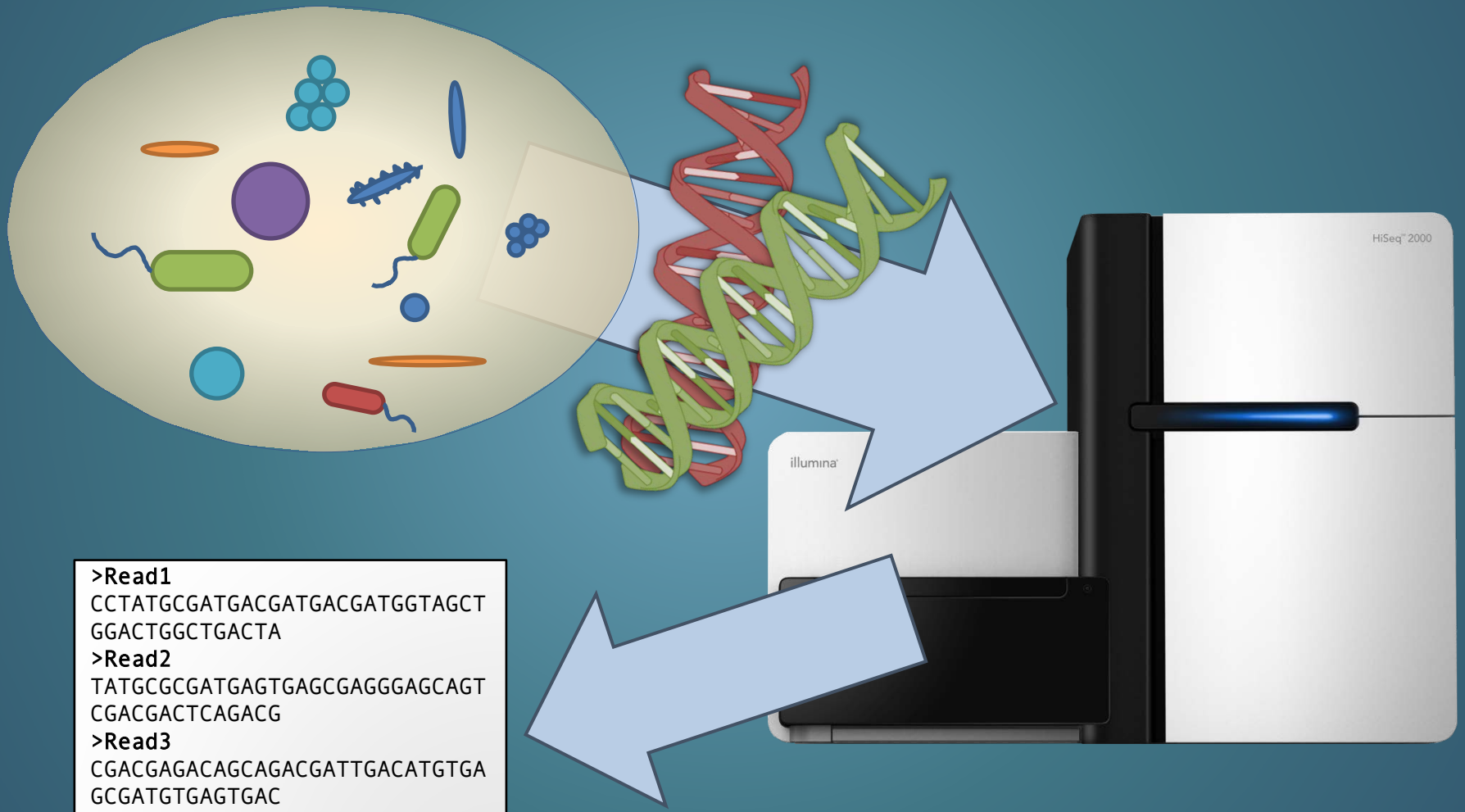
Framework for gene quantification in large-scale metagenomic data

