UNIVERSITY OF Nebraska Lincoln

SBBI@unl

A Large-scale Metagenomics Analysis Using OSG

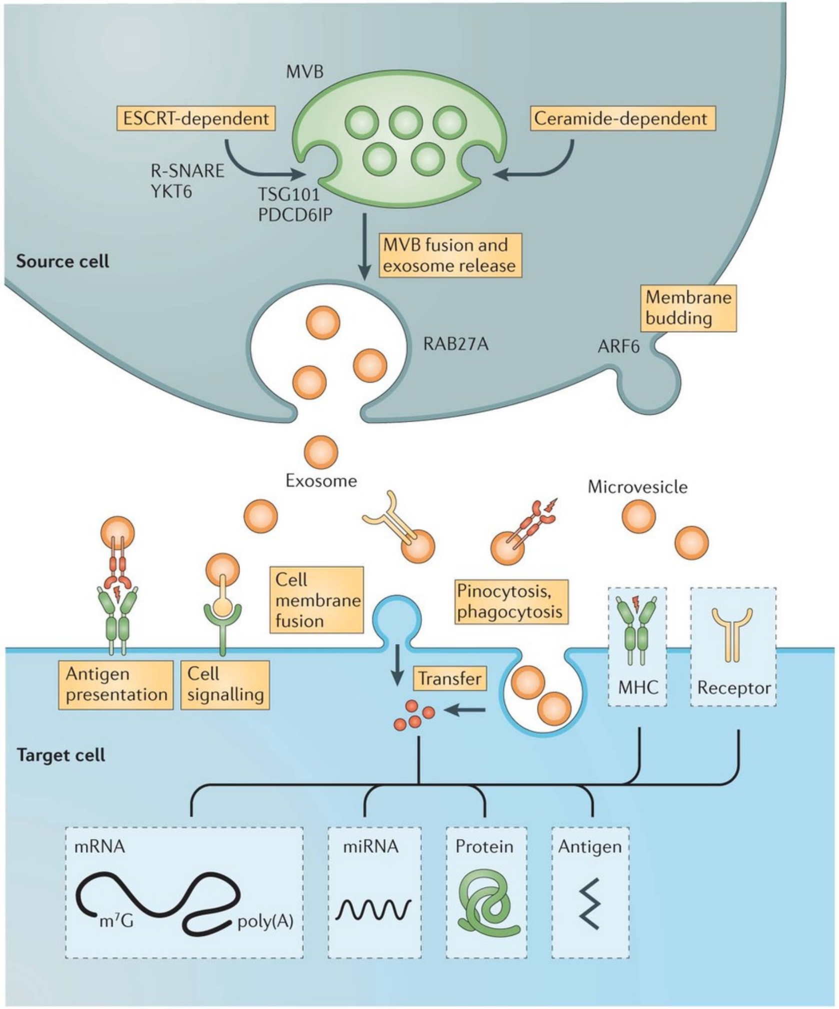
Jiang Shu

Systems Biology and Biomedical Informatics Laboratory
Department of Computer Science & Engineering
University of Nebraska-Lincoln

OSG All Hands Meeting @ UCSD
March 7th, 2017

Outline

- Background
  - Exosomes
  - RNA Sequencing Data Analysis
- Motivation
- Approach
  - Computational Challenges
  - Why OSG fits perfectly?
  - Files, scripts and submission
- Results
- OSG -- aftermath



