

Introduction to Metagenomics Analysis for High Throughput Sequencing Data

Noushin Ghaffari, PhD

Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research

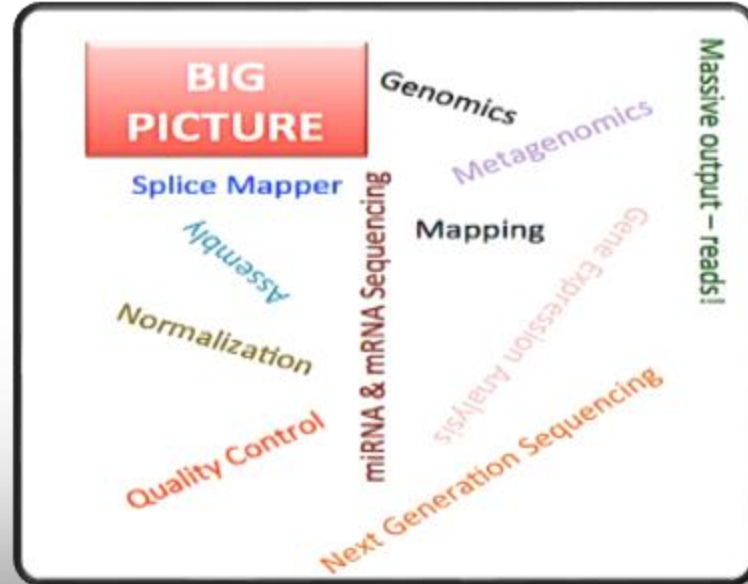
Research Scientist, Texas A&M High Performance Research Computing



DIVISION OF RESEARCH
TEXAS A & M UNIVERSITY

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 weeks {
- This class** →
- Next week →



Outline

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



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



Third Generation Sequencing Platforms

PacBio



Next Generation Sequencing Platforms

Illumina



© 2014 Illumina, Inc. All rights reserved.

MinION



NGS Sequencing Workflow

DNA/RNA extraction



Library creation/amplification



Sequencing (Illumina, PacBio, Oxford NanoPore)



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

Comparing Sequencing Platforms

Platform	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

Choosing among Illumina Sequencers

Metagenomics
16S rRNA

Metagenomics
WGS

MiniSeq

MiSeq

NextSeq

HiSeq 4000

HiSeq X Ten



MAX OUTPUT

8 Gb

MAX READ NUMBER

25 million

MAX READ LENGTH

2x150 bp

MAX OUTPUT

15 Gb

MAX READ NUMBER

25 million

MAX READ LENGTH

2x300 bp

MAX OUTPUT

120 Gb

MAX READ NUMBER

400 million

MAX READ LENGTH

2x150 bp

MAX OUTPUT

1500 Gb

MAX READ NUMBER

5 billion

MAX READ LENGTH

2x150 bp

MAX OUTPUT

1800 Gb

MAX READ NUMBER

6 billion

MAX READ LENGTH

2x150 bp

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

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

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 Techniques

1. Whole Genome Shotgun (WGS)

2. Marker Gene

- 16S Ribosomal RNA (rRNA)
 - Bacteria, Archaea
- 18S rRNA
 - Fungus, Eukaryotes

- Chen K and Pachter L, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities, *PLoS Comput Biol*, 1(2), 2005.



