Introduction to Metagenomics Analysis for Next Generation Sequencing Data

Noushin Ghaffari, PhD
Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research
Research Scientist, Texas A&M High Performance Research Computing

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

Outline

- Background
  - Sequencing
- Application of Next Generation Sequencing in Research
  - Metagenomics

Primary NGS Applications

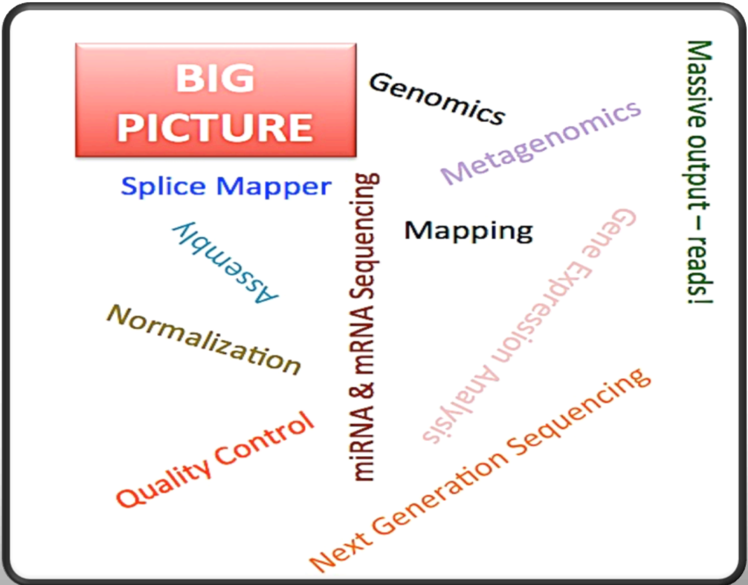
1. Alignment
2. Assembly(no reference/with a reference)
   - Genome
   - Transcriptome
3. RNA-Seq
4. **Metagenomics**
5. ChIP-Seq
6. RADSeq

Last week → 2. Assembly

Two weeks ago → 3. RNA-Seq

Today → 4. Metagenomics

Next Month → 6. RADSeq



Why sequencing?

Determining the sequence of nucleotides within a DNA (or RNA) fragment

How?

Using sequencing methods, such as Sanger sequencing, next generation sequencing and single-molecule techniques

Sanger

Classic Sequencing

Next Generation Sequencing Platforms

Illumina

© 2014 Illumina, Inc. All rights reserved.

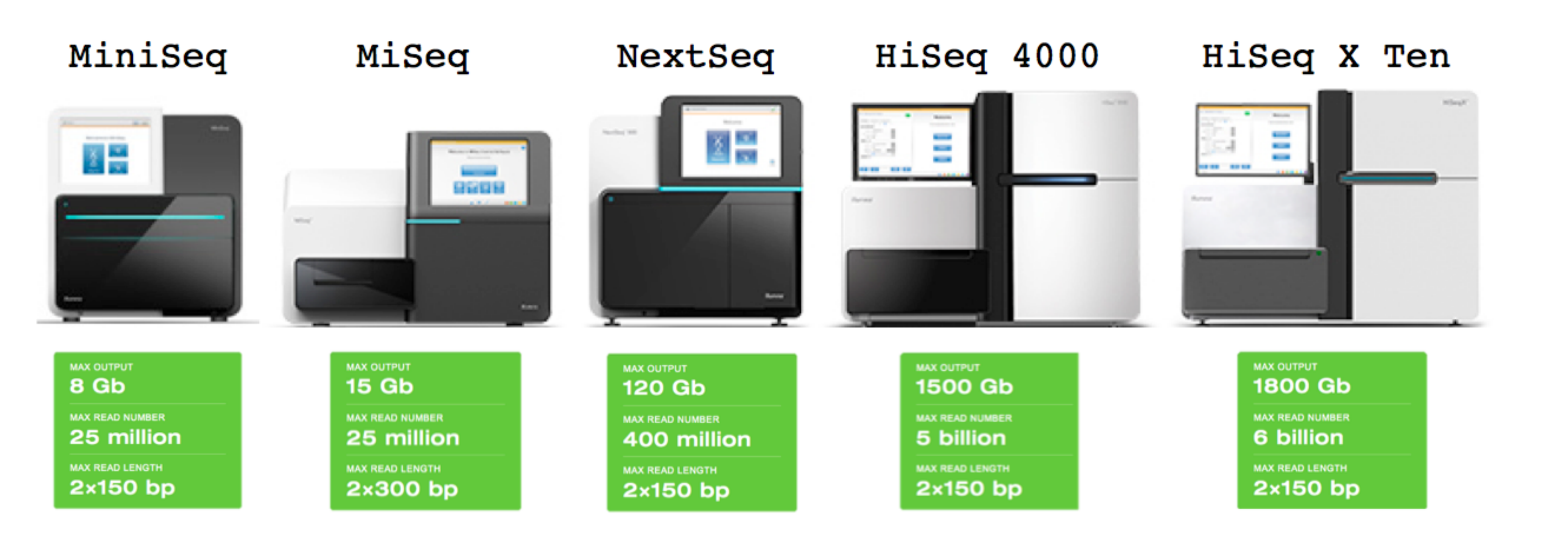
http://nextgenseek.com/2014/01/illumina-announces-new-sequencers-hiseqx-nextseq-500-at-jpm-2014/