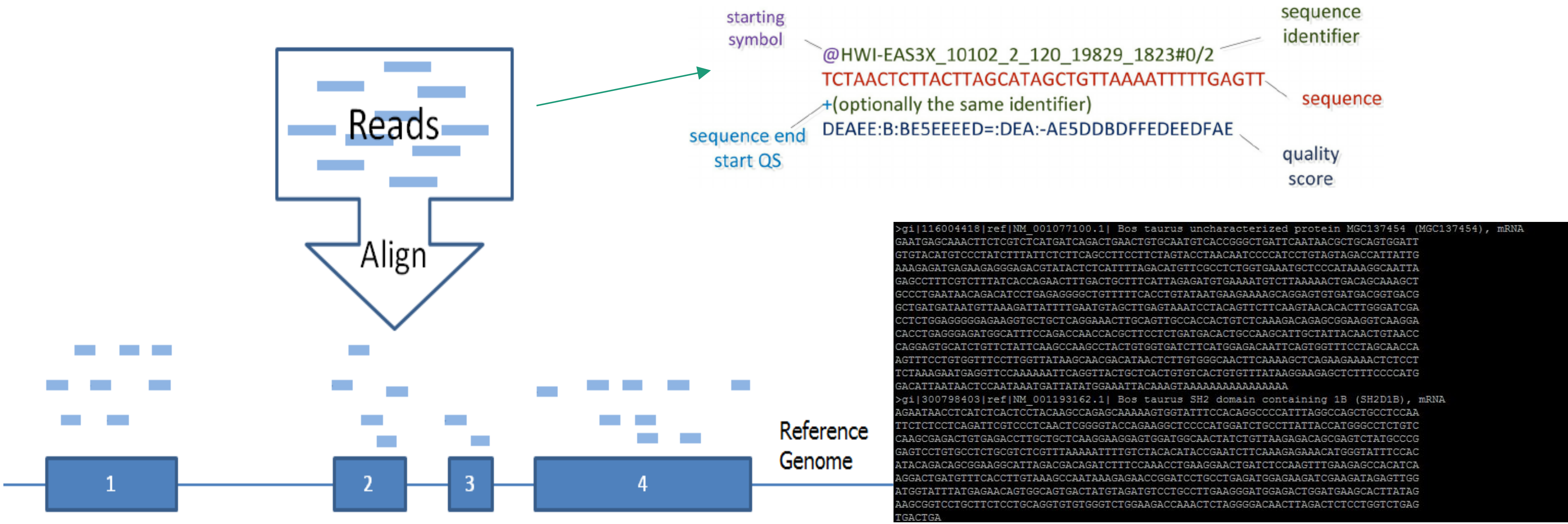
Nature Reviews | Drug Discovery

Background

- Exosomes
  - Nanoparticles (40-100 nm) present in biological fluids such as blood.
  - Play an important role in cell-to-cell communication.

Background

RNA sequencing data analysis



```
@HWI-EAS3X_10102_2_120_19829_1823#0/2
TCTAACTCTTACTTAGCATAGCTGTTAAAATTTTTGAGTT
+(optionally the same identifier)
DEAEE:B:BE5EEEED=:DEA:-AE5DDBDFFEDEEDFAE
```

```
>gi|116004418|ref|NM_001077100.1| Bos taurus uncharacterized protein MGC137454 (MGC137454), mRNA
GAATGAGCAAACTTCTCGTCTCATGATCAGACTGAACTGTGCAATGTCACCGGGCTGATTCAATAACGCTGCAGTGGATT
GTGTACATGTCCCTATCTTTATTCTCTTCAGCCTTCCTTCTAGTACCTAACAATCCCCATCCTGTAGTAGACCATTATTG
AAAGAGATGAGAAGAGGGAGACGTATACTCTCATTTTAGACATGTTCGCCTCTGGTGAAATGCTCCCATAAAGGCAATTA
GAGCCTTTCGTCTTTATCACCAGAACTTTGACTGCTTTCATTAGAGATGTGAAAATGTCTTAAAAACTGACAGCAAAGCT
GCCCTGAATAACAGACATCCTGAGAGGGGCTGTTTTTCACCTGTATAATGAAGAAAAGCAGGAGTGTGATGACGGTGACG
GCTGATGATAATGTTAAAGATTATTTTGAATGTAGCTTGAGTAAATCCTACAGTTCTTCAAGTAACACACTTGGGATCGA
CCTCTGGAGGGGGAGAAGGTGCTGCTCAGGAAACTTGCAGTTGCCACCACTGTCTCAAAGACAGAGCGGAAGGTCAAGGA
CACCTGAGGGAGATGGCATTTCCAGACCAACCACGCTTCCTCTGATGACACTGCCAAGCATTGCTATTACAACTGTAACC
CAGGAGTGCATCTGTTCTATTCAAGCCAAGCCTACTGTGGTGATCTTCATGGAGACAATTCAGTGGTTTCCTAGCAACCA
AGTTTCCTGTGGTTTCCTTGGTTATAAGCAACGACATAACTCTTGTGGGCAACTTCAAAAGCTCAGAAGAAAACTCTCCT
TCTAAAGAATGAGGTTCCAAAAAAATTCAGGTTACTGCTCACTGTGTCACTGTGTTTATAAGGAAGAGCTCTTTCCCCATG
GACATTAATAACTCCAATAAATGATTATATGGAAATTACAAAGTAAAAAAAAAAAAAAAA
>gi|300798403|ref|NM_001193162.1| Bos taurus SH2 domain containing 1B (SH2D1B), mRNA
AGAATAACCTCATCTCACTCCTACAAGCCAGAGCAAAAAGTGGTATTTCCACAGGCCCCATTTAGGCCAGCTGCCTCCAA
TTCTCTCCTCAGATTCGTCCCTCAACTCGGGGTACCAGAAGGCTCCCCATGGATCTGCCTTATTACCATGGGCCTCTGTC
CAAGCGAGACTGTGAGACCTTGCTGCTCAAGGAAGGAGTGGATGGCAACTATCTGTTAAGAGACAGCGAGTCTATGCCCG
GAGTCCTGTGCCTCTGCGTCTCGTTTAAAAATTTTGTCTACACATACCGAATCTTCAAAGAGAAACATGGGTATTTCCAC
ATACAGACAGCGGAAGGCATTAGACGACAGATCTTTCCAAACCTGAAGGAACTGATCTCCAAGTTTGAAGAGCCACATCA
AGGACTGATGTTTCACCTTGTAAAGCCAATAAAGAGAACCGGATCCTGCCTGAGATGGAGAAGATCGAAGATAGAGTTGG
ATGGTATTTATGAGAACAGTGGCAGTGACTATGTAGATGTCCTGCCTTGAAGGGATGGAGACTGGATGAAGCACTTATAG
AAGCGGTCCTGCTTCTCCTGCAGGTGTGTGGGTCTGGAAGACCAAACTCTAGGGGACAACTTAGACTCTCCTGGTCTGAG
TGACTGA
```

