Texas A&M

Supercomputing Facility

Boosting Productivity with Advanced User Services

Raffaele Montuoro

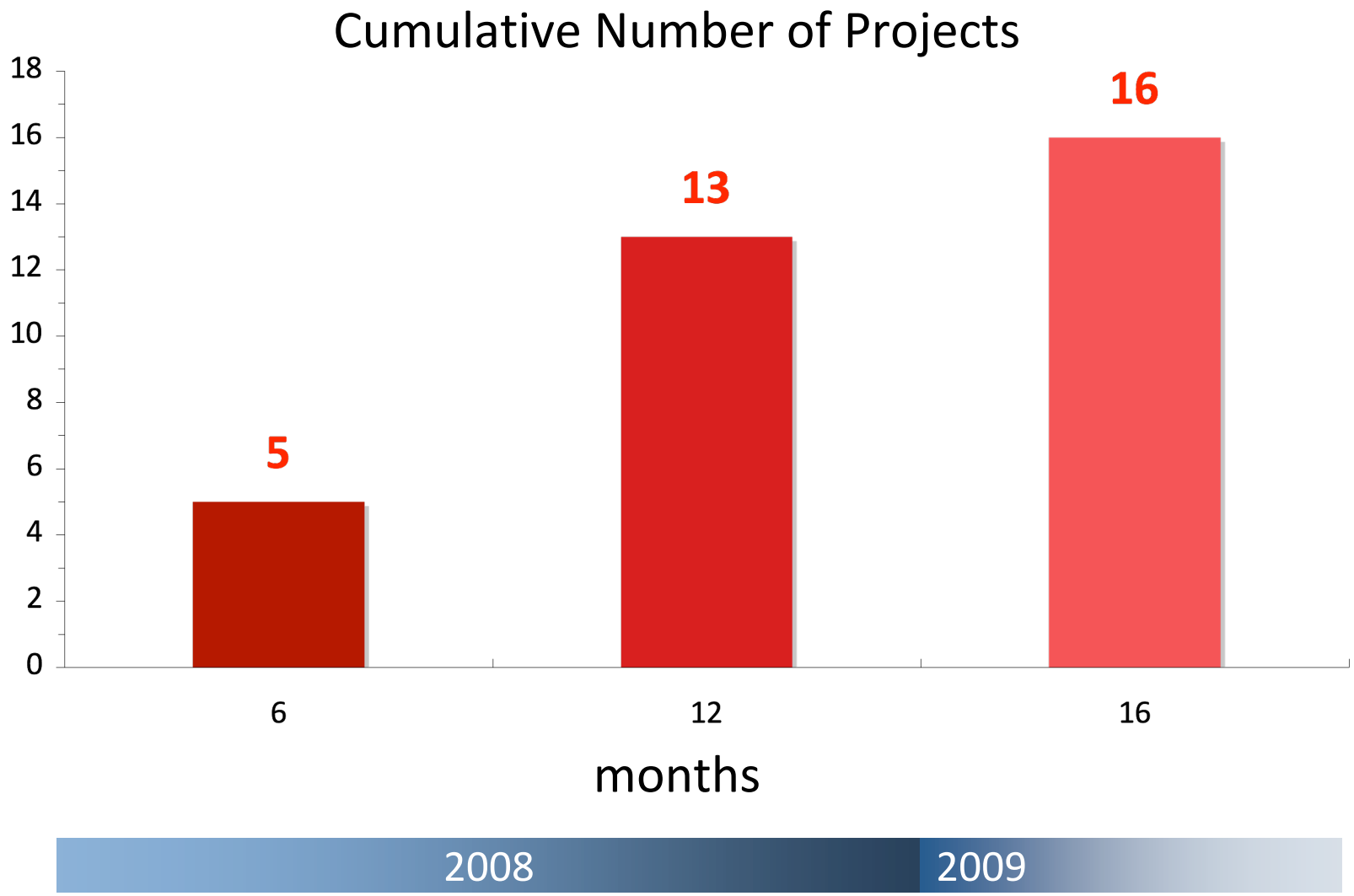
Advanced User Services: Mission

- Enhance and support computational sciences within Texas A&M University

Advanced User Services: Details

- Performance Analysis of computer codes
- Code Optimization
- Code Parallelization: OpenMP & MPI
- Consulting for code development on SC systems
- Code Porting on the SC systems
- Tune up of common scientific applications
- Benchmarking
- Design and configuration of small computer clusters