Third Generation Sequencing Platforms

PacBio

MinION

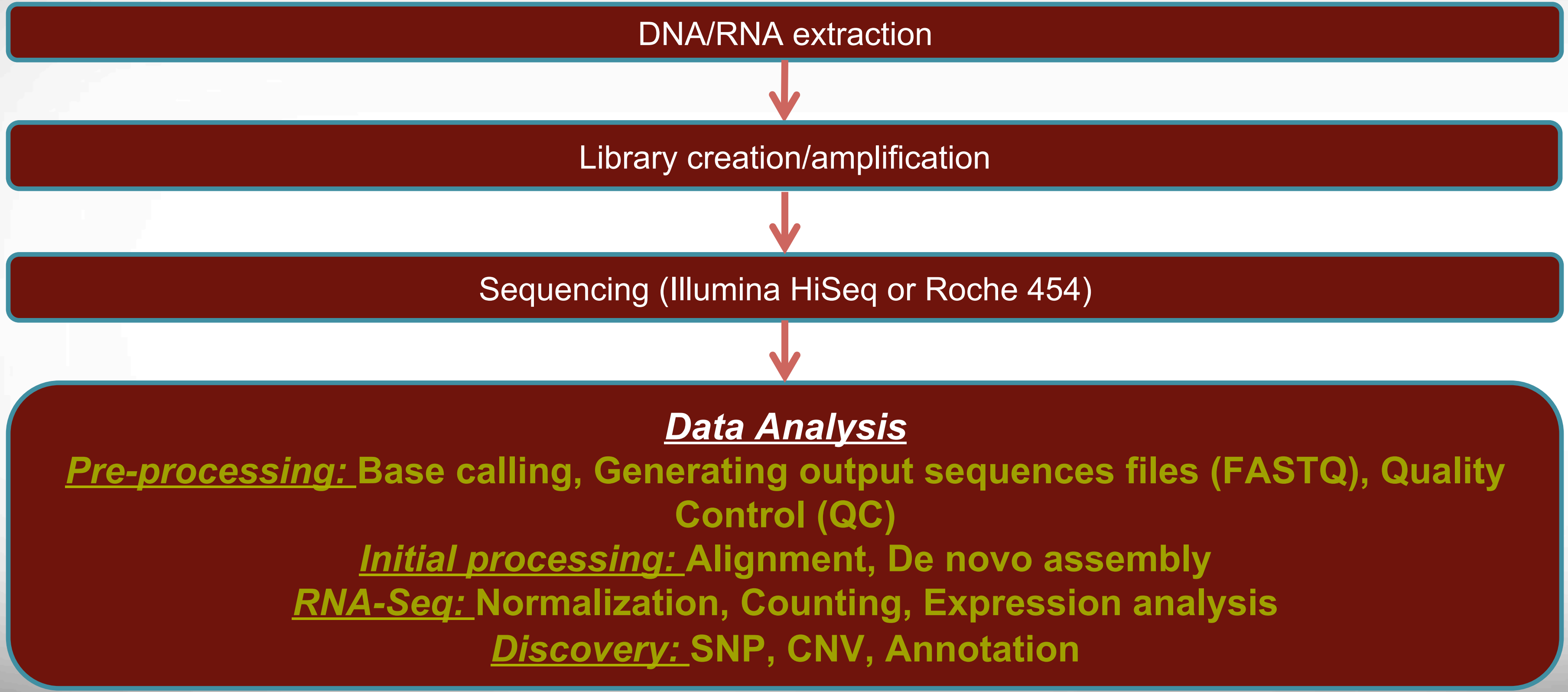
Choosing among Illumina Sequencers



	MiniSeq	MiSeq	NextSeq	HiSeq 4000	HiSeq X Ten
MAX OUTPUT	8 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
MAX READ NUMBER	25 million	25 million	400 million	5 billion	6 billion
MAX READ LENGTH	2x150 bp	2x300 bp	2x150 bp	2x150 bp	2x150 bp

http://core-genomics.blogspot.com/2016/01/meet-newest-members-of-family-miniseq.html

NGS Sequencing Workflow



DNA/RNA extraction

Library creation/amplification

Sequencing (Illumina HiSeq or Roche 454)

Data Analysis

Pre-processing: **Base calling, Generating output sequences files (FASTQ), Quality Control (QC)**