Fredrik Boulund
Chalmers University of Technology
27 February 2014

Presentation outline

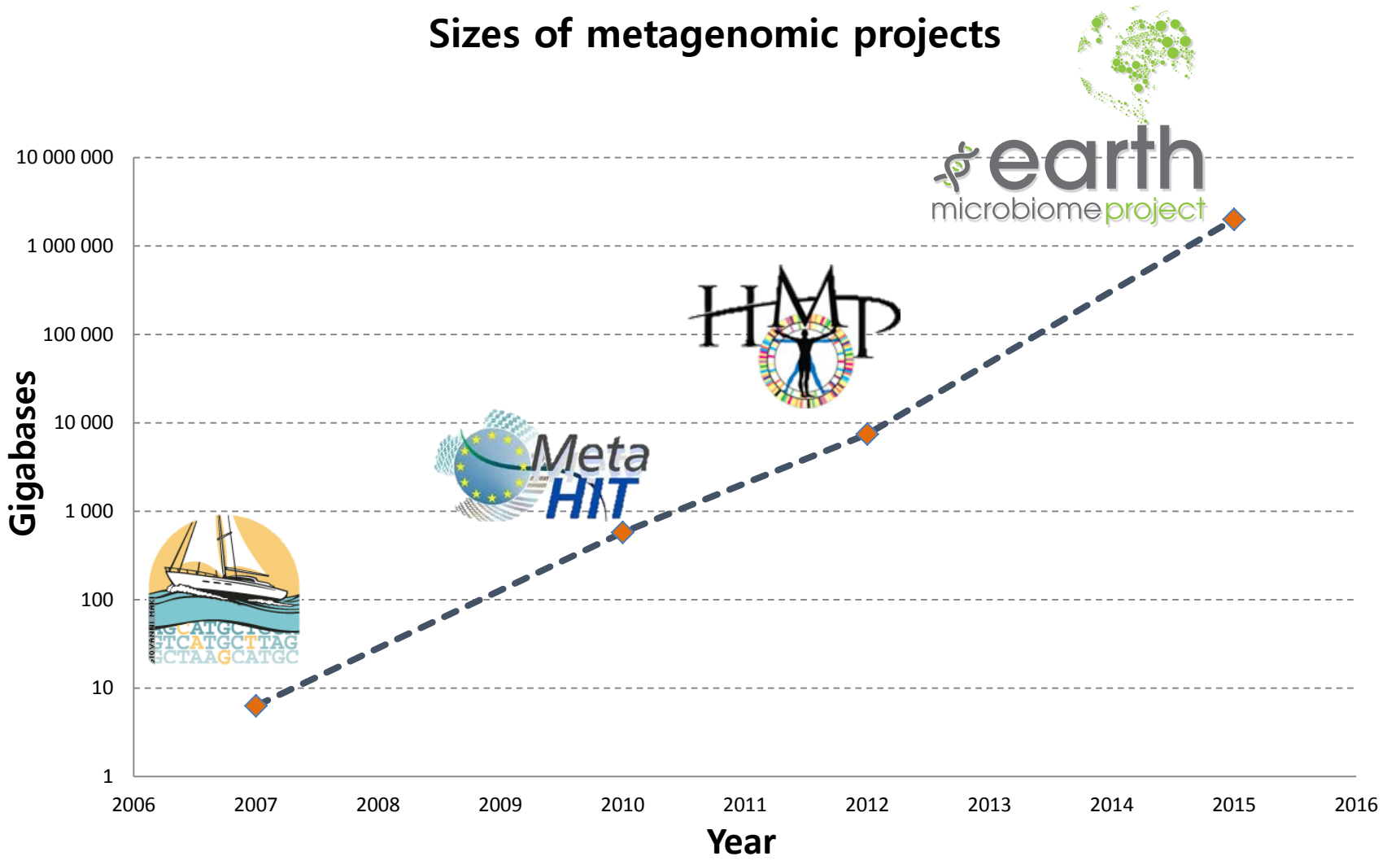
- Metagenomics
- Gene quantification
- Method
- Performance evaluation

Metagenomics




```
>Read1
CCTATGCGATGACGATGACGATGGTAGCT
GGACTGGCTGACTA
>Read2
TATGCGCGATGAGTGAGCGAGGGAGCAGT
CGACGACTCAGACG
>Read3
CGACGAGACAGCAGACGATTGACATGTGA
GCGATGTGAGTGAC
.........
```

Metagenomic data revolution



GOS: http://www.jcvi.org/cms/research/projects/gos/overview/, **Meta-HIT**: http://www.metahit.eu/, **HMP**: http://www.hmpdacc.org/, **EMP**: http://www.earthmicrobiome.org/. Accessed 2013-10-14.

