MaRS - Matrix of RNA-Seq

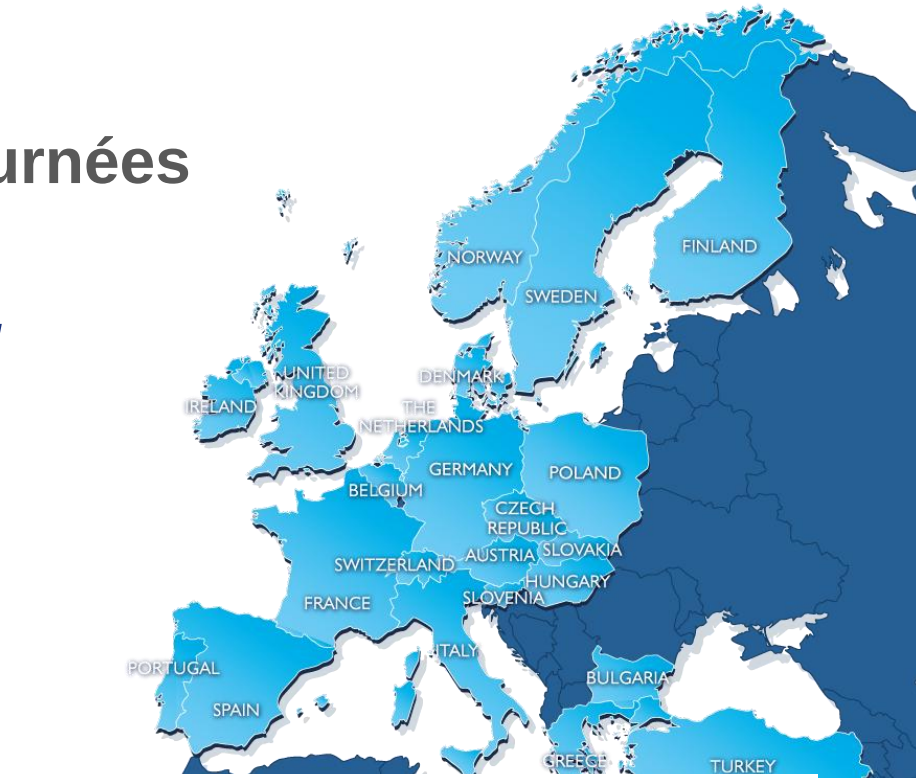
Cancéropôle Grand Sud-Ouest

12èmes Journées

Fabien PIERRAT – Bertrand CIROU

November, the 25th 2016

ACOBIOM – CINES – SHAPE



CINES Centre Informatique National de l'Enseignement Supérieur

ACOBIOM

PRACE SHAPE SME HPC Adoption Programme in Europe

MaRS - Matrix of RNA-Seq

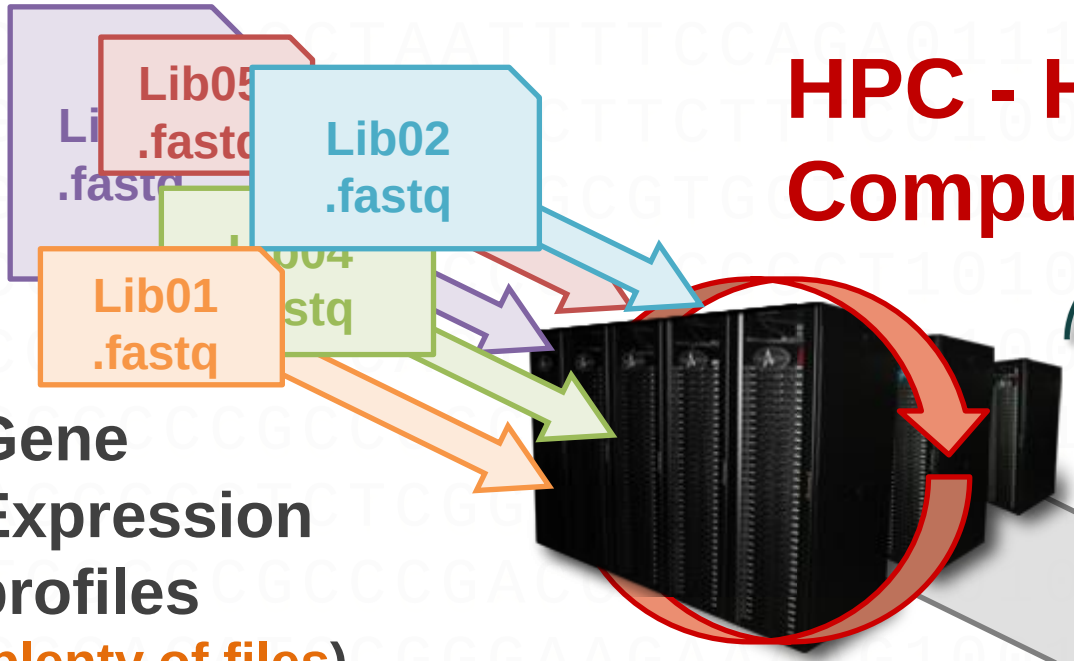
- A large set of gene expression data (RNA-Seq) is published & available