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

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

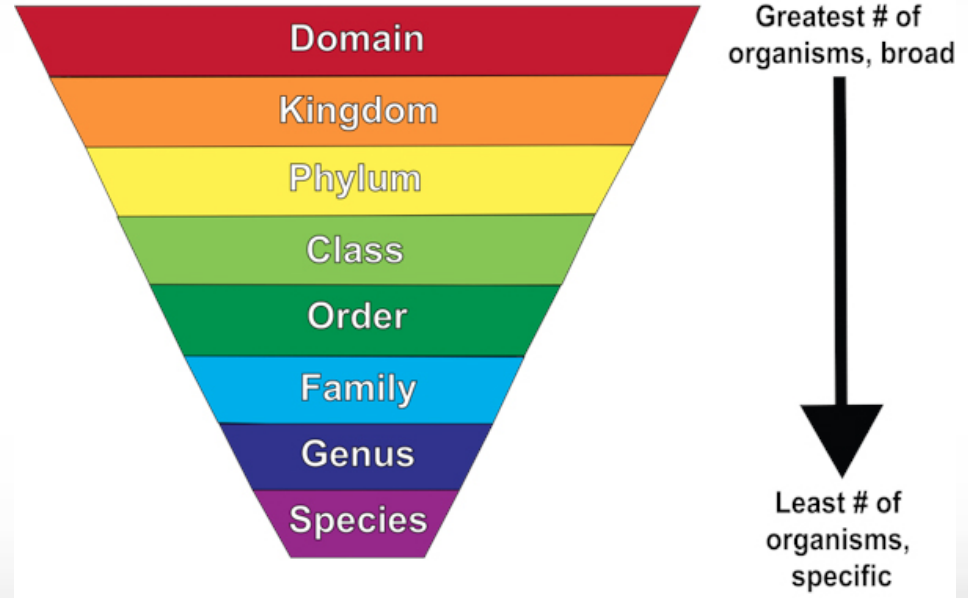
- MetaGeneAnnotator
- Glimmer-MG
- FragGeneScan
- MetaGeneMark
- Kraken

Marker Gene Metagenomics

- Usually based on 16S rRNA
 - 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

Metagenomics - Outcomes

- OUT clustering
- Taxonomic rank assignment
- Alpha diversity analysis – within sample diversity
- Beta diversity analysis – between sample diversity



<https://d2gne97vdumgn3.cloudfront.net/api/file/QooG1lg6RLGdDVli9oOg>

Marker Gene Metagenomics Tools

Microbial community analysis

- QIIME
- Mothur
- SILVAngs
- MG-RAST
- MEGAN

Diversity analysis

- Chao
- UniFrac
- PCoA

Visualization

- QIIME
- MEGAN
- FigTree

Metagenomics Web-Based Tool

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

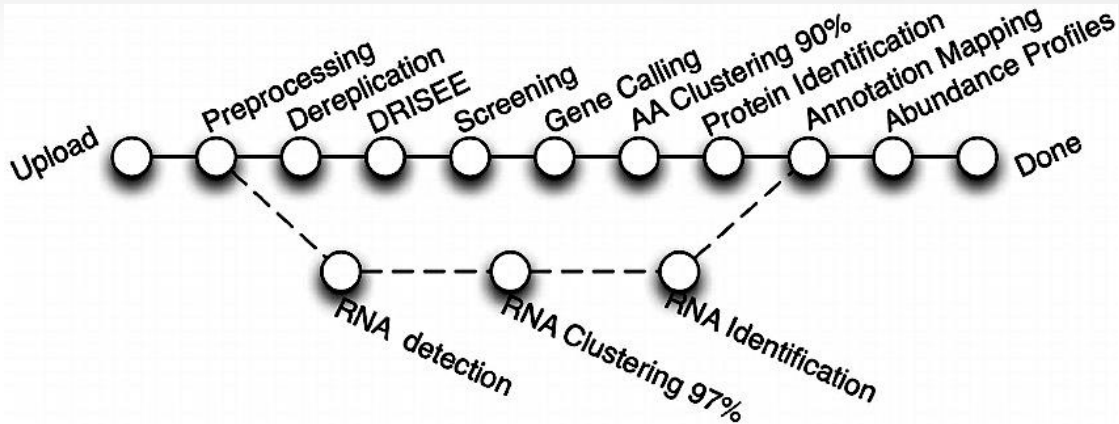
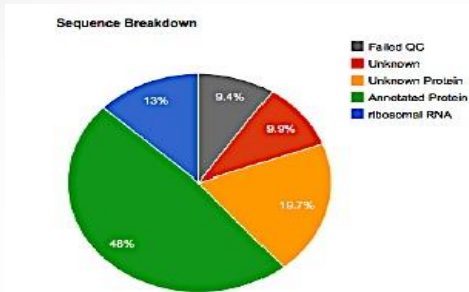
The screenshot shows the PATRIC (Pathosystems Resource Integration Center) website. The browser address bar displays 'www.patricrc.org/portal/portal/patric/RAST'. The main navigation bar includes 'ORGANISMS', 'SEARCHES & TOOLS', 'DOWNLOADS', and 'ABOUT'. A search bar is present at the top left. The 'SEARCHES & TOOLS' menu is expanded, showing options like 'Complete List of All Tools', 'Specialized Searches' (EC Search, GQ Search, Genome Finder, Feature Finder, BLAST, ID Mapping), 'Comparative Analyses' (Protein Protein Interactions, Protein Family Sorter, Genome Metadata, Comparative Pathway Tool), and 'Annotation Pipelines' (MG-RAST, RAST). Below the menu, there is a section for 'RAST Rapid Annotation using Subsystem Technology' with a brief description and a link to 'Info: To monitor RAST's load and view other news and statistics for RAST and the SEED...'. At the bottom, it states 'We have a number of presentations and tutorials available:' followed by a list of links: 'The RAST/SEED Workshop presentations', 'Registering for RAST', 'Downloading and installing the myRAST Toolkit', and 'The RAST batch submission interface (a part of myRAST)'.