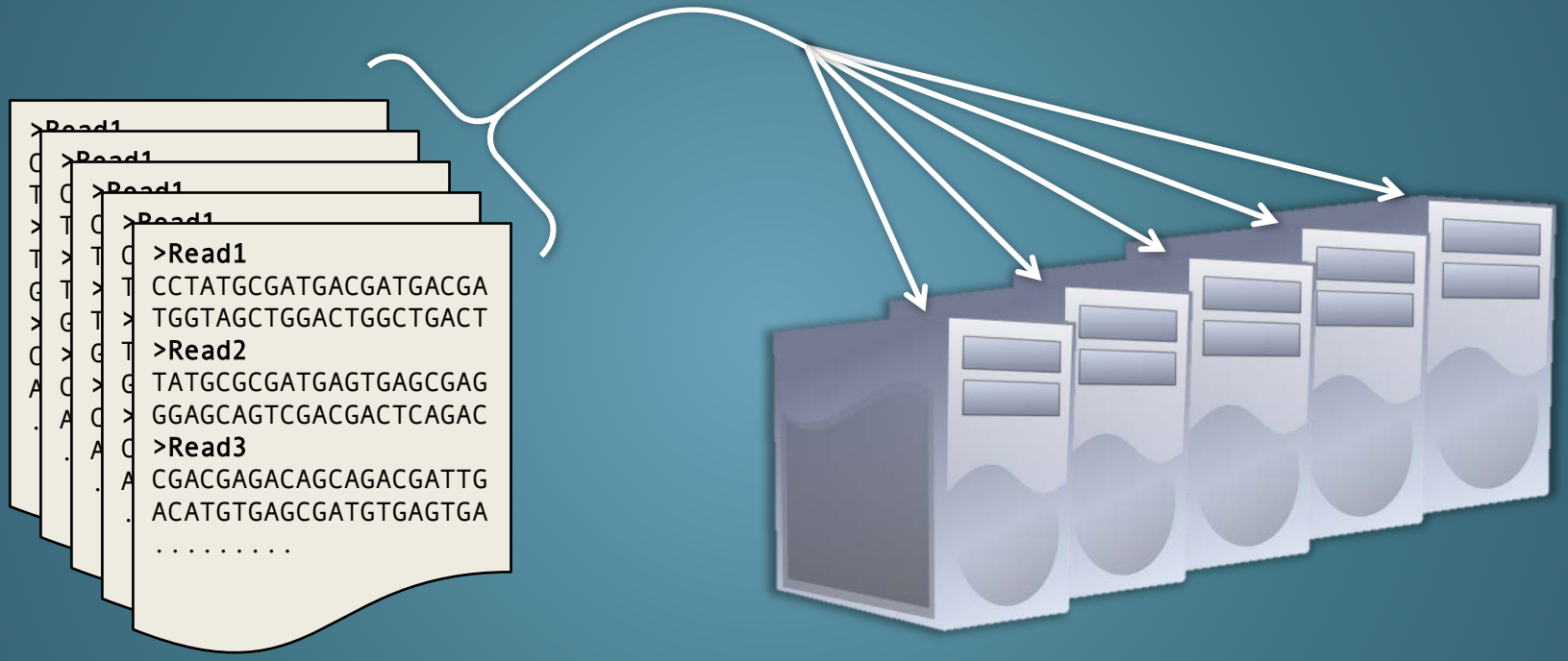
Gene quantification

- Statistical comparisons
- Measure gene abundance
- Mapping

Objectives

- Quantify the abundance of specific genes in large metagenomes (Terabases)
- Develop software for parallelized gene quantification
- Evaluate performance on High-Performance Computing (HPC) clusters