Advanced User Services: Young, but Tall



Advanced User Services: A Closer Look

- SODA: A Simple Ocean Data Assimilation Model

  Dr. Benjamin Giese, Dept. of Oceanography

- MST: Material Simulation Tool

  Dr. Tahir Cagin, Dept. of Chemical Engineering

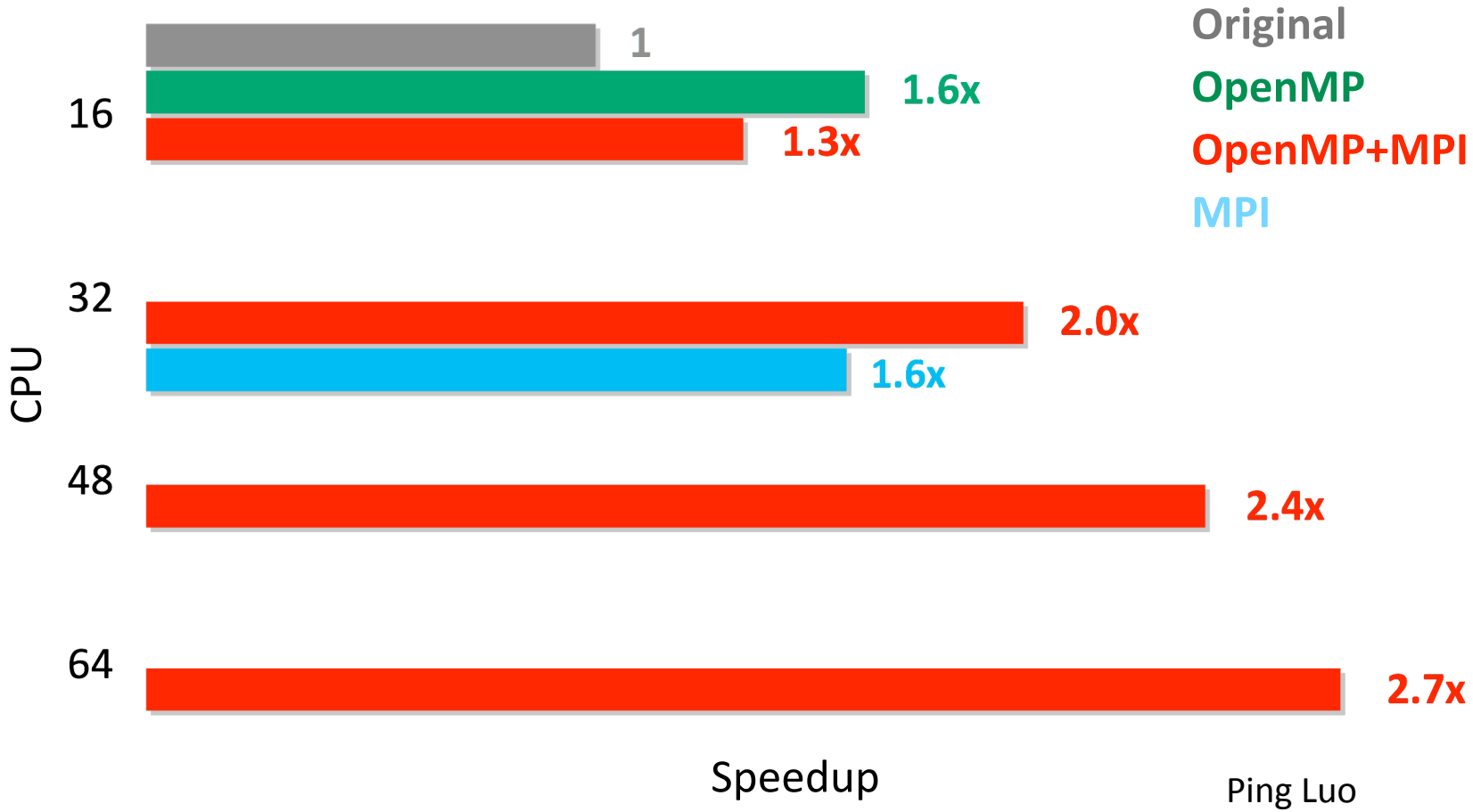
