Tazro Ohta
DBCLS, ROIS
twitter.com/inutano
github.com/inutano
speakerdeck.com/inutano

A system to validate published data and studies

Tazro Ohta, Ryota Yamanaka, Osamu Ogasawara, Yoshinobu Masatani, Shigetoshi Yokoyama, Kento Aida

Reproduction of results from a published study or Meta-analysis using public data is painful

What happened during the data analysis

3. Data Processing of Heliscope CAGE data

Sequenced Heliscope reads have a high sequencing error rate (~5%), vary in length and lack an estimation of base qualities. Combined these factors make the data processing challenging. As an initial step we removed reads corresponding to ribosomal RNA. We accomplish this by directly aligning each read against the whole human (mouse) ribosomal DNA complete repeating unit and discarding all reads with an edit distance smaller or equal to two. For this purpose we implemented Myers' bit parallel dynamic programming algorithm[13] in the program rRNAdust (author: T. Lassmann). For computational efficiency we further parallelized this algorithm using both SIMD instructions and threads. All remaining CAGE reads were mapped to the genome (hg19 and mm9) using Delve, a probabilistic mapper[14]. In brief, Delve uses a pair hidden Markov model to iteratively map reads to the genome and estimate position dependent error probabilities. After all error probabilities are estimated, individual reads are placed to a single position on the genome where the alignment has the highest probability to be true according to the pHMM model. Phred scaled mapping qualities[15], reflecting the likelihood of the alignment at a given genome position, are also reported. Reads mapping with a quality of less than 20 (<99% chance of true) were discarded. Furthermore, we discarded all reads that map to the genome with a sequence identity of less than 85%.

What exactly happened during the data analysis

```
# Get input sequence
wget ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/DRA000/DRA000991/DRX008610/DRR009482.fastq.bz
bunzip2 DRR009482.fastq.bz2

head -300000 DRR009482.fastq > input.fastq

input=input.fastq
output=output.sam
genome=hg19.fa
ribosomal=human_rDNA_U13369.1.fa

rRNAdust=rRNAfilter/rRNAdust
delve=delve/src/delve
samtools=samtools-1.2/samtools

# Cleaning
$rRNAdust -s $ribosomal $input -e 2 > tmp1

# Alignment & Mapping (FASTQ to SAM)
$delve index $genome
$delve seed  tmp1 $genome -o tmp2 -t 8 -l 12 -s 8
$delve align tmp2 $genome -u 1 -o $output -t 8

# SAM to BED
$samtools view -bS $output > $output\.bam
```

Missing description:

- *OS/Library dependencies*
- *Versions of tools, references, databases*
- *Arguments of tool executions*
- *Tiny scripts or one liner for data format conversion*