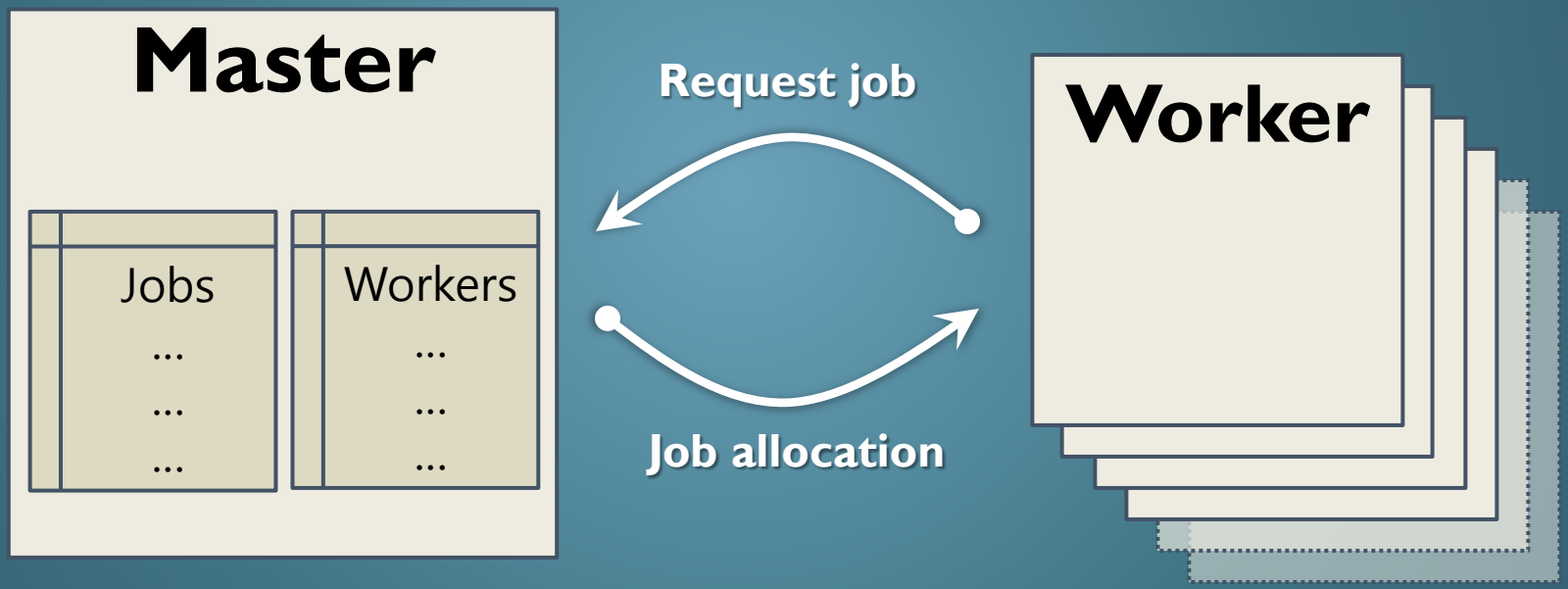
Distributed read mapping



Many samples

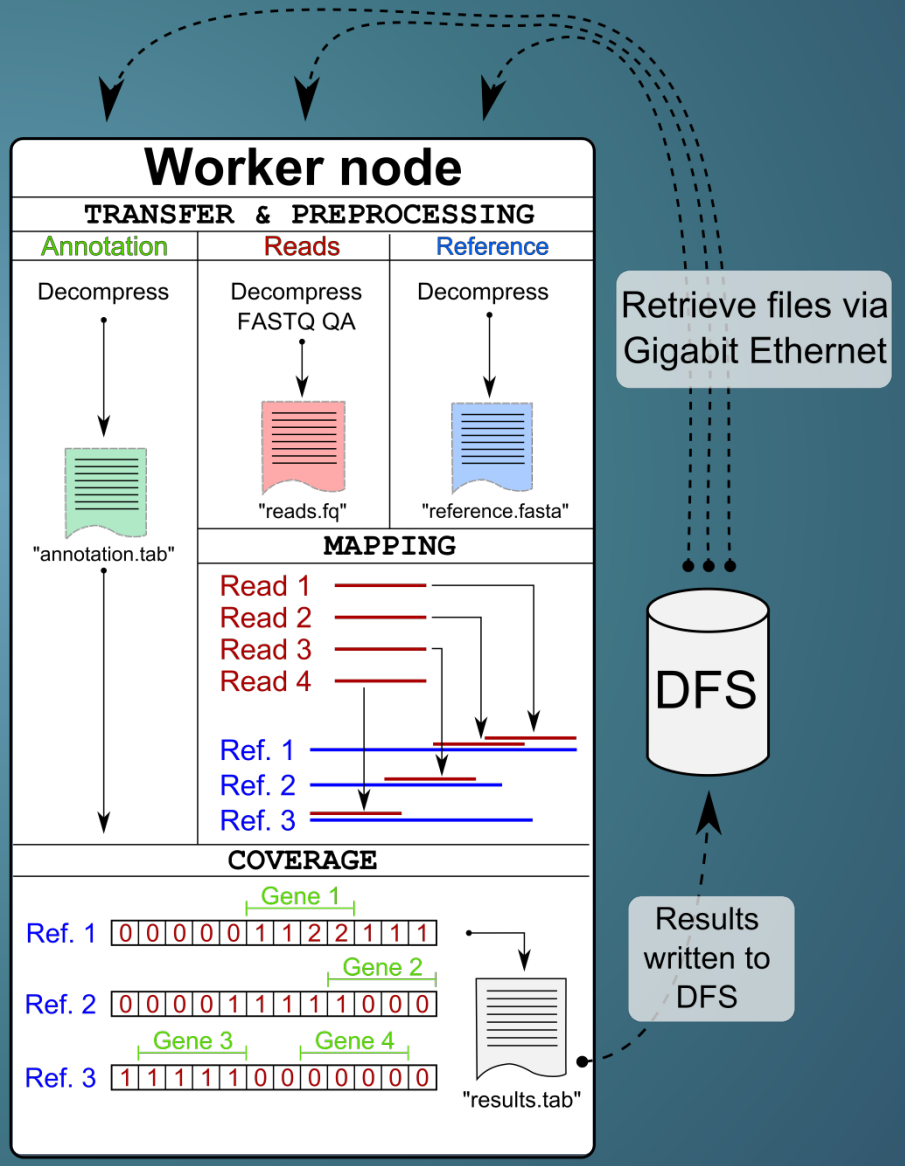
HPC cluster

Master-Worker



Worker node

- Preprocessing:
  - Annotations
  - Reads
  - References
- Mapping
  - GEM, PBLAT, Bowtie2, BLAST, RazerS3
- Coverage
  - Custom implementation



Implementation

- Python 2.7

- ZeroMQ

- Chalmers Centre for Computational Science and Engineering