Initial processing: **Alignment, De novo assembly**

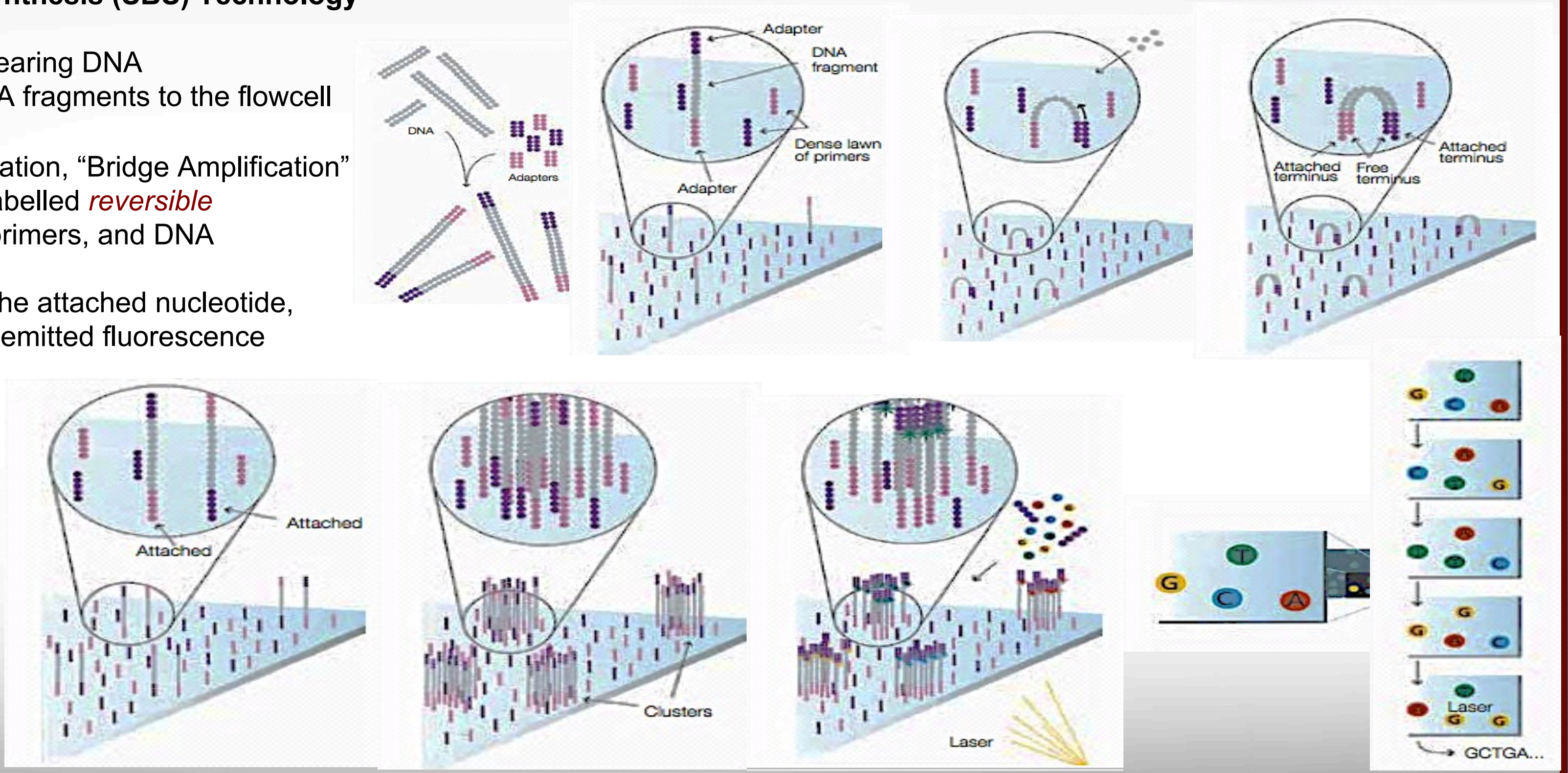
RNA-Seq: **Normalization, Counting, Expression analysis**

Discovery: **SNP, CNV, Annotation**

Illumina next-generation sequencing

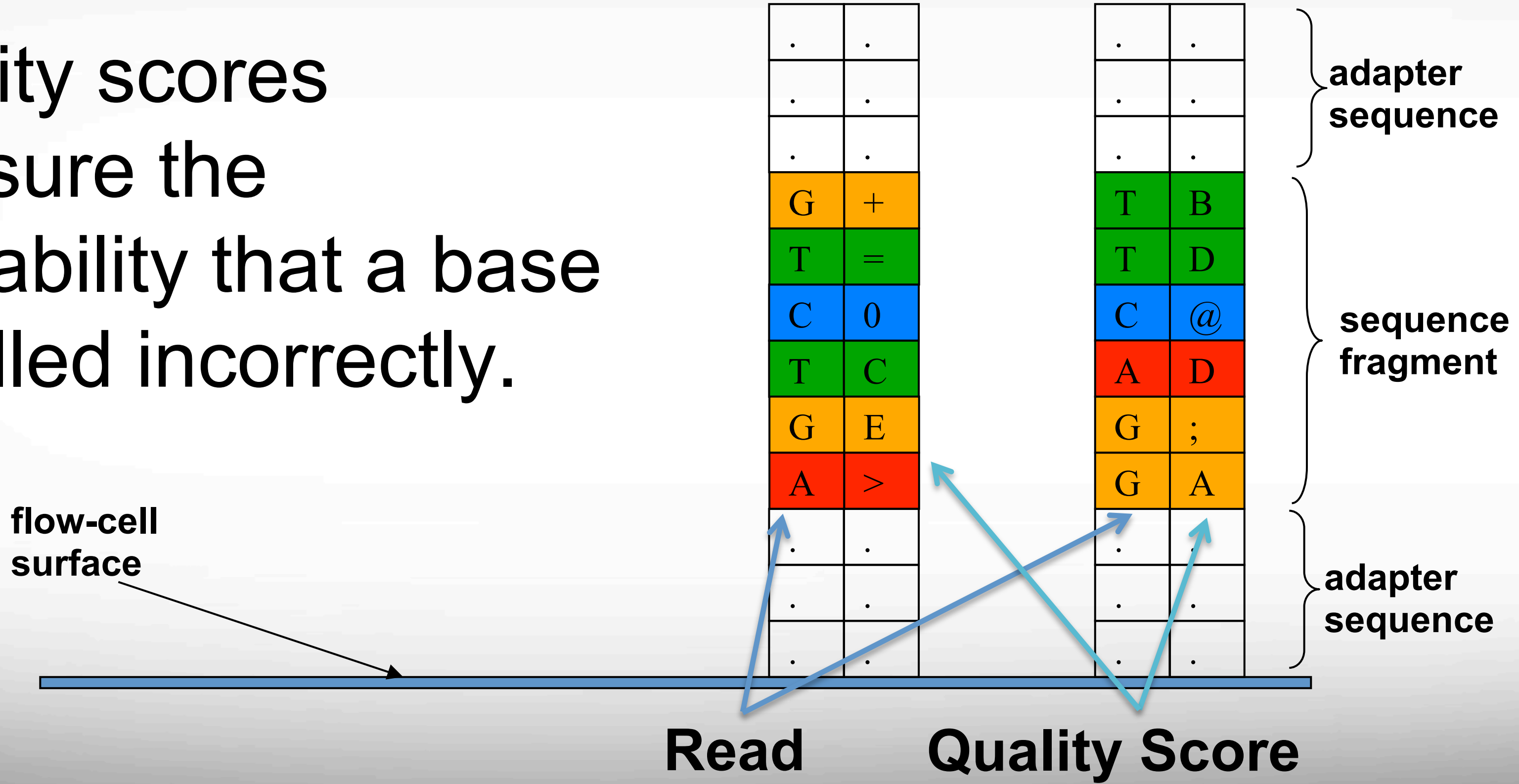
Sequencing by Synthesis (SBS) Technology