- Represents a big source of information for:
  - the discovery of new Biomarkers
  - the validation of Biomarkers

- But, due to
  - Huge amount of data, huge size of files
  - various methods of analysis used

- These data are:
  - challenging to compare
  - unexploited all together

The MaRS Project



HPC - High Performance Computing (only 1 method)

Gene Expression profiles (plenty of files)

MaRS (only 1 matrix)

	Lib01	Lib02	Lib03	Lib04	Lib05
Gene001	0	5	76	10	0
Gene002	23	0	2	6	9
Gene003	2	95	2	0	2
Gene004	11	4	0	33	1
Gene005	0	5	5	0	53

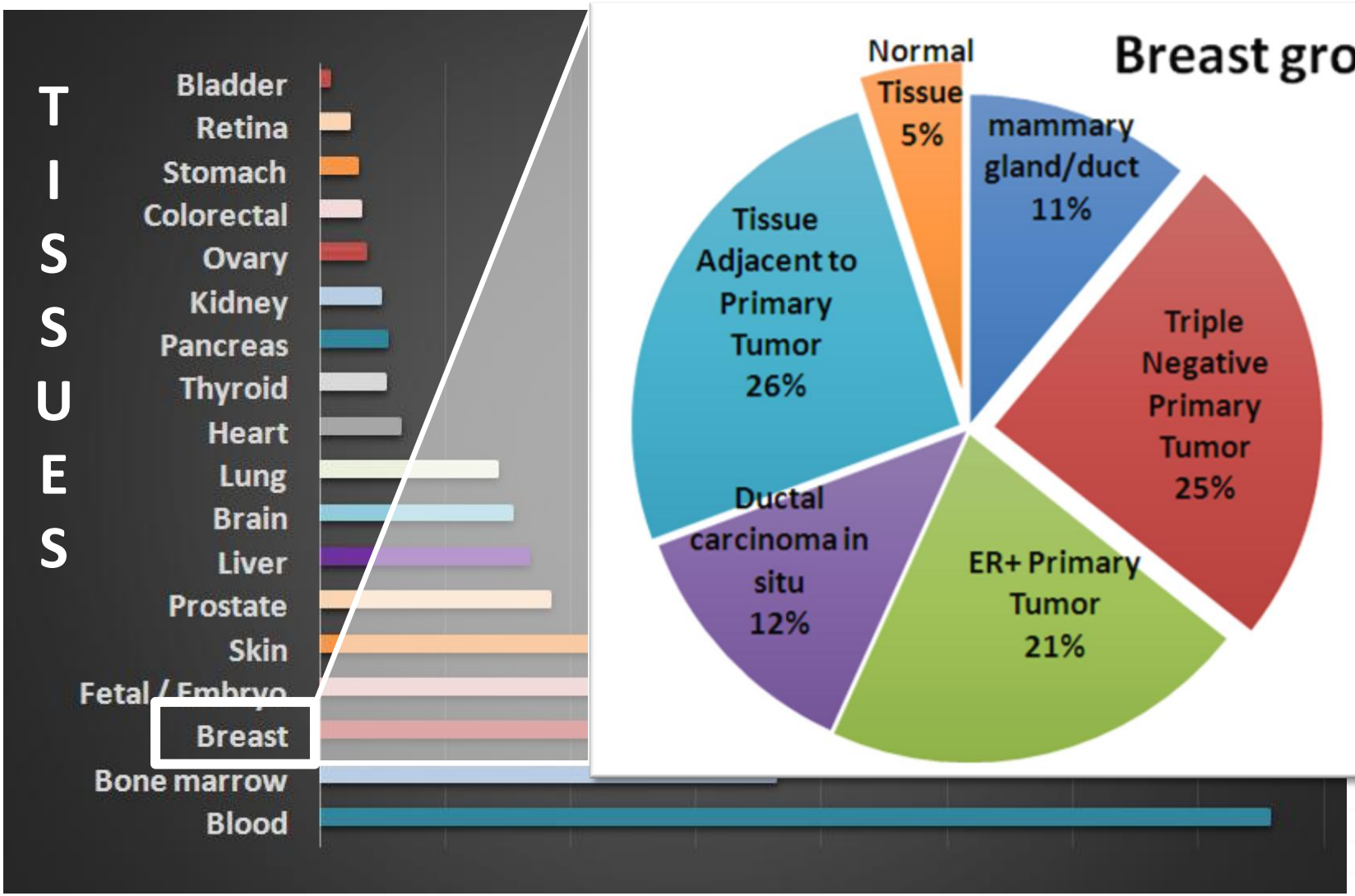
RNA-Seq collection

- **RNA-Seq libraries are collected from public databases (NCBI, EBI)**

- **We selected libraries:**
  - **with comparable data :**
    - same type of data
    - same technology of sequencing
  - from multiple sources (laboratories, countries)

➔ **~27000 Human RNA-Seq selected**