Motivation

- In a pervious project, we have isolated exosomes from one type of body fluid of one host species and assessed the molecules inside the exosomes.
- Moreover, we also found many unmapped reads are from microbial species.
- Thus, we designed a follow-up study to **understand the origin of microbial sequences in the exosomes of this type of body fluid.**
  - **Metagenomics analysis**

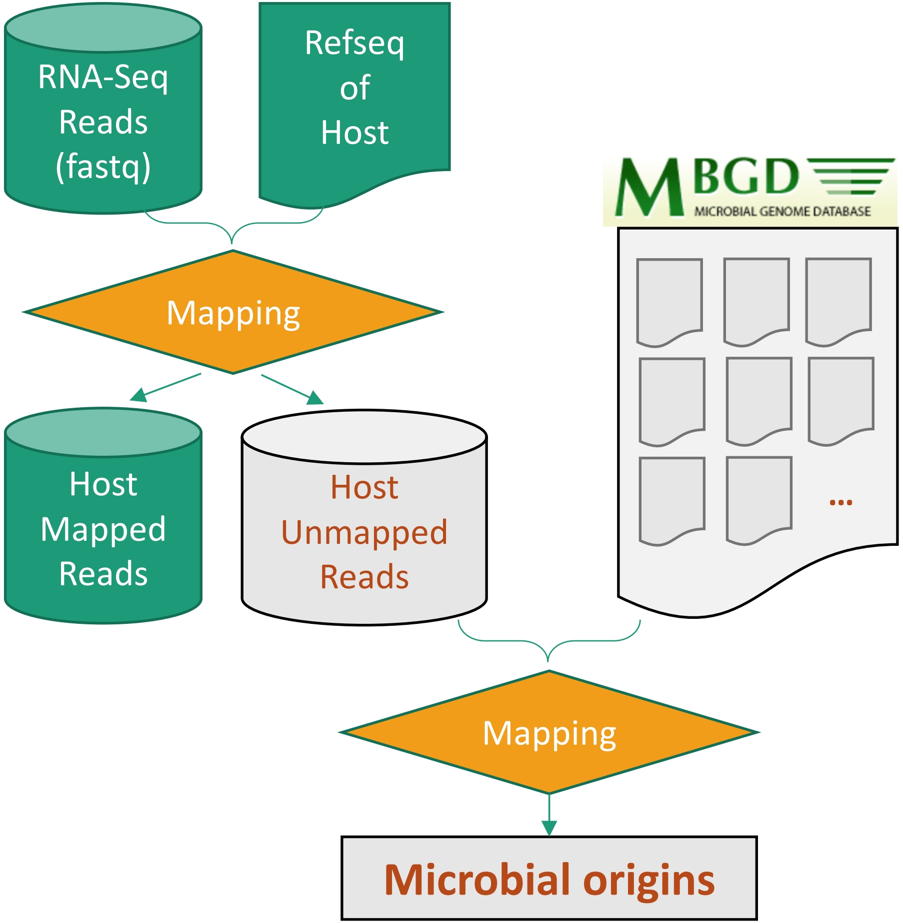
Approach

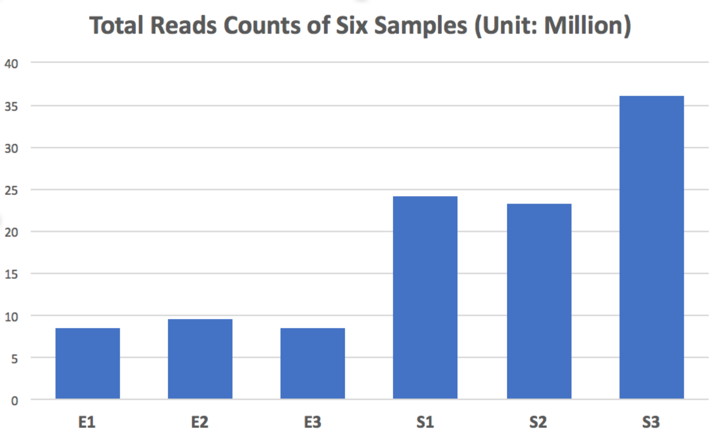
Two layers of analysis:

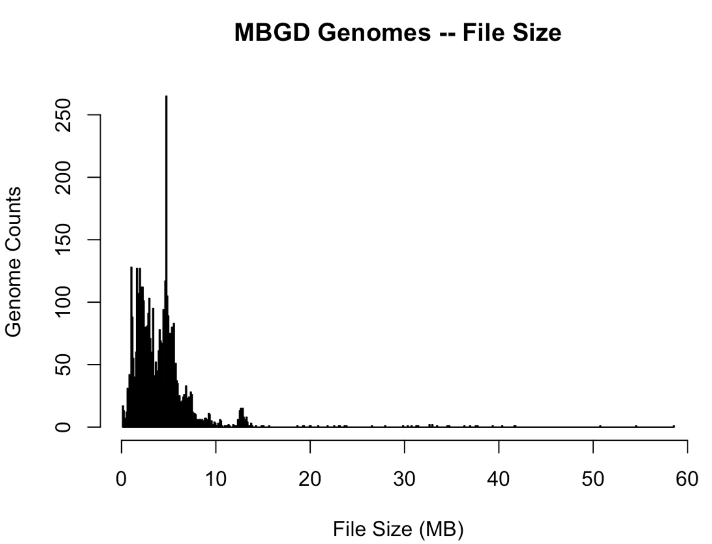
