X_{blk_{i+3}}$
COMP(X_{blk_i}, y)
COMP($X_{blk_{i+1}}$, y)
COMP($X_{blk_{i+2}}$, y)

To take away

- Non-overlapping: 10% - 15% overhead

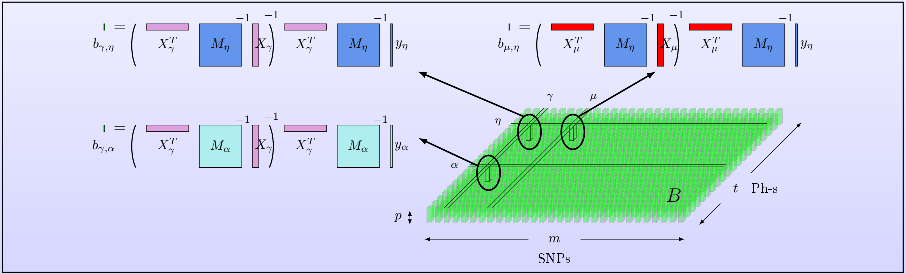
To take away

- Non-overlapping: 10% - 15% overhead
- Try to overlap as much as possible to minimize overhead

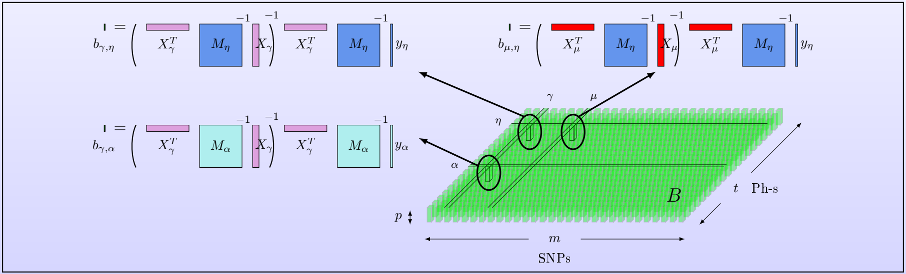
To take away

- Non-overlapping: 10% - 15% overhead
- Try to overlap as much as possible to minimize overhead
- Perfect overlapping:
  - Data on disk but...
  - Efficiency as if data in RAM!

Multiple phenotype analysis ($t \approx 10^5$)



Multiple phenotype analysis ($t \approx 10^5$)



- Traditionally: run single-phenotype routines for each phenotype
- CLAK-EIG considers the whole 2D sequence in its entirety

First step:

- 1 eigendecomposition vs t Cholesky factorizations
- SNPs premultiplied ($Z^T X_i$) only once and reused

First step:

- 1 eigendecomposition vs t Cholesky factorizations
- SNPs premultiplied ($Z^T X_i$) only once and reused

Second step:

- Cost of traditional algorithms: $t(mn^2)$
- CLAK-EIG linear with all dimensions: $t(mn)$

First step:

- 1 eigendecomposition vs t Cholesky factorizations
- SNPs premultiplied ($Z^T X_i$) only once and reused

Second step:

- Cost of traditional algorithms: $t(mn^2)$
- CLAK-EIG linear with all dimensions: $t(mn)$
- There is much more: fine tuning, parallelism, ...

First step:

- 1 eigendecomposition vs t Cholesky factorizations
- SNPs premultiplied ($Z^T X_i$) only once and reused