  - Glenn

Case studies

- Metagenomic data set from Meta-HIT study (Qin et al. 2010)
- Illumina sequencing data
- 400 Gigabyte compressed (2.2 Terabyte uncompressed)
- 1.2 billion fragments
- 512 samples

Case 1

- Mapper: pblat
- References: contigs
- Ref. size: ~160 MiB
- Sequences: 50 k - 100 k
  6.6 million total

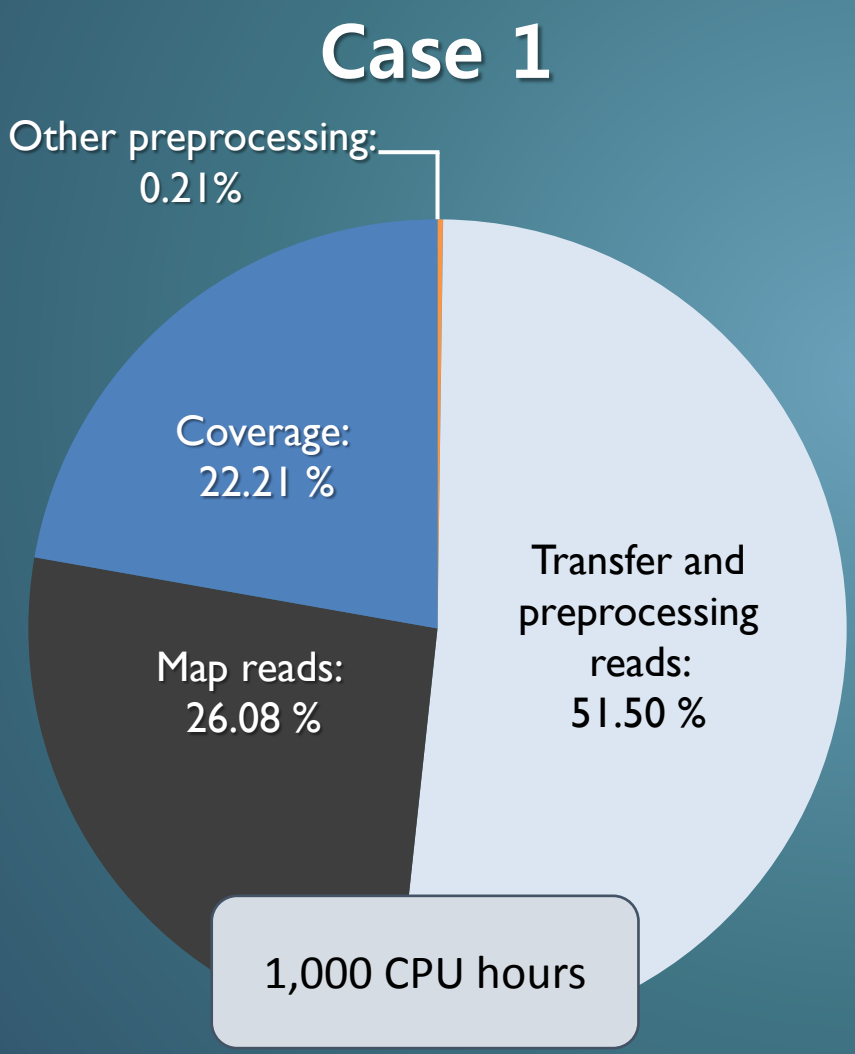
- Several small references
- High similarity expected