- Illumina Genome Analysis Pipeline

  Dr. James Sacchettini, Dept. of Biochemistry and Biophysics

SODA: Simple Ocean Data Assimilation

B. Giese

Original Code: FORTRAN/OpenMP Case: 2004-01-20



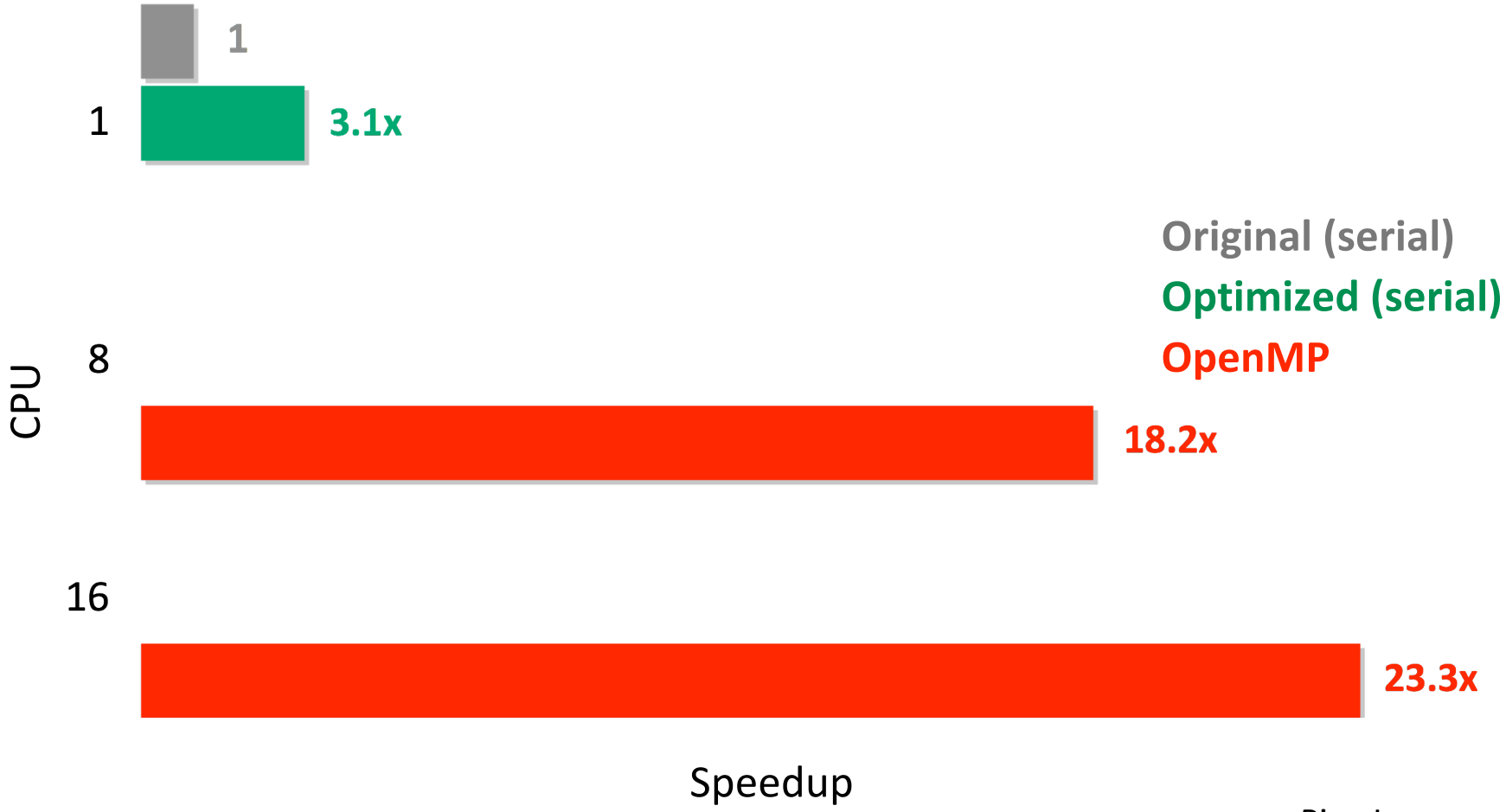
Ping Luo
Xiandong Meng
SC staff

Material Simulation Tool

T. Cagin

Original Code: C++, serial