The screenshot shows the EBI Metagenomics website. The browser address bar displays 'https://www.ebi.ac.uk/metagenomics/'. The main navigation bar includes 'Home', 'Search', 'Sequence search', 'Submit data', 'Projects', 'Samples', 'Comparison tool', 'About', and 'Help'. A search bar is present at the top right. The main content area features a large blue banner with the text 'Submit, analyse, visualize and compare your data.' and a 'SUBMIT DATA' button. Below the banner, there are statistics for data sets, assemblies, metatranscriptomes, runs, samples, and projects, categorized by Public and Private. A 'Browse projects' link is visible at the bottom.

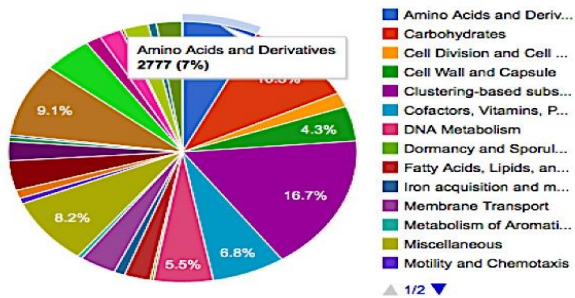
MG-RAST
metagenomics analysis server

MG-RAST Pipeline

19



Subsystems [Download chart data](#)
 has 42,515 predicted functions
 79.8% of predicted proteins
 104.4% of annotated proteins
[View Subsystems interactive chart](#)



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 4478543.3

Annotation Sources MSNR

Max. e-Value Cutoff 1e-5

Min. % Identity Cutoff 80 %

Min. Alignment Length Cutoff 15

Workbench use features from workbench

3 Data Visualization

bar chart tree table heatmap PCoA rarefaction

MG-RAST Example

20

Amplicon Based 16S Ribosomal RNA Sequencing and Genus Identification

*J. Risinger, *L. Renken, +J. Hill,
+N. Ghaffari, +R P. Metz, PhD,
+C. D. Johnson, *M. M. Toloue,
*Bioo Scientific