(from multiple organs / pathologies / ethnic groups…)

RNA-Seq diversity



HPC Toolchain

compressed FASTQ

EMBL-EBI

NCBI

1 Downloading

Trimmomatic

Picard

build passing

2 Quality & Filters

Cleaning

3 Mapping

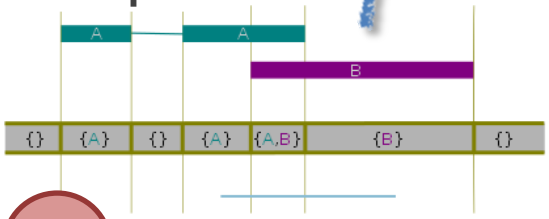
TopHat

A spliced read mapper for RNA-Seq

BOW TIE

Organisation & Aligment

4 Counting



HTSeq

5 Matrix building

	Lib01	Lib02	Lib03	Lib04	Lib05
Gene001	0	5	76	10	0
Gene002	23	0	2	6	9
Gene003	2	95	2	0	2
Gene004	11	4	0	33	1
Gene005	0	5	5	0	53

MaRS

Why HPC?

1 RNA-Seq

(average)

1. **Download** — **6 Go** & **0.5 hour**
2. **Cleaning**
3. **Mapping**
4. **Counting**

4 hours on 12 CPU = 48 hours/core

+ 5. **Matrix Building**

X 27000

MaRS

120 To & **416 Days**

1.2 million hours/core

Many Challenges

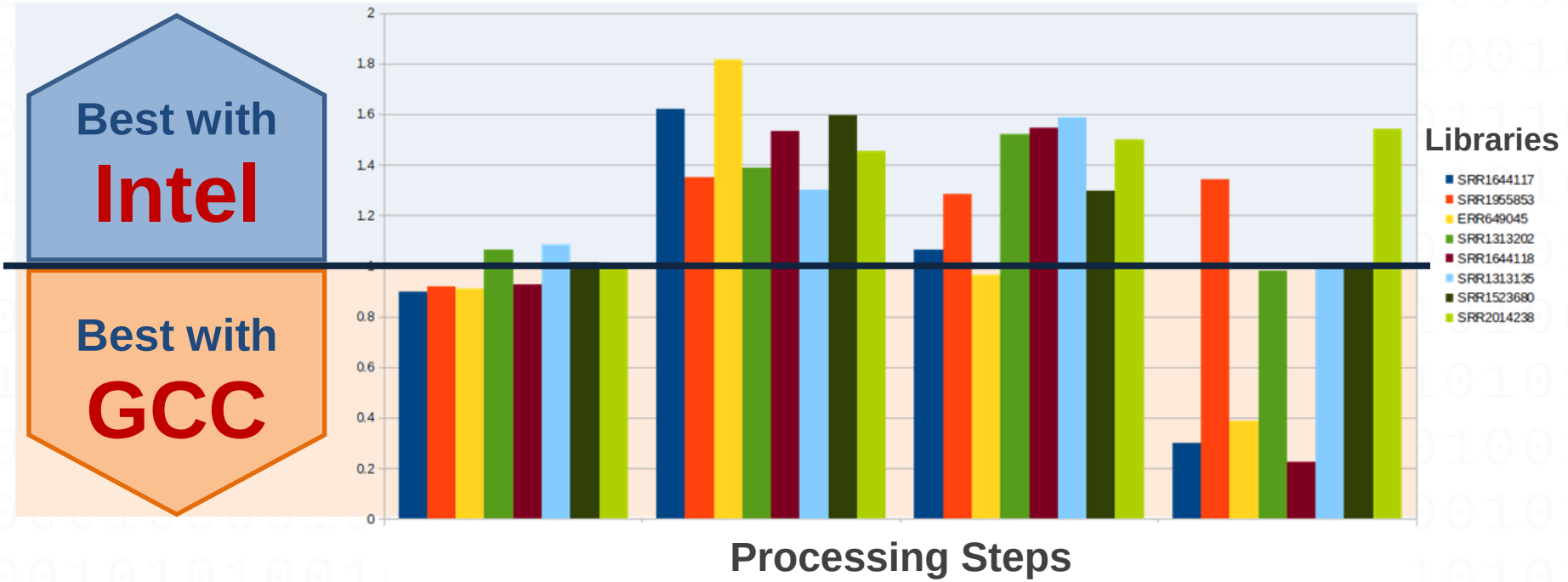
- **Very Large amount of data:**
  - **Downloading**
  - **Processing**