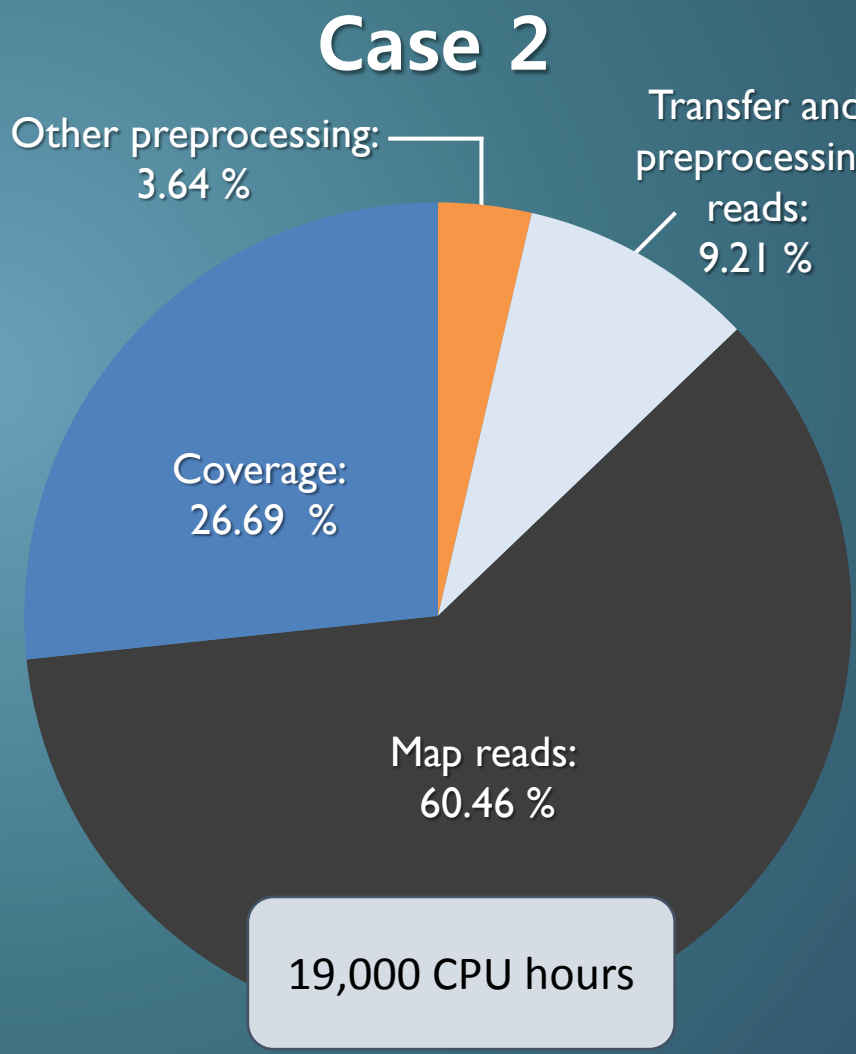
Case 2

- Mapper: GEM
- References: large DB
- Ref. size: 7.5 GiB
- Sequences: 11.6 million