- Randomly shearing DNA
- Attaching DNA fragments to the flowcell surface
- Cluster generation, “Bridge Amplification”
- Adding four labelled *reversible terminators*, primers, and DNA polymerase
- Determining the attached nucleotide, based on the emitted fluorescence



Sequence and Quality Scores

Quality scores measure the probability that a base is called incorrectly.



Comparing Sequencing Platforms

	Read length	Error rates	Technology	Portable?
Illumina	< 400 bp	Low	Sequencing by synthesis	No
PacBio	~ 10-15 Kb	High	SMRT – ZMW	No
Oxford Nanopore Technologies	~ 5-8 Kb	High	Nanopore protein – strand sequencing	Yes

Metagenomics

What is Metagenomics?

Study of communities of microbial organisms directly in their natural environments
Without the need for isolation and lab cultivation of individual species

Moved from traditional BAC cloning to NGS long reads or high coverage short reads

- Chen K and Pachter L, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities, *PLoS Comput Biol,*

Metagenomics Techniques

1. Whole Genome Shotgun (WGS)
2. Marker Gene
   - 16S rRNA
     - Bacteria, Archaea
   - 18S rRNA
     - Fungus, Eukaryotes

- Chen K and Pachter L, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities, *PLoS Comput Biol,*

Whole Genome Shotgun (WGS) Metagenomics

- Sequencing the whole genome of the organisms present in the sample
- Facilitates discovering gene/gene function, genome structure
- Studying the evolutionary relationships for microbiomes

- Chen K and Pachter L, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities, *PLoS Comput Biol,*

Whole Genome Shotgun (WGS) Metagenomics

- Sequencing the whole genome of the organisms present in the sample
- Facilitates discovering gene/gene function, genome structure
- Studying the evolutionary relationships for microbiomes
- Steps
  - Genome Assembly
  - Binning
  - Predicting and Annotating Genes

- Chen K and Pachter L, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities, *PLoS Comput Biol,*

WGS Metagenomics Tools

Assembly

- Velvet, MetaVelvet, MetaVelvet-SL
- IDBA-UD
- MetAMOS pipeline: selecting assembly, scaffolding, annotating
- Genome Assemblers such as ALLPATHS, SOAP and ABySS

Binning

- LikelyBin
- PHYSCIMM
- MetaCluster
- MetaWatt
- MetaPhyler
- PhymmBL

Annotation

- MetaGeneAnnotator
- Glimmer-MG
- FragGeneScan
- MetaGeneMark
- Kraken

Marker Gene Metagenomics

- Usually based on 16S RNA
  - Conserved within species
  - Greatly different between species
  - Widely used for microbial ecology
- Needs a reference database to match the Operational Taxonomic Units (OTU)
  - Silva
  - Ribosomal Database Project
  - Unite
- Steps
  - Preprocessing to remove noise
  - OUT clustering and taxonomic assignment
  - Alpha diversity analysis – within sample diversity
  - Beta diversity analysis – between sample diversity

Marker Gene Metagenomics Tools

Microbial community analysis

- QIIME
- Mother
- SILVAngs
- MG-RAST
- MEGAN

Diversity analysis

- Chao
- UniFrac
- PCoA

Visualization

- QIIME
- MEGAN
- FigTree

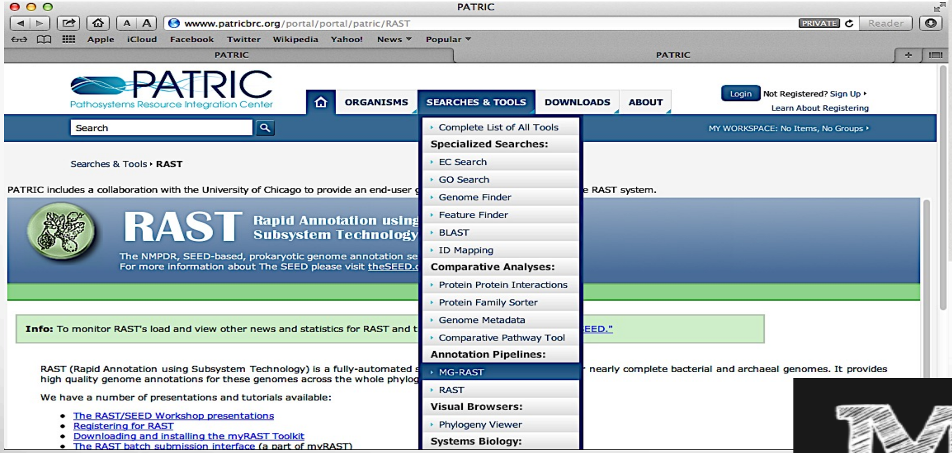
Metagenomics Studies

- PathoMap
  - Research project by Weill Cornell Medical College to study the microbiome and metagenome of the built environment of NYC
- Cow rumen microbiome study
  - 220 bacterial and archaeal genomes assembled directly from 768 GB rumen sequenced data
  - Majority unsequenced strains and species of bacteria and archaea
  - Over 13,000 proteins predicted to be involved in carbohydrate metabolism, over 90% of which do not have a good match in the public databases
  - Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism

Metagenomics Web-Based Tools

MG-RAST

- Available tools, via PATRIC
- RAST: Rapid Annotation using Subsystem Technology
- Annotating the assembled contigs of a bacterial and archaeal genomes
- Quantitative insights for microbial populations, based on NGS data



PATRIC includes a collaboration with the University of Chicago to provide an end-user … e RAST system.