- Extract the reads cannot map to host genome
- Identify the microbial species through another level reads mapping based on host unmapped reads

Microbial genome database:

- 4,742 microbial genomes were downloaded from ***MBGD***.







Computational challenges

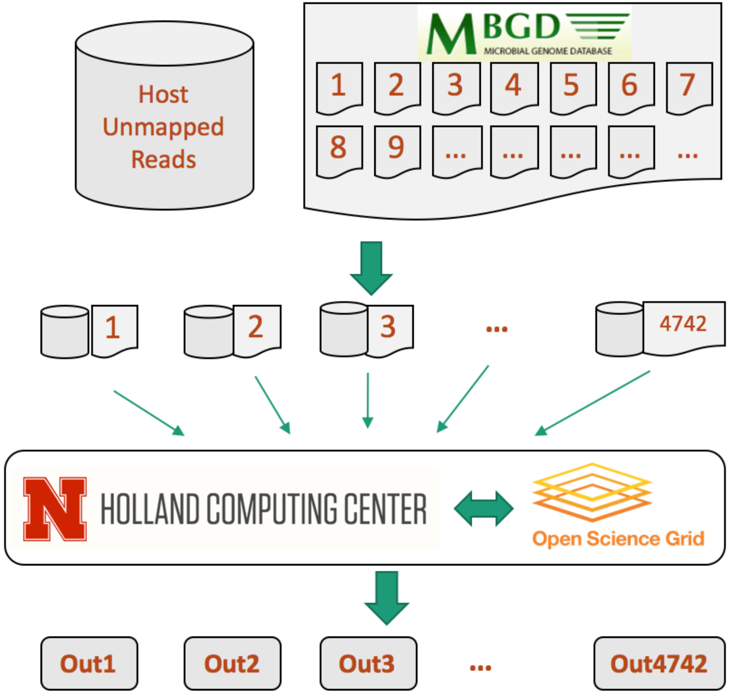
- A large number of target genomes
  - 4,742 genomes (size: 100KB ~ 58MB)
- Six samples contain over 100 million of host unmapped reads
  - Increasing the computing time
- In total, 6 x 4,742 = ***28,452*** mapping tasks

Question: Where to execute this many of jobs?