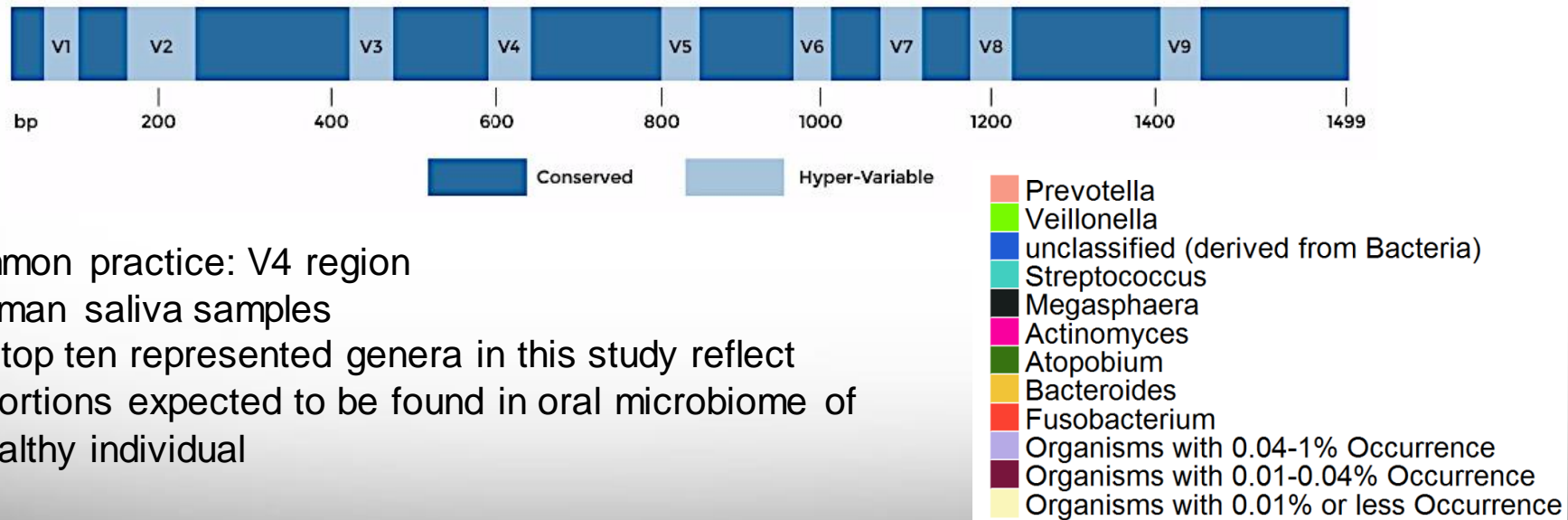
+AgriLife Genomics and Bioinformatics Service, Texas A&M University

Presented at PAG 2015



MG-RAST Example - 2

We demonstrate the utility of the NEXTflex™ 16S **V1-V3 Amplicon**-Seq Kit combined with the longer read chemistry of Illumina MiSeq (2x300) for enabling accurate identification of genera present in highly complex microbial communities across a vast number of samples



- Common practice: V4 region
- 7 human saliva samples
- The top ten represented genera in this study reflect proportions expected to be found in oral microbiome of a healthy individual

Metagenomics Tools - 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



Metagenomics Tools – Qiime 2

- Qiime: Quantitative Insights Into Microbial Ecology
- Open-source, community developed
- NGS microbial bioinformatics platform
 - Interactive visualizations and data exploration
 - Automatically tracks analysis
 - Facilitate easy sharing
 - Plug-in based
- Multiple interfaces
 - Command line interface: [q2cli](#)
 - Data scientist's interface: Artifact API
 - the graphical user interface: [q2studio](#) (PROTOTYPE)
- Artifact: contain data and metadata



Qiime2 - Continued

Input Data

- SE or PE FastQ files, multiplexed or demultiplexed
 - Name of the input files should have a specific format: Sample1_Barcode1_L001_R1_001.fastq.gz
 - sample identifier_barcode sequence/barcode identifier_the lane number_read number_set number.fastq(.gz)
- FastQ Manifest
 - CSV manifest file, columns are
 - Sample ID, file-path, direction of sequencing (forward/reverse)
- Feature Table Data
 - BIOM format, based on [HDF5](#)
 - Id, type, format-url, format-version, generated-by, creation-date, shape, nnz (non-zero elements)
- Per-feature unaligned sequence data
- Phylogenetic trees (unrooted)

Qiime2 - Continued

Meta Data

- Tab-separated text (TSV) file
 - Example:
<https://docs.google.com/spreadsheets/d/1bHXutGx07HnYUGE1O4IFn9yltt6BEXQEzn276xqPid0/edit#gid=0>