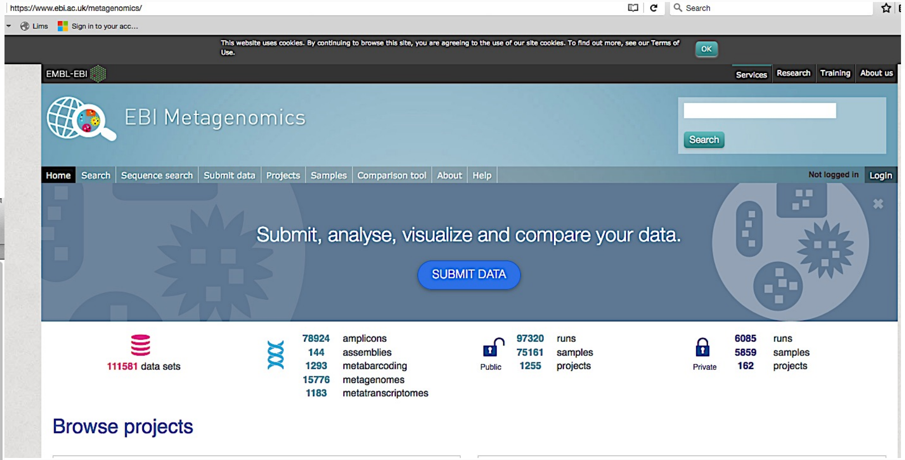
The NMPDR, SEED-based, prokaryotic genome annotation se…
For more information about The SEED please visit theSEED.…

Info: To monitor RAST's load and view other news and statistics for RAST and t… EED."

RAST (Rapid Annotation using Subsystem Technology) is a fully-automated … nearly complete bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylog…

We have a number of presentations and tutorials available:

- The RAST/SEED Workshop presentations
- Registering for RAST
- Downloading and installing the myRAST Toolkit
- The RAST batch submission interface (a part of myRAST)



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

Submit, analyse, visualize and compare your data.

111581 data sets

78924	amplicons
144	assemblies
1293	metabarcoding
15776	metagenomes
1183	metatranscriptomes

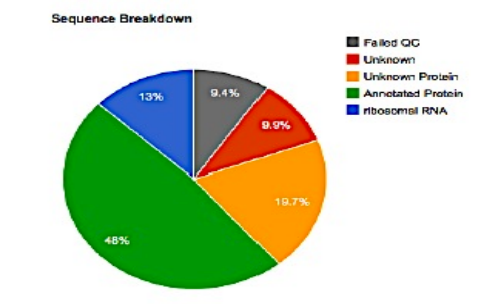
Public	
97320	runs
75161	samples
1255	projects

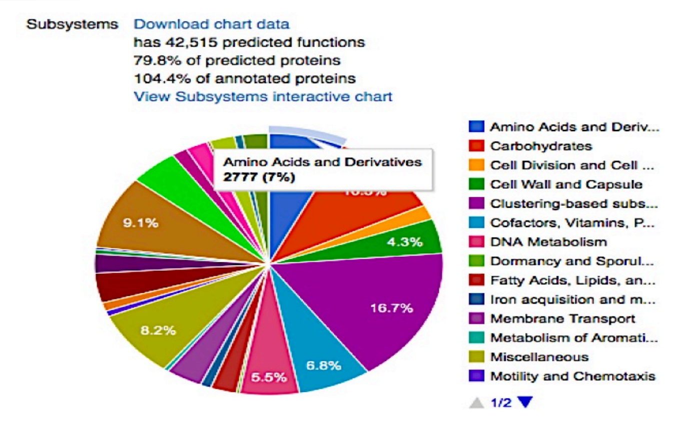
Private	
6085	runs
5859	samples
162	projects