- Impossible for the lab-server (32 cores)
- Long pending time if submitted it to HCC clusters
  - Dynamic priority scheduling of users/groups
  - More jobs completed -> longer queue time

Perfect Fit of Open Science Grid (OSG)

- The tasks are independent to each other
- Limited file transfer
  - Total size of transferred files ~1GB
- Small memory consumptions
  - Memory < 2GB
- Short running time for each task
  - Maximum: 3 hours (HCC@UNL-Crane)
- Software is available on OSG
  - Pre-installed Bowtie and Tophat
  - No further configuration needed



OSG Preparation: Files

- Input files that transfer to the executing node on OSG
  - Fastq files of each sample
  - Target genome file: `*.dnaseq`

- Output files that transfer back from the executing node on OSG
  - Mapping results file: `accepted_hits.bam` (~30MB)
  - Mapping summary file: `align_summary.txt` (~0.1KB)

- Standard system files:
  - `*.out, *.err, *.log` (~10KB)

OSG Preparation: Scripts

exe.sh

```
#!/bin/bashsource
/cvmfs/oasis.opensciencegrid.org/osg/modules/lmod/5.6.2/init/bash
module load libgfortran/4.4.7
module load bowtie
module load tophat
bowtie2-build "$2".dnaseq "$2”
tophat -o ./  "$2" "$1"_R1.fastq.gz "$1"_R2.fastq.gz
echo `hostname`
```

Load the pre-installed software

$1 -> sample name,
$2 -> target genome name

Software commands

job.submit

```
universe = vanilla
executable = exe.sh
arguments  = "E1 afd”
```

$1 and $2 in exe.sh

```
error = E1_afd.err
log = E1_afd.log
output = E1_afd.out
```

System files, help on debugging.

```
should_transfer_files = YES
when_to_transfer_output = ON_EXIT

transfer_input_files = exe.sh, afd.dnaseq, E1_R1.fastq, E1_R2.fastq.gz

transfer_output_files = accepted_hits.bam, align_summary.txt

Requirements = (HAS_MODULES =?= TRUE)
on_exit_hold = (ExitBySignal == True) || (ExitCode != 0)
periodic_release =  (NumJobStarts < 2) && ((CurrentTime - EnteredCurrentStatus) > 60)

queue
```

OSG Preparation: Submission

- All jobs were submitted from login nodes of HCC@UNL-Crane to Open Science Grid

    -
      ```
      $ Condor_submit job.submit
      ```

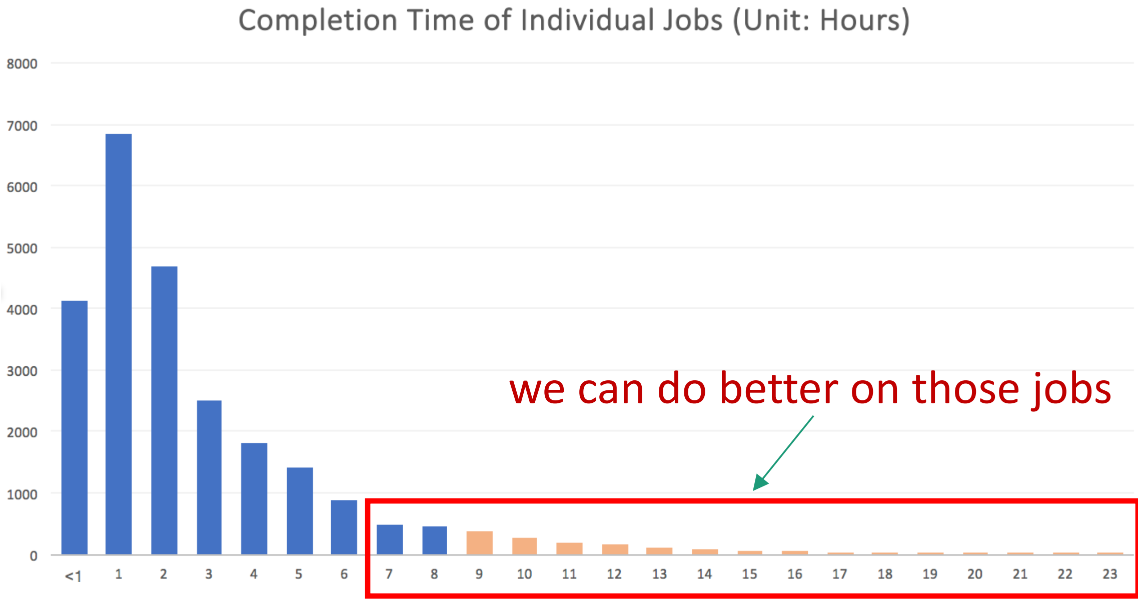
HOLLAND COMPUTING CENTER
Open Science Grid