Required: executable materials and methods

Research Activity Time Course

Details of Project Design

Sampling

Wet Experiments

Primary Data

Data Processing & Analysis

Publication Text, Figs

DDBJ DNA Data Bank of Japan

BioProject

DDBJ DNA Data Bank of Japan

BioSample

DDBJ DNA Data Bank of Japan

Genbank, DRA

PubMed Central

figshare credit for all your research

zenodo

(GIGA)n SCIENCE

Research Activity Time Course

Details of Project Design

Sampling

Wet Experiments

Primary Data

Data Processing & Analysis

Publication Text, Figs

DDBJ DNA Data Bank of Japan

BioProject

DDBJ DNA Data Bank of Japan

BioSample

Ge

Galaxy PROJECT

(GIGA) SCIENCE

Research Activity Time Course

Details of Project Design

Sampling

Wet Experiments

Primary Data

Data Processing & Analysis

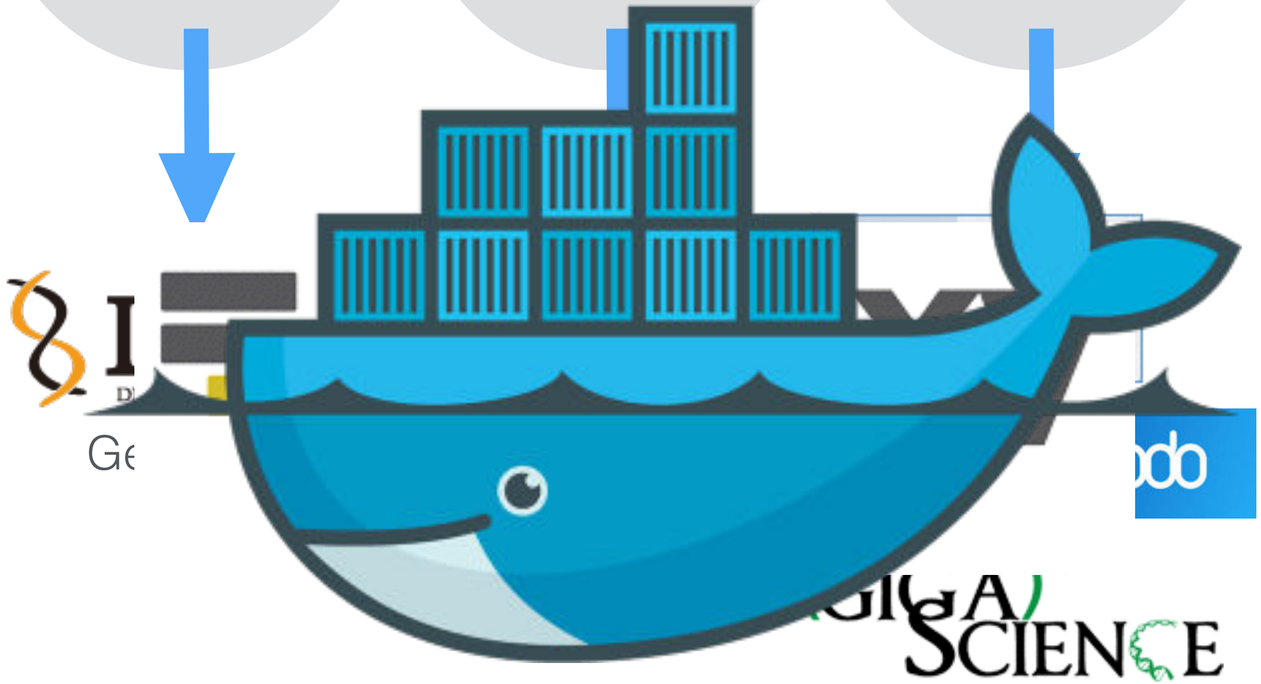
Publication Text, Figs

DDBJ DNA Data Bank of Japan

BioProject

DDBJ DNA Data Bank of Japan

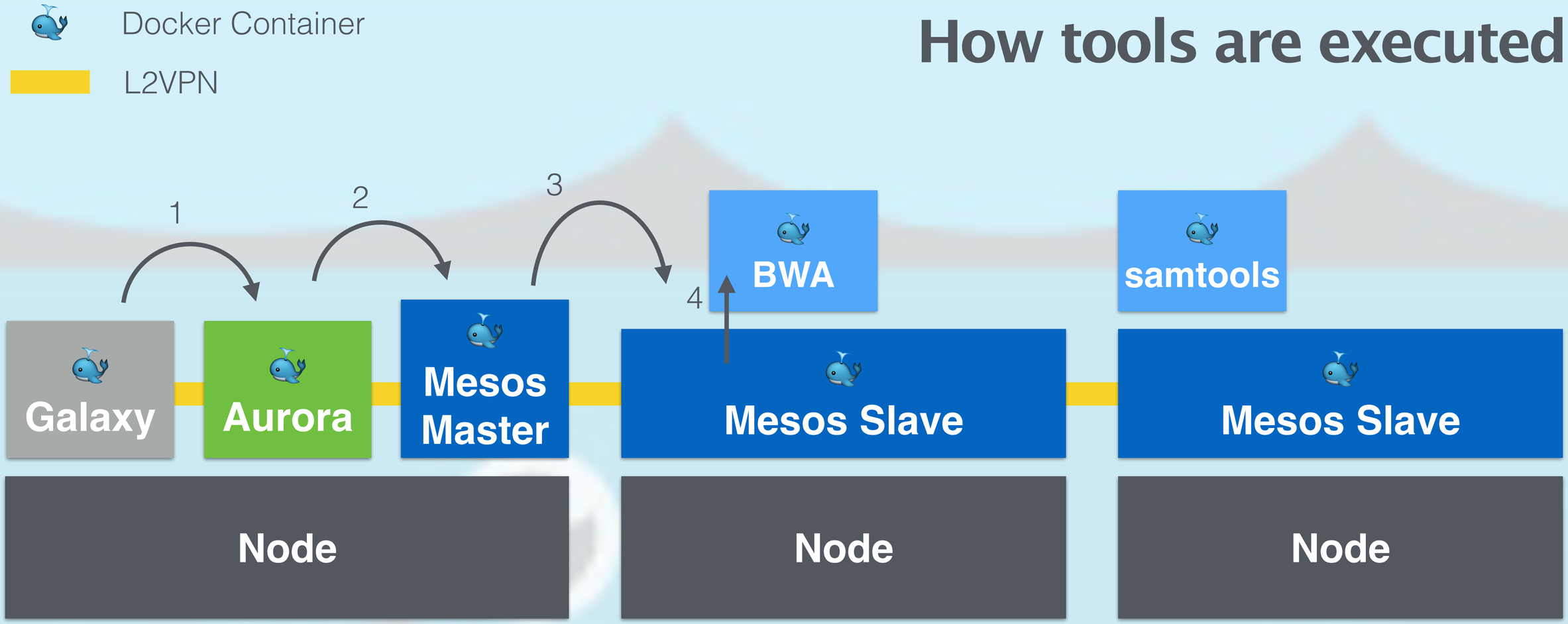
BioSample



SCIENCE


proof-of-concept development:
portable, scalable infrastructure for galaxy container

How tools are executed



How data fetched & stored