Case: 101010



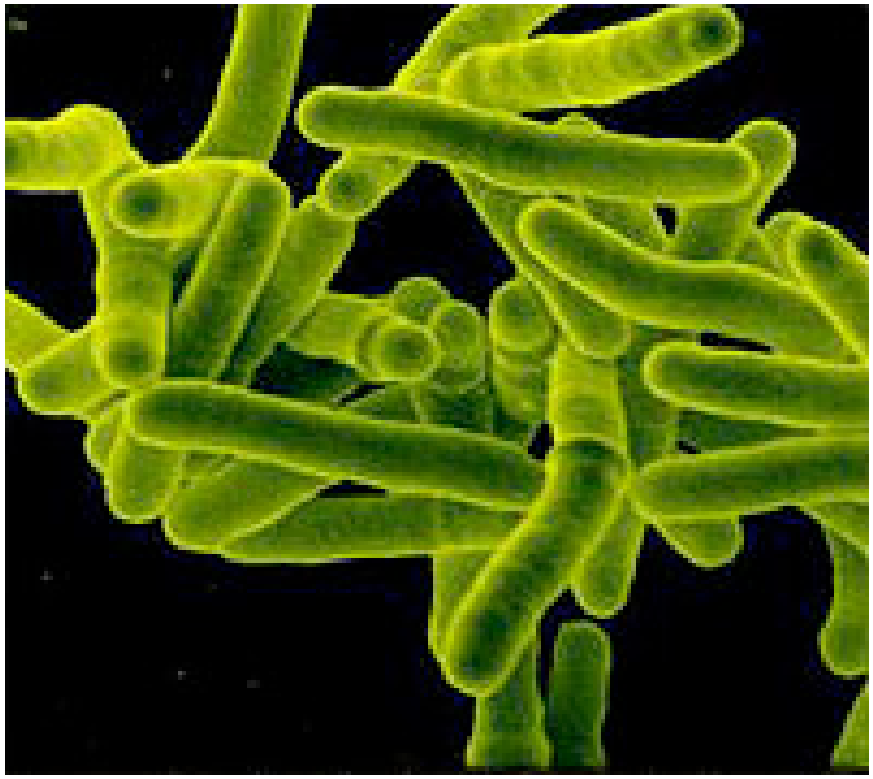
Ping Luo

SC staff

Parallel Genome Analysis Pipeline

J. Sacchettini
T. Ioerger

Bio- and Chemi- Informatics:
Studies on *M. tuberculosis* Drug Action & Resistance



Whole genome sequencing can be used to define the mechanism of drug action and resistance.