Browse projects

MG-RAST
metagenomics analysis server

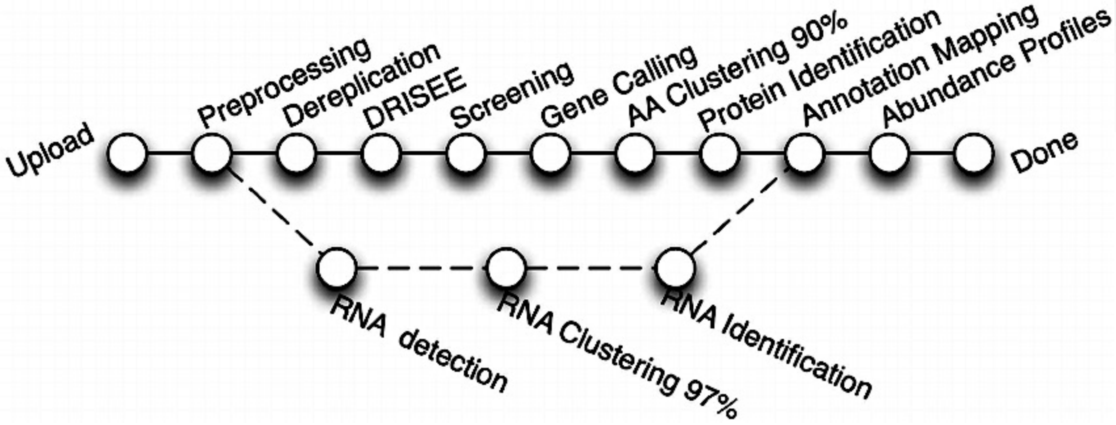
MG-RAST Pipeline

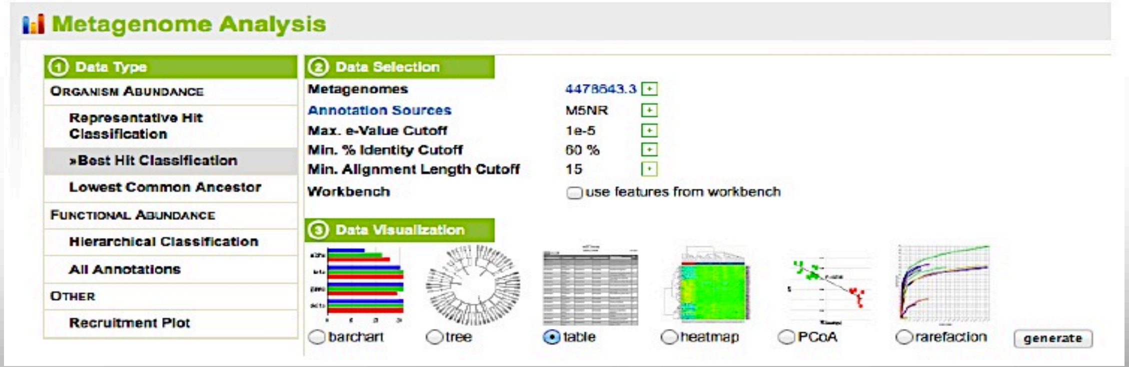




Subsystems Download chart data
has 42,515 predicted functions
79.8% of predicted proteins
104.4% of annotated proteins
View Subsystems interactive chart

Amino Acids and Derivatives
2777 (7%)
10.5%
9.1%
4.3%
16.7%
8.2%
6.8%
5.5%
■ Amino Acids and Deriv...
■ Carbohydrates
■ Cell Division and Cell ...
■ Cell Wall and Capsule
■ Clustering-based subs...
■ Cofactors, Vitamins, P...
■ DNA Metabolism
■ Dormancy and Sporul...
■ Fatty Acids, Lipids, an...
■ Iron acquisition and m...
■ Membrane Transport
■ Metabolism of Aromati...
■ Miscellaneous
■ Motility and Chemotaxis
▲ 1/2 ▼





Metagenome Analysis

① Data Type

ORGANISM ABUNDANCE

Representative Hit Classification

»Best Hit Classification

Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

Hierarchical Classification

All Annotations

OTHER

Recruitment Plot

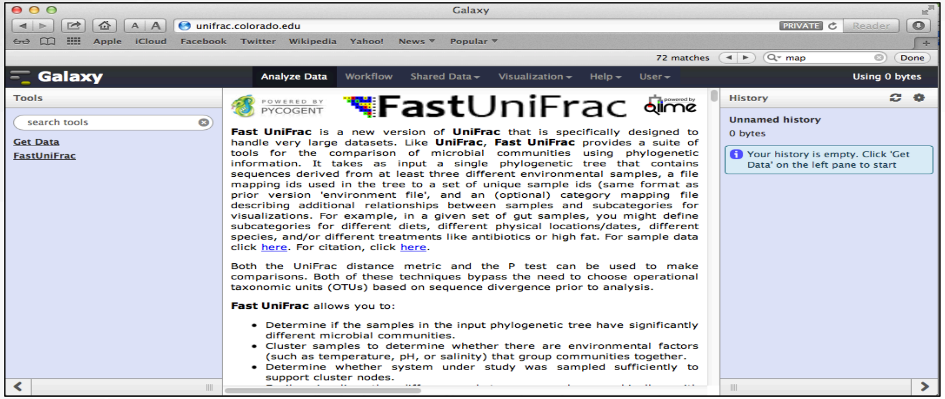
② Data Selection

Metagenomes	4478643.3	+
Annotation Sources	M5NR	+
Max. e-Value Cutoff	1e-5	+
Min. % Identity Cutoff	60 %	+
Min. Alignment Length Cutoff	15	+
Workbench	□ use features from workbench	

③ Data Visualization

○ barchart ○ tree ● table ○ heatmap ○ PCoA ○ rarefaction generate

From UniFrac to FastUniFrac



Your history is empty. Click 'Get Data' on the left pane to start

Fast UniFrac is a new version of **UniFrac** that is specifically designed to handle very large datasets. Like **UniFrac**, **Fast UniFrac** provides a suite of tools for the comparison of microbial communities using phylogenetic information. It takes as input a single phylogenetic tree that contains sequences derived from at least three different environmental samples, a file mapping ids used in the tree to a set of unique sample ids (same format as prior version 'environment file', and an (optional) category mapping file describing additional relationships between samples and subcategories for visualizations. For example, in a given set of gut samples, you might define subcategories for different diets, different physical locations/dates, different species, and/or different treatments like antibiotics or high fat. For sample data click here. For citation, click here.