➔ **HPC on cluster (Occigen)**

- **Software's Installation, Settings on the cluster**
- **Optimization**
  - compiler choice
  - packages choice
- **Jobs duration & Limitations management**
  - Adaptation of software's parameters for all libraries

Tests & Benchmarks

Compiler GCC vs compiler Intel



Performance & Optimization

Vtune analysis highlights the high consumption of compression & decompression steps

trimmomatic

Function (Full)	CPU Time:... Effective T... Idle	Poor	CPU Tim... Effective ... Idle	Poor	Instru... Retired: Total
	99.9%		0s		100.0%
org::usadellab::trimmomatic::trim::IlluminaClippingTrimmer$IlluminaLongClippingSeq::readsSeqCompare(org::usadellab::trimmomatic::fastq::FastqRecord)	24.8%		53.193s		33.5%
deflate_slow	**32.2%**		68.992s		26.6%
memscan	3.2%		6.892s		6.9%
[obdclass]	6.0%		12.769s		4.1%
org::usadellab::trimmomatic::trim::IlluminaClippingTrimmer$IlluminaPrefixPair::palindromeReadsCompare(org::usadellab::trimmomatic::fastq::FastqRecord, org::usadellab::trimmomatic::fastq::FastqRecord)	4.7%		10.093s		3.9%
inflate_fast	3.4%		7.292s		2.9%
[libcfs	2.6%		5.532s		2.4%
org::us	1.3%		2.719s		2.2%
sun::n	0.5%		1.030s		1.3%
sun::n	0.4%		0.813s		1.3%
compr	1.2%		2.558s		1.2%
org::us ...matic::fastq::...	0.9%		1.986s		1.2%
org::us	0.8%		1.665s		1.2%
[lov]	1.4%		3.026s		1.0%
org::...	0.5%		0.973s		0.7%
org::us	0.7%		1.469s		0.6%
[lustre					0.6%
[lvfs]					0.6%
jshort_					0.5%
org::us					0.5%
org::us					0.3%
fill_wi					0.3%
[osc]					0.2%
org::us					0.1%

Picard SortSam

Function (Full)	CP.. Eff... Idle	CP. Eff... Idle	Instr... Retir... Total
	99.9%	0s	100.0%
deflate_slow	25.5%	19.796s	28.2%
func@0x54ea0	12.7%	9.854s	9.2%
htsjdk::samtools::BAMRecordCodec::encode(htsjdk::samtools::SAMRecord)			
inflate_fast			
htsjdk::samtools::SamReader$AssertingIterator::next()			
func@0x5624b			
compress_block			
java::util::TimSort::mergeHi(int, int, int, int)			