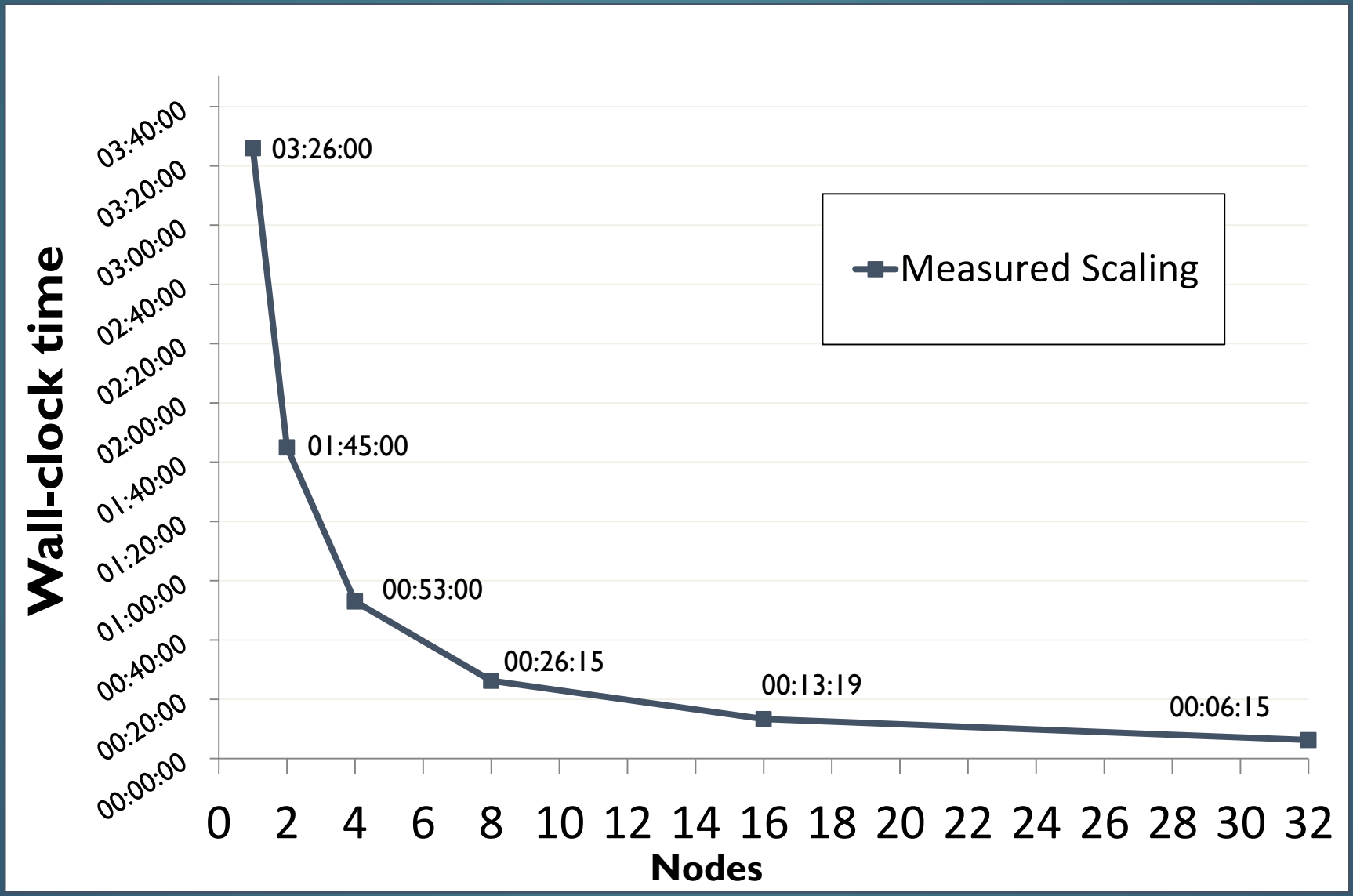
- One common reference
- Filter human sequences
- Collaboration with Fredrik Karlsson