Results

- Several microbial species were identified in the exosomes of this type of body fluid in the host species
- Although some microbial species have been reported in this host species before, this is the first time of identifying microbial sequences in the exosomes of this specific body fluid
- Based on the findings from this analysis, we have designed two experiments to further our understanding in this subject

OSG -- aftermath

- Total computation
  - ~84K CPU hours or 9.2 years
- Completed in
  - 408 hours or 17 days
- At average, ~2,500 jobs were running simultaneously
- 93% of jobs could be completed in 8 hours



Completion Time of Individual Jobs (Unit: Hours)

we can do better on those jobs

NEBRASKA CENTER FOR THE PREVENTION OF OBESITY DISEASES

NATIONAL INSTITUTES OF HEALTH

HOLLAND COMPUTING CENTER

Open Science Grid

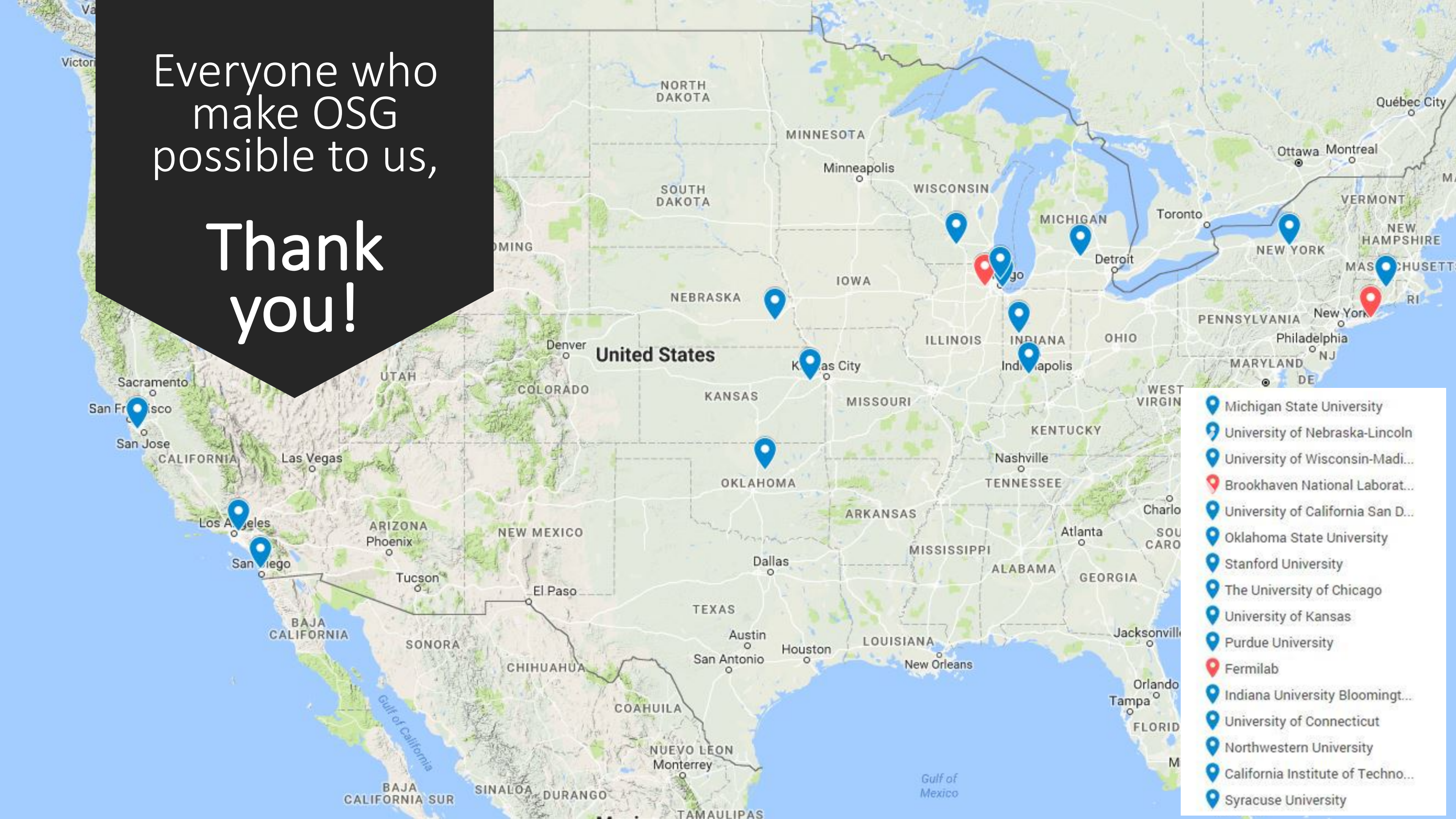
OSG USER SCHOOL 2016
Harness the power of distributed computing