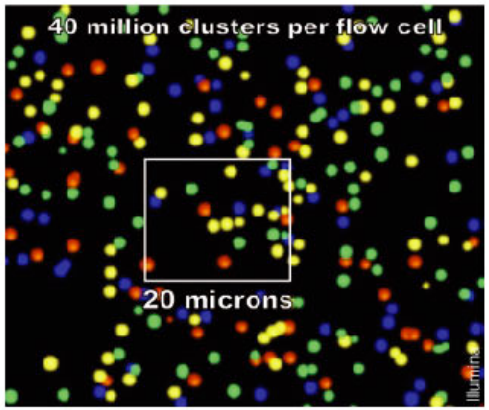
Parallel Genome Analysis Pipeline

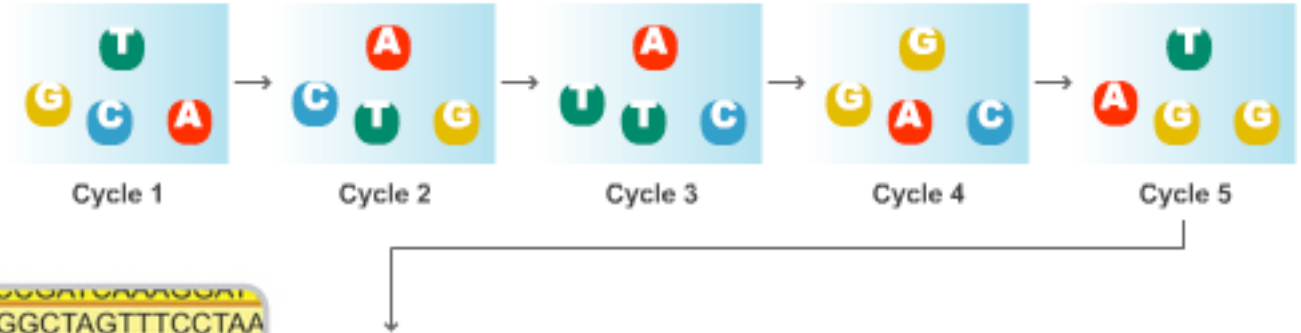
J. Sacchettini
T. Ioerger

How Solexa sequencing works

Single-ended sequencing:

1) Fragment gDNA select 200-300 base fragments
2) Spread and attach fragments to a lane on a chip, amplify
3) Press go





4) Just align your short reads, allowing gaps/mismatches, against a reference genome, and look for snps and indels