Subprocess time consumption

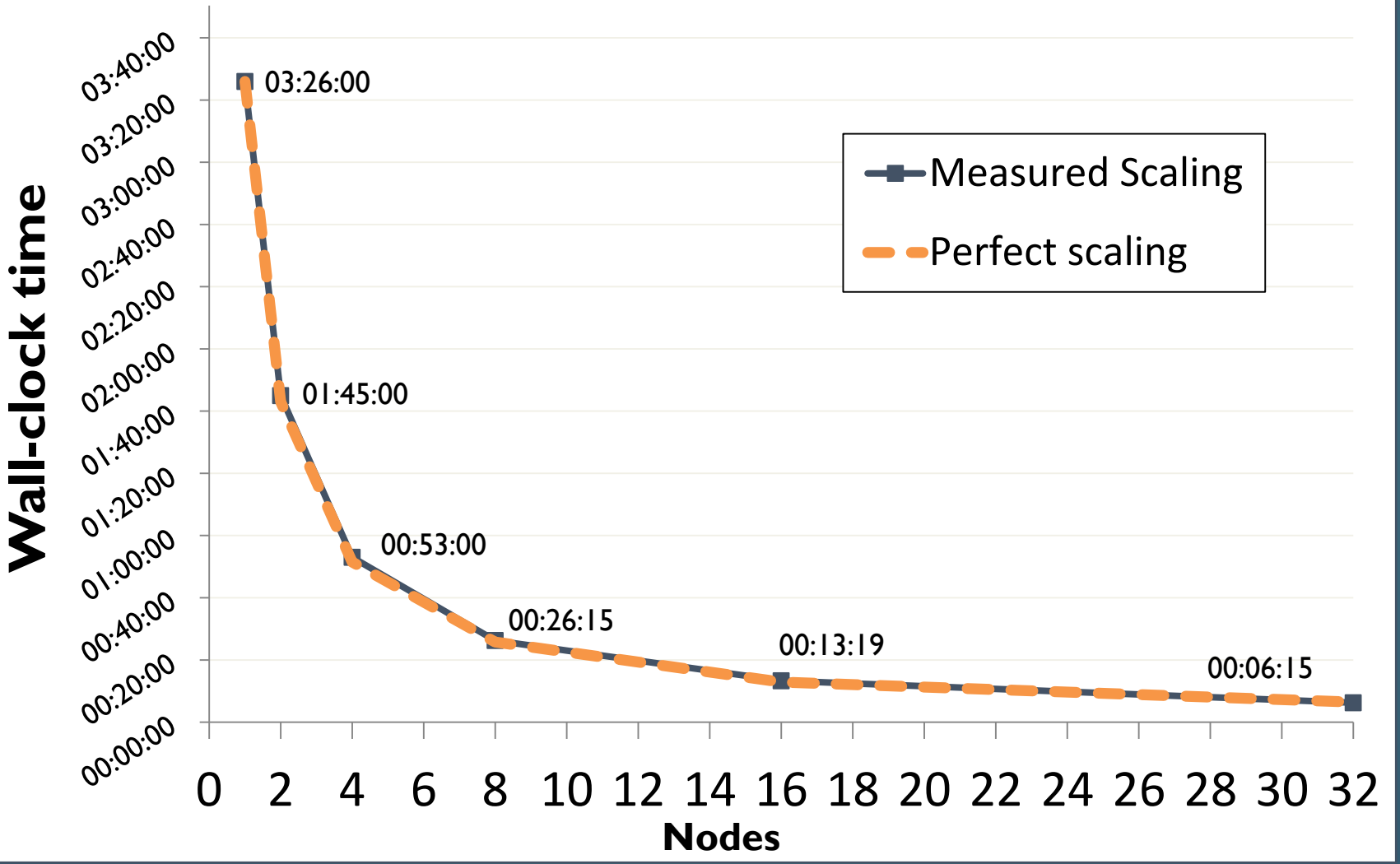




Parallel scaling



Parallel scaling



Conclusions

- Method for gene quantification in terabase metagenomes
- Excellent parallel scaling
- Flexible; applicable to wide range of studies

Acknowledgements

- Collaborators:
  - Chalmers Statistics: Anders Sjögren, Erik Kristiansson
  - Chalmers SysBio: Fredrik Karlsson, Intawat Nookaew
  - Gothenburg University, Sahlgrenska Academy; Joakim Larsson's group: Carl-Fredrik Flach, Anna Johnning, Johan Bengtsson-Palme
- Funding:

LIFE SCIENCE
A CHALMERS AREA OF ADVANCE

CHALMERS
e-Science Centre

ADLERBERTSKA STIFTELSERNA

Vetenskapsrådet