Acknowledgements

- Dr. Janos Zempleni’s group at UNL for sample preparation
  - Di Wu and Dr. Bijaya Upadhyaya
- This work is support by National Institutes of Health (1P20GM104320)
- The discussions and encouragements from HCC staffs that greatly assisted the work
  - *Dr. Emelie Harstad*
  - *Dr. David Swanson*
  - *Natasha Pavlovikj*
  - *Dr. Derek Weitzel*
  - *Dr. Jingchao Zhang*
- *And …*

Everyone who make OSG possible to us,

Thank you!

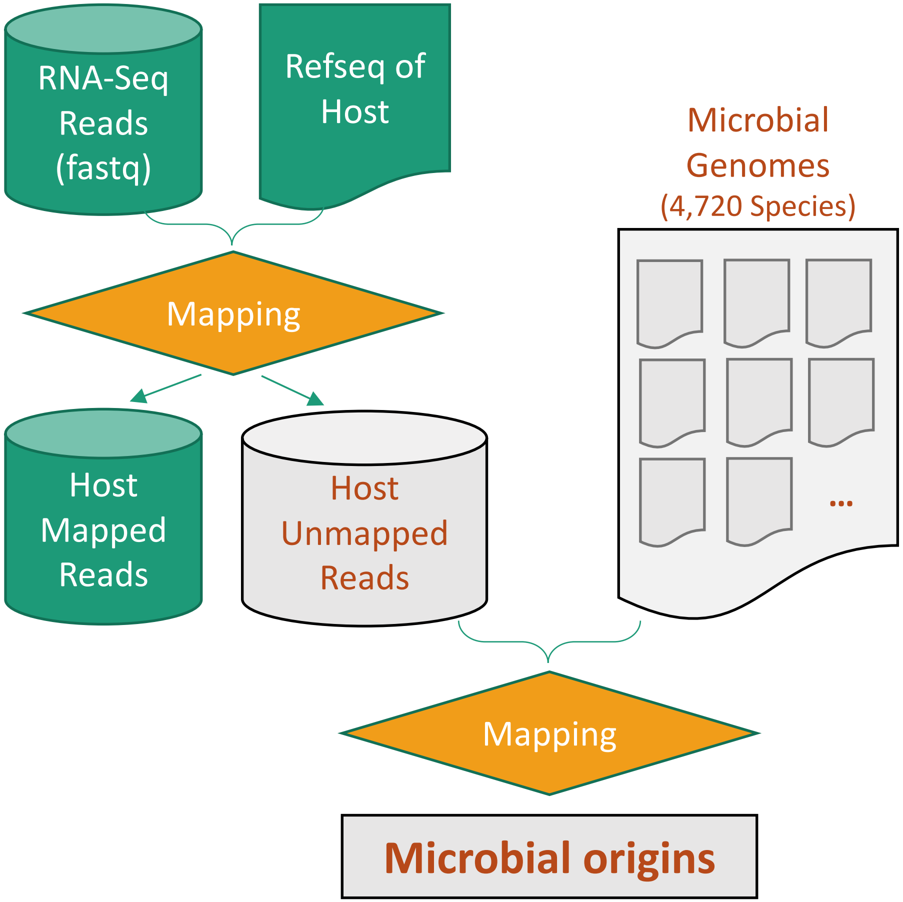


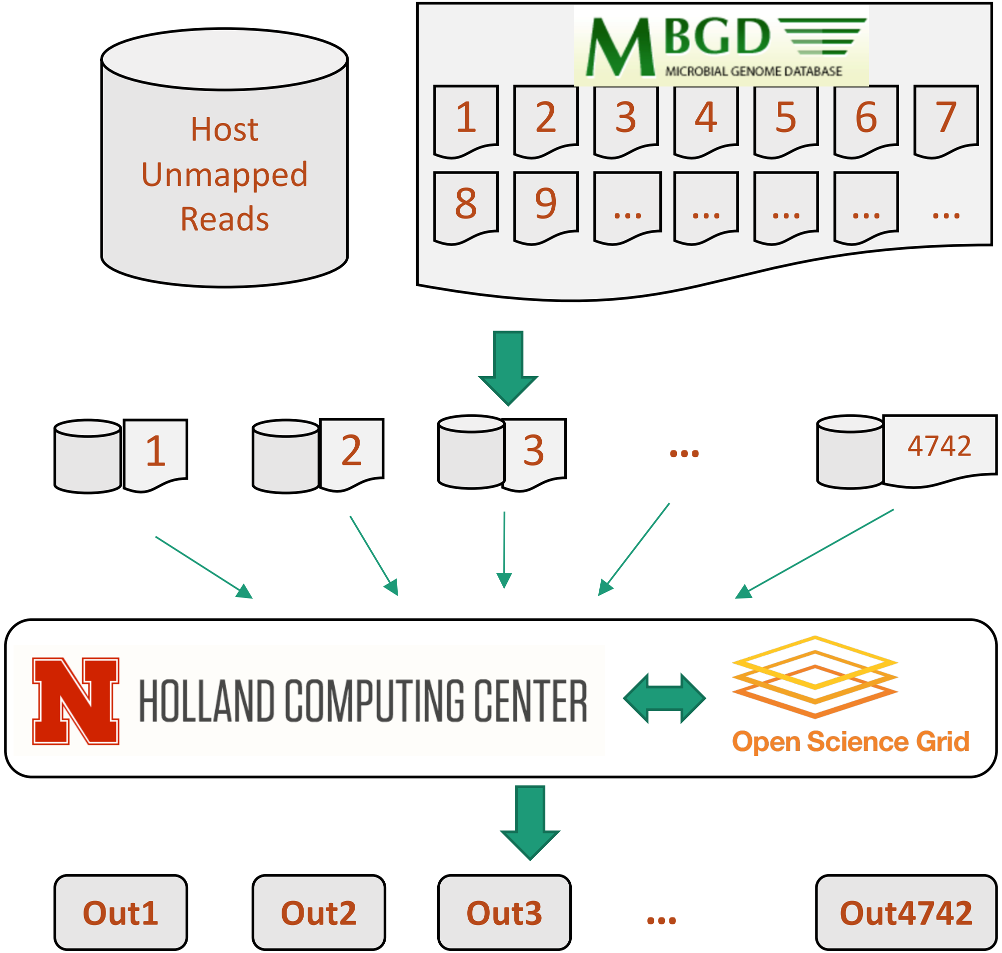
- Michigan State University
- University of Nebraska-Lincoln
- University of Wisconsin-Madi...
- Brookhaven National Laborat...
- University of California San D...
- Oklahoma State University
- Stanford University
- The University of Chicago
- University of Kansas
- Purdue University
- Fermilab
- Indiana University Bloomingt...
- University of Connecticut
- Northwestern University
- California Institute of Techno...
- Syracuse University

SBBI and OSG

- In 2016, our team used 2 million CPU hours on OSG on following projects:
  - Metagenomics analysis
    - Bioinformatics analysis, Bowtie and Tophat
  - microRNA target prediction at genome scale
    - Machina Learning, Python
  - Gene regulatory network prediction in cancers
    - Statistical modeling, R
- We appreciate the continued support from HCC@UNL and OSG.

Jiang Shu
jshu2@unl.edu


RNA-Seq Reads (fastq)
Refseq of Host
Microbial Genomes (4,720 Species)
Mapping
Host Mapped Reads
Host Unmapped Reads
...
Mapping
Microbial origins


Host Unmapped Reads
MBGD MICROBIAL GENOME DATABASE
1
2
3
4
5
6
7
8
9
…
…
…
…
…
1
2
3
…
4742
HOLLAND COMPUTING CENTER
Open Science Grid
Out1
Out2
Out3
…
Out4742