Parallel Genome Analysis Pipeline

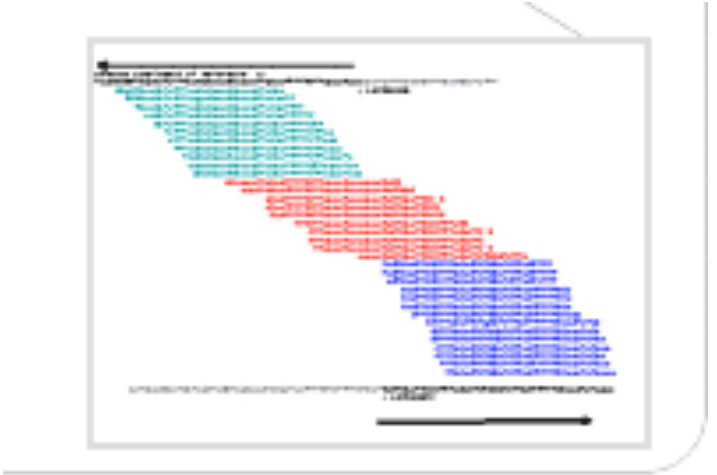
J. Sacchettini
T. Ioerger

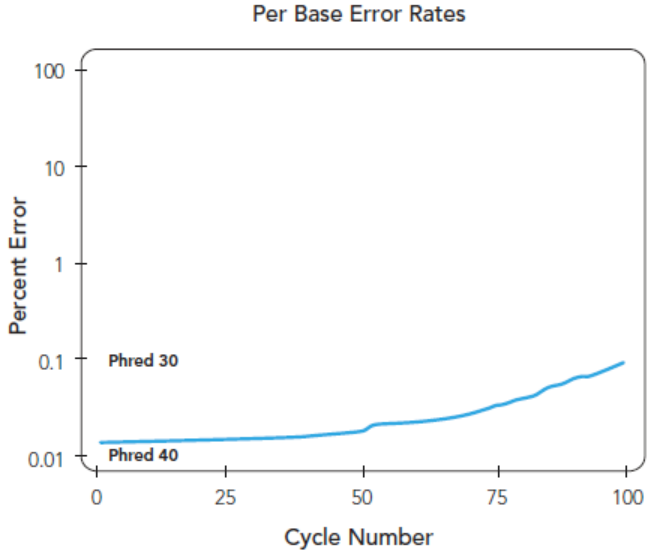
What can you do with all that sequence?

Sequence an entire genome of Mtb
—4M bases in the Mtb genome

Therefore with 20Mx36 base reads you get 100% coverage with 200-fold redundancy- (depth of coverage) per genome; 7 genomes per chip

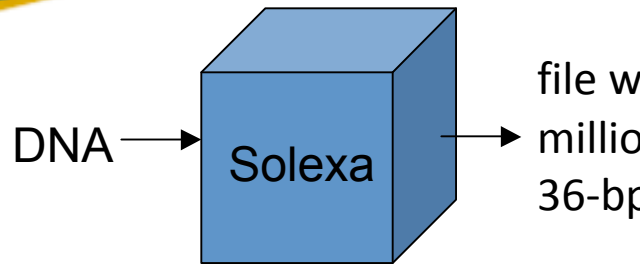
Or you can add a tag to each of 4 genomes and run them on a single lane -50 fold depth of cover per genome- 28 genomes per chip





Parallel Genome Analysis Pipeline

J. Sacchettini
T. Ioerger



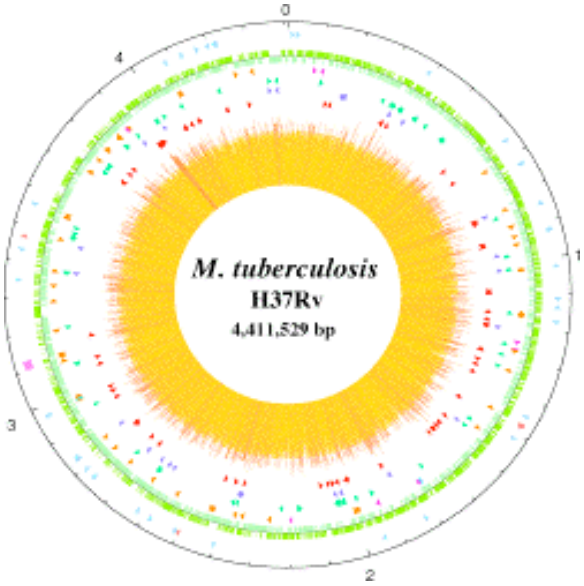
file with millions of 36-bp reads

align reads against reference genome

identify *putative SNPs* where majority base differs from expected base

build *local contigs* in surrounding ~200 bp and align against reference genome to identify true SNPs vs. indels