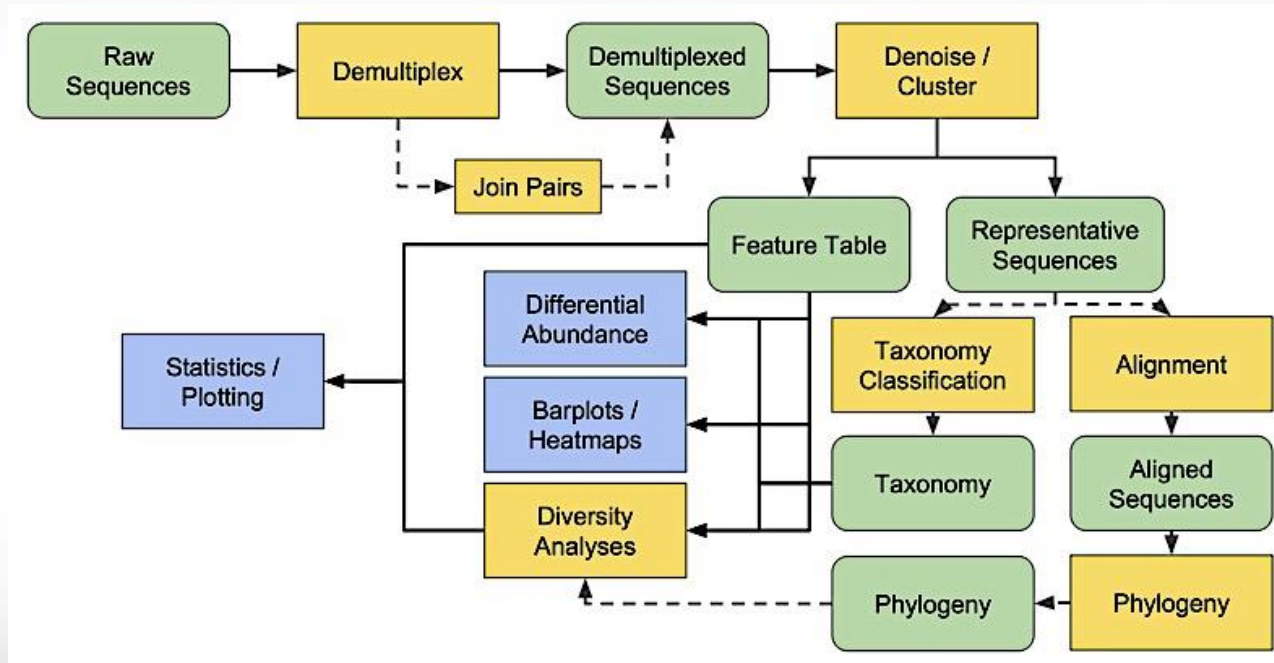
Artifacts

- Name.qza, zipped archives files containing data and its related files
- Name.qzv, visualization files
 - Visualize online at: <https://view.qiime2.org/>
 - Command: "qiime tools view file.qzv" on HPRC portal > VNC session by logging into: portal.hprc.tamu.edu

Classification

- Naive Bayes classifier can be trained based on sequenced data or can be downloaded based pre-trained from Qiime2 “Data resources”: <https://docs.qiime2.org/2018.8/data-resources/>

Qiime2 - Continued



- Sample QC: DADA2 R Package. Only SE, thus, ran for R1 and R2 files separately and then results are merged.

<https://docs.qiime2.org/2018.8/tutorials/overview/>

Qiime 2 - Practice

- <https://docs.qiime2.org/2018.2/tutorials/>
- **Practical portion** (based on different tutorials)
 - Data: Fecal microbiota transplant (FMT)
 - Children under the age of 18 with autism and gastrointestinal disorders
 - Treated with fecal microbiota transplant in attempt to reduce the severity of their behavioral and gastrointestinal symptoms
 - collection of weekly fecal swab samples
 - stool samples
 - 18 treated individuals, 20 control
 - Subset data for exercise: **5 treated, 5 control**: Between six and sixteen samples are included per individual, including stool and fecal swab samples for each individual, and samples before and after FMT treatment. Five samples of the transplanted fecal material are also included.
 - 2 Illumina MiSeq sequencing runs

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)

The screenshot shows the MobaXterm application interface. On the left, a file explorer window displays the directory structure of the user's home directory. The main terminal window shows the following output:

```
whomps@login5~
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Settings Help
Quick connect...
/generalhome/whomps/
Name Size (KB) Last Modified
..
..env_fea2.015.0_cache 4 2015
..env_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
.jmod.d 4 2016
.local 4 2015
.lsbatch 4 2017
.matlab 4 2016
.mozilla 4 2015
.mw 4 2016
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net
```

```
=====
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 @_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 ~]$
```

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:
```