Both the UniFrac distance metric and the P test can be used to make comparisons. Both of these techniques bypass the need to choose operational taxonomic units (OTUs) based on sequence divergence prior to analysis.

Fast UniFrac allows you to:

- Determine if the samples in the input phylogenetic tree have significantly different microbial communities.
- Cluster samples to determine whether there are environmental factors (such as temperature, pH, or salinity) that group communities together.
- Determine whether system under study was sampled sufficiently to support cluster nodes.

UniFrac

Calculates distance between microbial communities, using phylogenic trees

- FastUniFrac
  - Adapted for NGS data
  - Incorporated with Galaxy tools
  - The same idea as UniFrac

UniFrac vs FastUniFrac

- Input: tree
- Newick (PHYLIP package output) or Nexus (TAXA, CHARACTER, DATA,TREE blocks (Newick format))
- Tagging each sequence's environment, Creating Sample ID map
- Analysis
- Measuring the overall difference between each pair of environments
- Clustering the environments
- Principal Coordinates Analysis (3D in FastUniFrac)

- M Hamady, C Lozupone and R Knight, "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data", *The ISME Journal* 4, 17–27 2010.

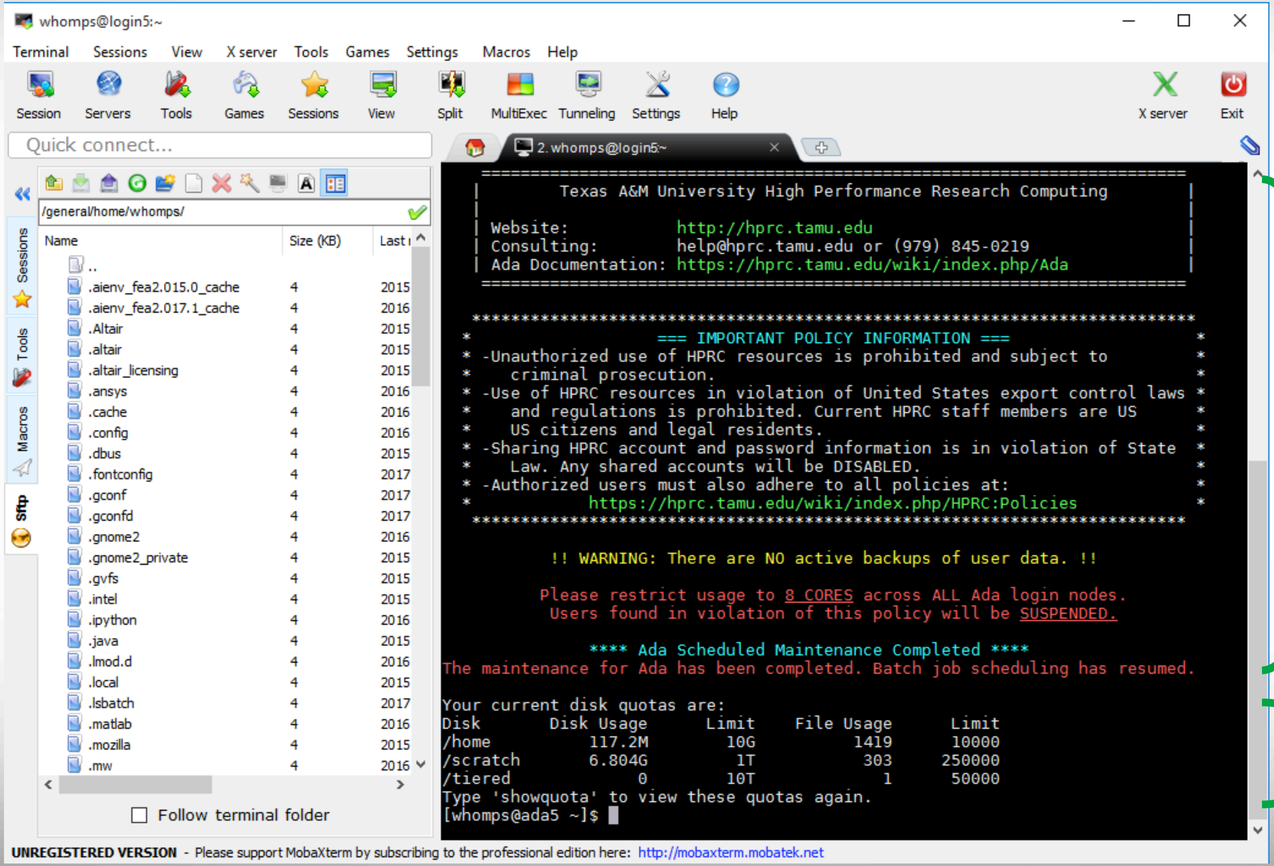
Practical Portion

Logging in to the system

- SSH (secure shell)
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- For Microsoft Windows PCs, use *MobaXterm*
    - https://hprc.tamu.edu/wiki/HPRC:MobaXterm
      - You are able to view images and use GUI applications with MobaXterm
  - or *Putty*
    - https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY
      - You can not view images or use GUI applications with PuTTY

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently

Using SSH - MobaXterm (on Windows)