edit genome sequence and realign reads



Ioerger Sequencing Protocol

Parallel Genome Analysis Pipeline

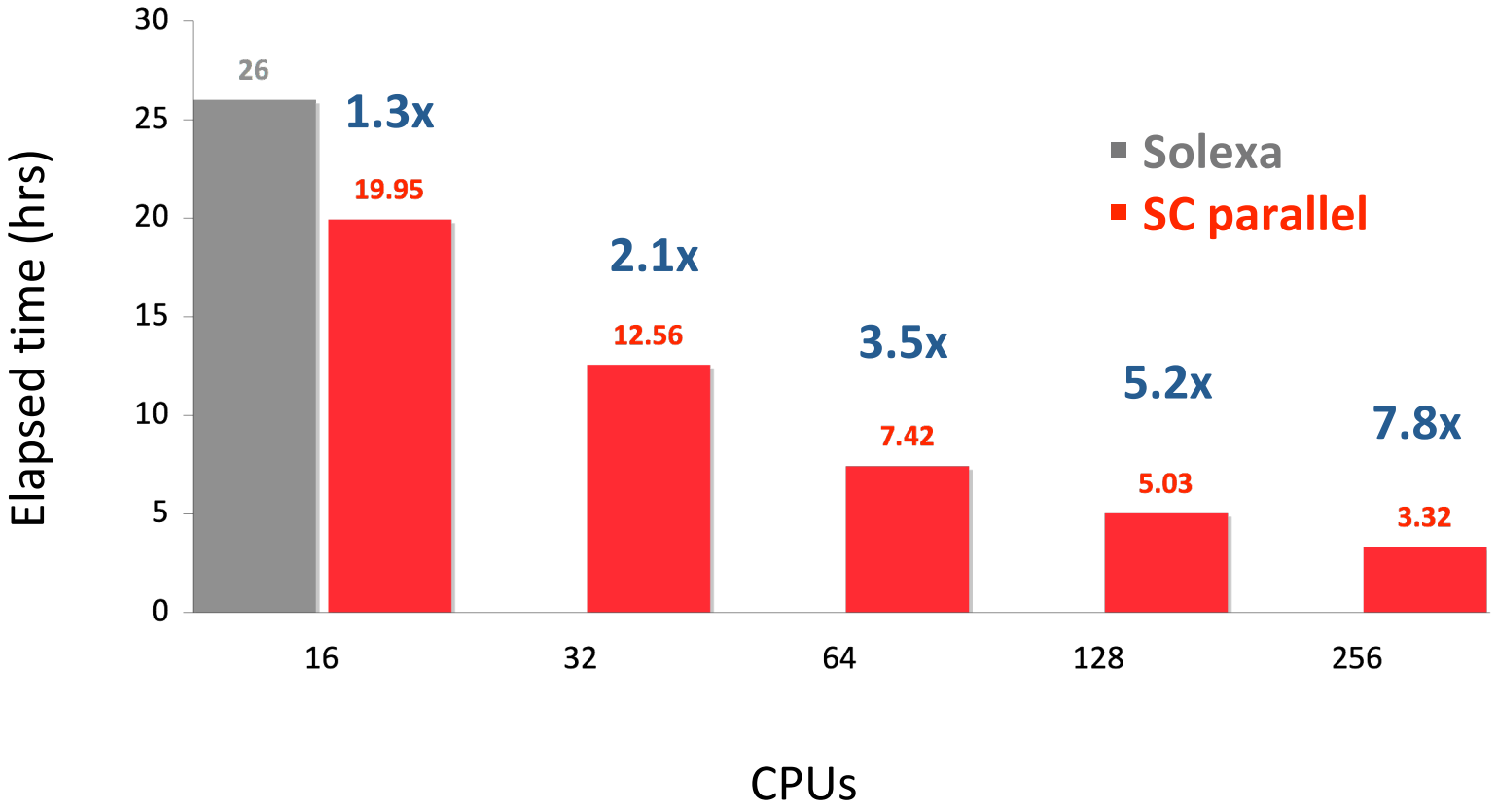
J. Sacchettini
T. Ioerger

- Solexa's Genome Analysis Pipeline is a customizable analysis engine capable of taking the raw image data generated by the Genome Analyzer and producing intensity scores, base calls, and quality metrics, and quality scored alignments
- Based on Makefile
- Scales up to 8 shared-memory tasks (`gmake -j 8`)
- Typical problem size: 8 lanes x 36 cycles x 4 bases x 100 images/base/cycle = **115,200** images to be processed