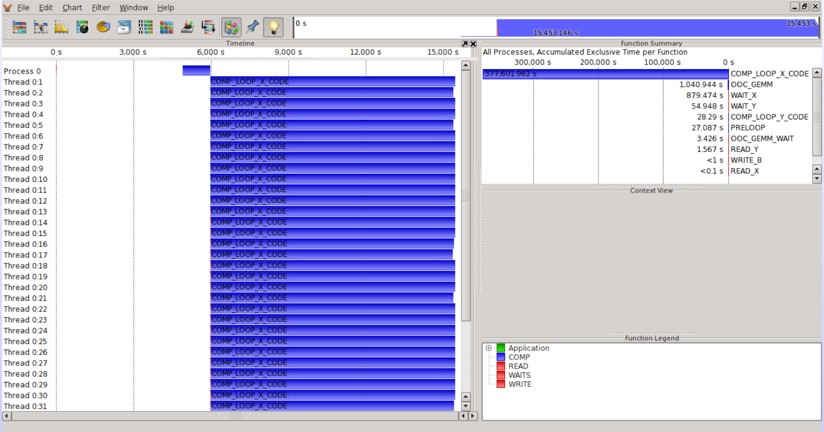
Second step:

- Cost of traditional algorithms: $t(mn^2)$
- CLAK-Eig linear with all dimensions: $t(mn)$
- There is much more: fine tuning, parallelism, ...

Out-of-core: a careful tuning of the overlapping is **VERY** important.

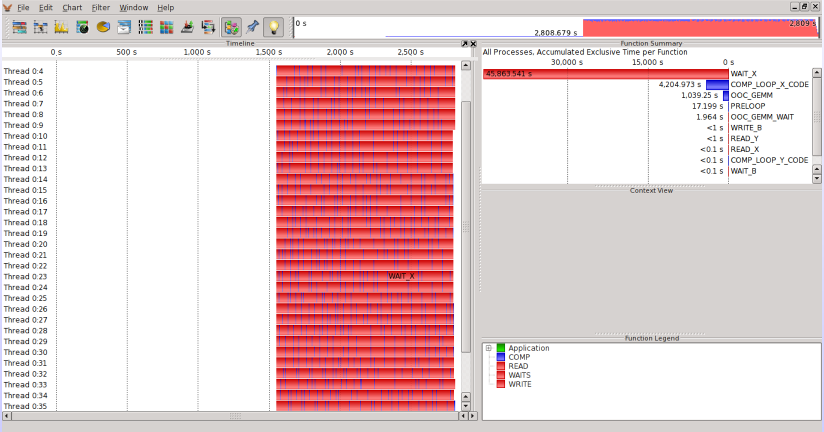
Algorithms - CLAK-EIG

Multiple phenotype analysis ($t \approx 10^5$)



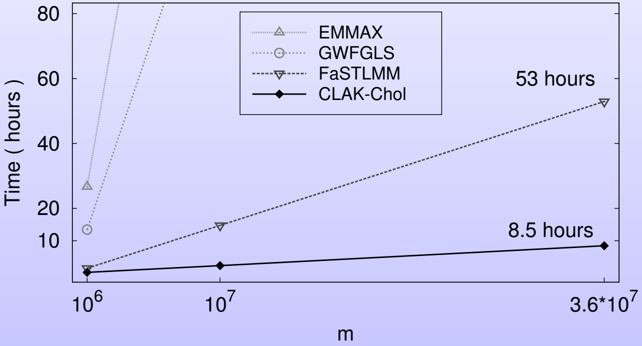
RWTH AACHEN UNIVERSITY

Multiple phenotype analysis ($t \approx 10^5$)



AICES

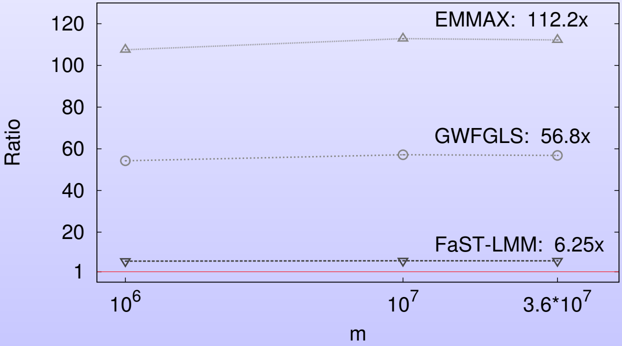
- Sample size: $10{,}000$
- # covariates: 2
- 12 cores