/general/home/whomps/

Name	Size (KB)	Last r
..		
.aienv_fea2.015.0_cache	4	2015
.aienv_fea2.017.1_cache	4	2016
.Altair	4	2015
.altair	4	2015
.altair_licensing	4	2015
.ansys	4	2016
.cache	4	2016
.config	4	2016
.dbus	4	2015
.fontconfig	4	2017
.gconf	4	2017
.gconfd	4	2017
.gnome2	4	2016
.gnome2_private	4	2015
.gvfs	4	2015
.intel	4	2015
.ipython	4	2016
.java	4	2015
.lmod.d	4	2016
.local	4	2015
.lsbatch	4	2017
.matlab	4	2016
.mozilla	4	2015
.mw	4	2016

```
  ==========================================================================
 |        Texas A&M University High Performance Research Computing          |
 |                                                                           |
 | Website:          http://hprc.tamu.edu                                    |
 | Consulting:       help@hprc.tamu.edu or (979) 845-0219                    |
 | Ada Documentation: https://hprc.tamu.edu/wiki/index.php/Ada               |
  ==========================================================================

 ****************************************************************************
 *                    === IMPORTANT POLICY INFORMATION ===                  *
 * -Unauthorized use of HPRC resources is prohibited and subject to         *
 *    criminal prosecution.                                                 *
 * -Use of HPRC resources in violation of United States export control laws *
 *    and regulations is prohibited. Current HPRC staff members are US      *
 *    US citizens and legal residents.                                      *
 * -Sharing HPRC account and password information is in violation of State  *
 *    Law. Any shared accounts will be DISABLED.                            *
 * -Authorized users must also adhere to all policies at:                   *
 *           https://hprc.tamu.edu/wiki/index.php/HPRC:Policies             *
 ****************************************************************************

          !! WARNING: There are NO active backups of user data. !!

          Please restrict usage to 8 CORES across ALL Ada login nodes.
           Users found in violation of this policy will be SUSPENDED.

               **** Ada Scheduled Maintenance Completed ****
The maintenance for Ada has been completed. Batch job scheduling has resumed.

Your current disk quotas are:
Disk          Disk Usage      Limit    File Usage      Limit
/home             117.2M        10G          1419      10000
/scratch          6.804G         1T           303     250000
/tiered                0        10T             1      50000
Type 'showquota' to view these quotas again.
[whomps@ada5 ~]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

message of the day

your quotas

Using SSH to Access Ada

```
ssh –X user_NetID@ada.tamu.edu
```

https://hprc.tamu.edu/wiki/Ada:Access

You may see something like the following the first time you connect to the remote machine from your local machine:

```
Host key not found from the list of known hosts.
Are you sure you want to continue connecting (yes/no)?
```

Type yes, hit enter and you will then see the following:

```
Host 'ada.tamu.edu' added to the list of known hosts.
user_NetID@ada.tamu.edu's password:
```

Metagenomics Practice - Tool

Mothur

- Open-source
- Serves the microbial ecology community
- DOTUR and SONS programs
- Data: Sanger, PacBio, IonTorrent, 454, and Illumina MiSeq and HiSeq
- Most cited bioinformatics tool for analyzing 16S rRNA gene sequences
- Our practical session will use Mothur to demonstrate a typical MiSeq data analysis

Metagenomics Practice - Data

- Objective: Understanding the effect of normal variation in the gut microbiome on host health
- Collected 365 (daily basis) fresh feces from mice, post weaning
- No treatment in first 150 days post weaning (dpw)

- Question: rapid change in weight observed during the first 10 dpw affected the stability microbiome when compared with microbiome in days 140 – 150 or not?
- Mock community composed of genomic DNA from 21 bacterial strains

one animal at 10 time points

group	**time**
F3D0	Early
F3D1	Early
F3D141	Late
F3D142	Late
F3D143	Late
F3D144	Late
F3D145	Late
F3D146	Late
F3D147	Late
F3D148	Late
F3D149	Late
F3D150	Late
F3D2	Early
F3D3	Early
F3D5	Early
F3D6	Early
F3D7	Early
F3D8	Early
F3D9	Early

Looking at Data!

```
cd /scratch/training/NGS_metagenomics
ls -l
cd Data
ls -l
cd MiSeq_SOP
ls -l
head –n 16 F3D1_S189_L001_R1_001.fastq
```

Running Mothur

- Loading modules and calling the program

```
module load VSEARCH/2.3.0-intel-2016a \
Mothur/1.38.1.1-intel-2016a-Python-2.7.11
mothur
```

- To run the preprocessing script

```
cd /scratch/training/NGS_metagenomics/Data/MiSeq_SOP
mothur preprocessing.batch
```

Login and Set up

- Login to Ada using SSH or MobaXterm
- Let’s take a look at the path and create appropriate directories

```
echo $SCRATCH
cd $SCRATCH
Pwd
mkdir NGS_assembly_Oct17
mkdir NGS_assembly_Oct17/Data
mkdir NGS_assembly_Oct17/Scripts
mkdir NGS_assembly_Oct17/Outputs
```

Preprocessing of the Data

```
cd /scratch/training/NGS_metagenomics/Outputs
ls -l
```

Processes done by pre-processing script:

- Making contigs for PE input data
- Mapping to reference
- Checking the alignment output
- Removing chimeras
- Assessing error rates

Analyzing Data

```
cd /scratch/training/NGS_metagenomics/Scripts
ls -l
```

Use either methods to look at the file and copy/paste the codes

```
cp /scratch/training/NGS_metagenomics/\
Scripts/Analysis_Commands.txt $SCRATCH
cd $SCRATCH
cat Analysis_Commands.txt
```

```
Scp username@ada.tamu.edu:path-to-script .
```

FigTree Visualization

- Login to Ada using SSH using “-X”
- Move to the correct directory (alternatively, you can add the path to your $PATH)

```
cd /scratch/training/NGS_metagenomics/FigTree/lib
java -Xms64m -Xmx512m -jar figtree.jar $*
```

- FigTree window will appear on your screen
- Use File → Open to open a tree file (.tre)

Any question?

nghaffari@tamu.edu