Parallel Genome Analysis Pipeline

J. Sacchettini
T. Ioerger

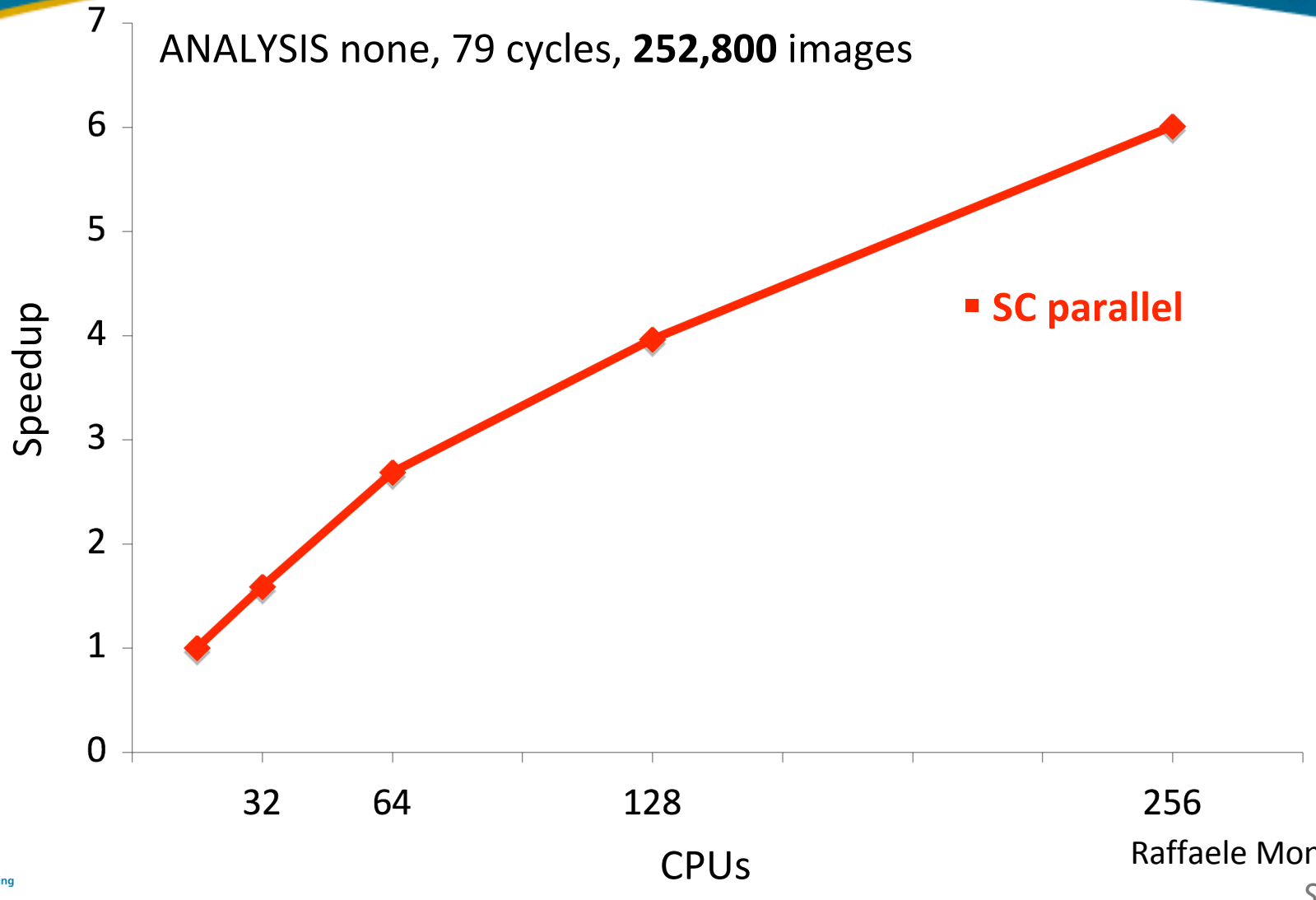
ANALYSIS none, 79 cycles, **252,800** images



Raffaele Montuoro
SC staff

Parallel Genome Analysis Pipeline

J. Sacchettini
T. Ioerger



Raffaele Montuoro
SC staff

Advanced User Services

Q: How to apply?

A: E-mail the Supercomputing Help Desk:

help@sc.tamu.edu