Experimental results

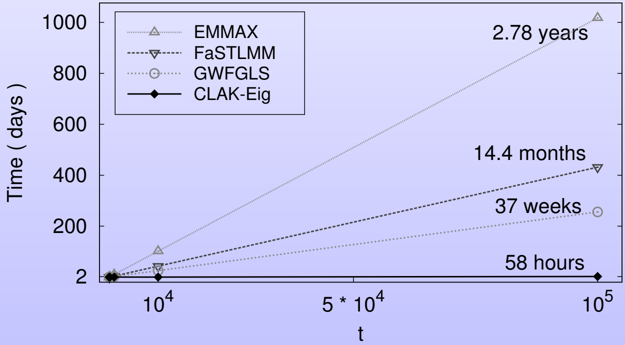
Scenario 1: Single phenotype

- Sample size: $10{,}000$
- # covariates: 2
- 12 cores



Scenario 2: Multiple phenotype

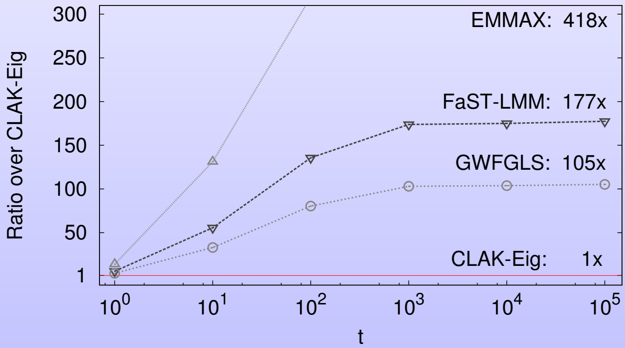
- Sample size: $1{,}000$
- # SNPs: $1{,}000{,}000$
- # covariates: 2
- 12 cores



Experimental results

Scenario 2: Multiple phenotype

- Sample size: $1{,}000$
- # SNPs: $1{,}000{,}000$
- # covariates: 2
- 12 cores



Conclusions and Future work (I)

Two different scenarios: Two different algorithms

Single phenotype: CLAK-CHOL

Multiple phenotype: CLAK-EIG

Conclusions and Future work (I)

RWTH AACHEN UNIVERSITY

Two different scenarios: Two different algorithms

Single phenotype: CLAK-Chol

Multiple phenotype: CLAK-Eig

Guidelines for High Performance

- Asymptotical cost is not enough
- Number of arithmetic operations
- Efficiency and scalability of the operations
- Perfect overlapping of I/O with computation $\rightarrow$ no stalls

Conclusions and Future work (I)

Two different scenarios: Two different algorithms

Single phenotype: CLAK-CHOL

Multiple phenotype: CLAK-EIG

Guidelines for High Performance

- Asymptotical cost is not enough
- Number of arithmetic operations
- Efficiency and scalability of the operations
- Perfect overlapping of I/O with computation $\rightarrow$ no stalls
- **Very important**: look at the problem as a whole

Conclusions and Future work (II)

Results

- Single phenotype: CLAK-CHOL - Speedup > 6x
- Multiple phenotype: CLAK-EIG - Speedup > 100x
  - ➥ Years/Months to hours!!!

Conclusions and Future work (II)

Results

- Single phenotype: CLAK-CHOL - Speedup > 6x
- Multiple phenotype: CLAK-EIG - Speedup > 100x
  - ➥ Years/Months to hours!!!

Future Work

- Reduction of complexity by exploiting sparsity
- More computational power: GPU, MPI

Thanks to:

- Dr. Edoardo Di Napoli
- Matthias Petschow
- Roman Iakymchuk
- Elmar Peise
- Lucas Beyer

Financial support from the **Deutsche Forschungsgemeinschaft** (German Research Association) through grant GSC 111 is gratefully acknowledged.

Deutsche Forschungsgemeinschaft
DFG