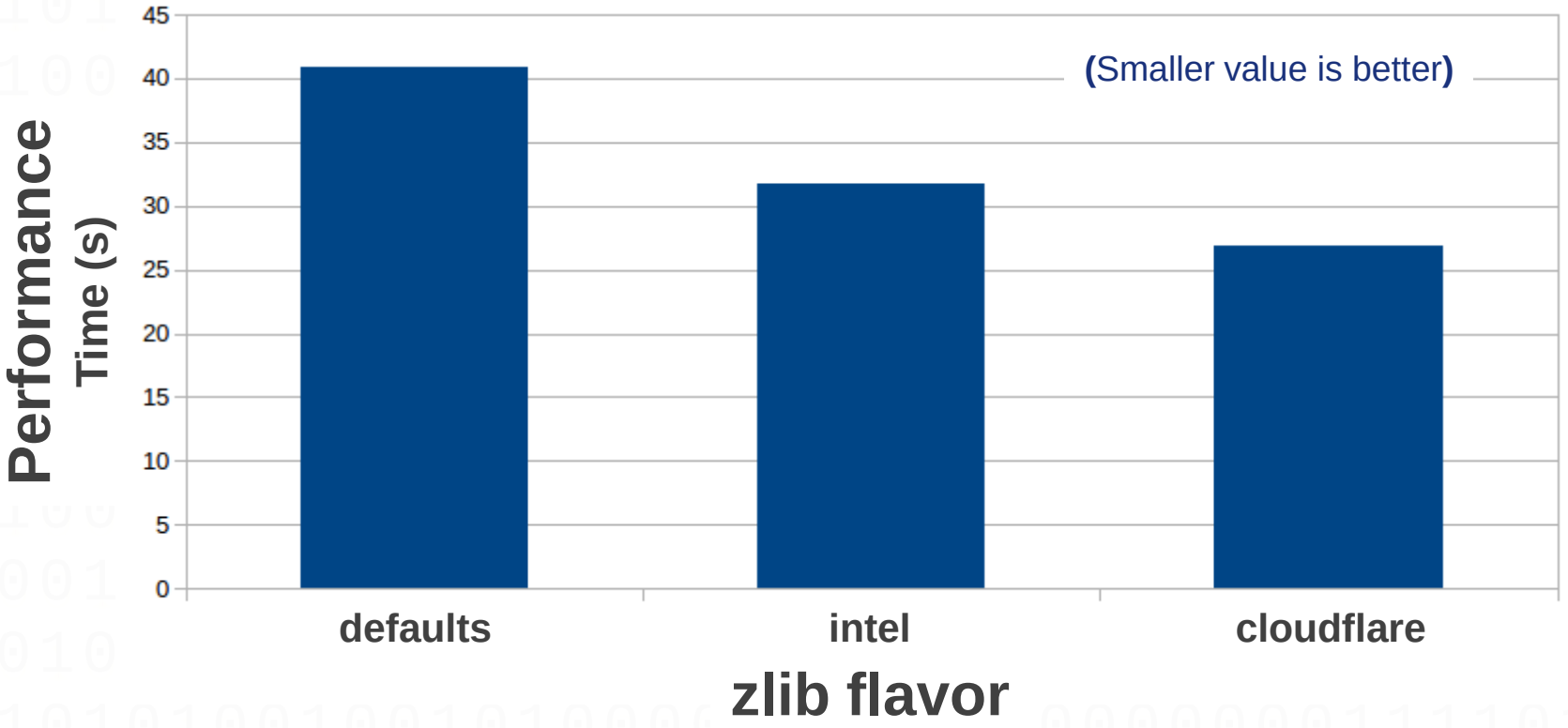
Samtools view

Function (Full)	CP.. Eff... Idle	CP. Eff... Idle	Instr... Retir... Total
	100.0%	0s	100.0%
longest_match	66.8%	66.535s	56.6%
deflate_slow	14.2%	14.155s	14.7%
sam_parse1	3.4%	3.415s	8.6%
compress_block	4.7%	4.721s	6.0%

Performance & Optimization

zlib decompression performance

~750 MB compressed lib



Conclusion & Perspectives

- **NOW : ➔ 95% of the Human MaRS matrix is built**
- **NEXT :**
  - **➔ MaRS *Exploration***
    Matrix hosting & Queries allowing:
    - Web interface / Database
    - integrated analysis tools
  - **➔ RNA-Seq Data production increases continuously**
    - Integration of new profiles in MaRS
  - **➔ Similar work on other species: Mouse**
- **BIOMARKERS discovery**

Partners

ACOBIOM

- French SME
- Discovery & validation of new biomarkers

INES
Centre Informatique National de l'Enseignement Supérieur

- Multiple missions:
  HPC, digital archiving, hosting
- *Occigen* cluster: 2.1 Pflops

(Top 100 of the most powerful computers in the world)

PRACE SHAPE
SME HPC Adoption Programme in Europe

PRACE - SHAPE

SME HPC Adoption Programme in Europe

Acknowledgments

THANK YOU for your attention

THANKS to all collaborators:

- Fabien PIERRAT
- Laurent MANCHON
- Roman BRUNO
- David PIQUEMAL
- Bertrand CIROU
- Victor CAMEO PONZ
- Francis DAUMAS*