The Elizabeth H. and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Washington University in St. Louis
School of Medicine

PMBIO Module 01

Setup. Cloud computing, the command line, and tool installation

Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia

Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)

29 October – 2 November, 2018

Glasgow

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

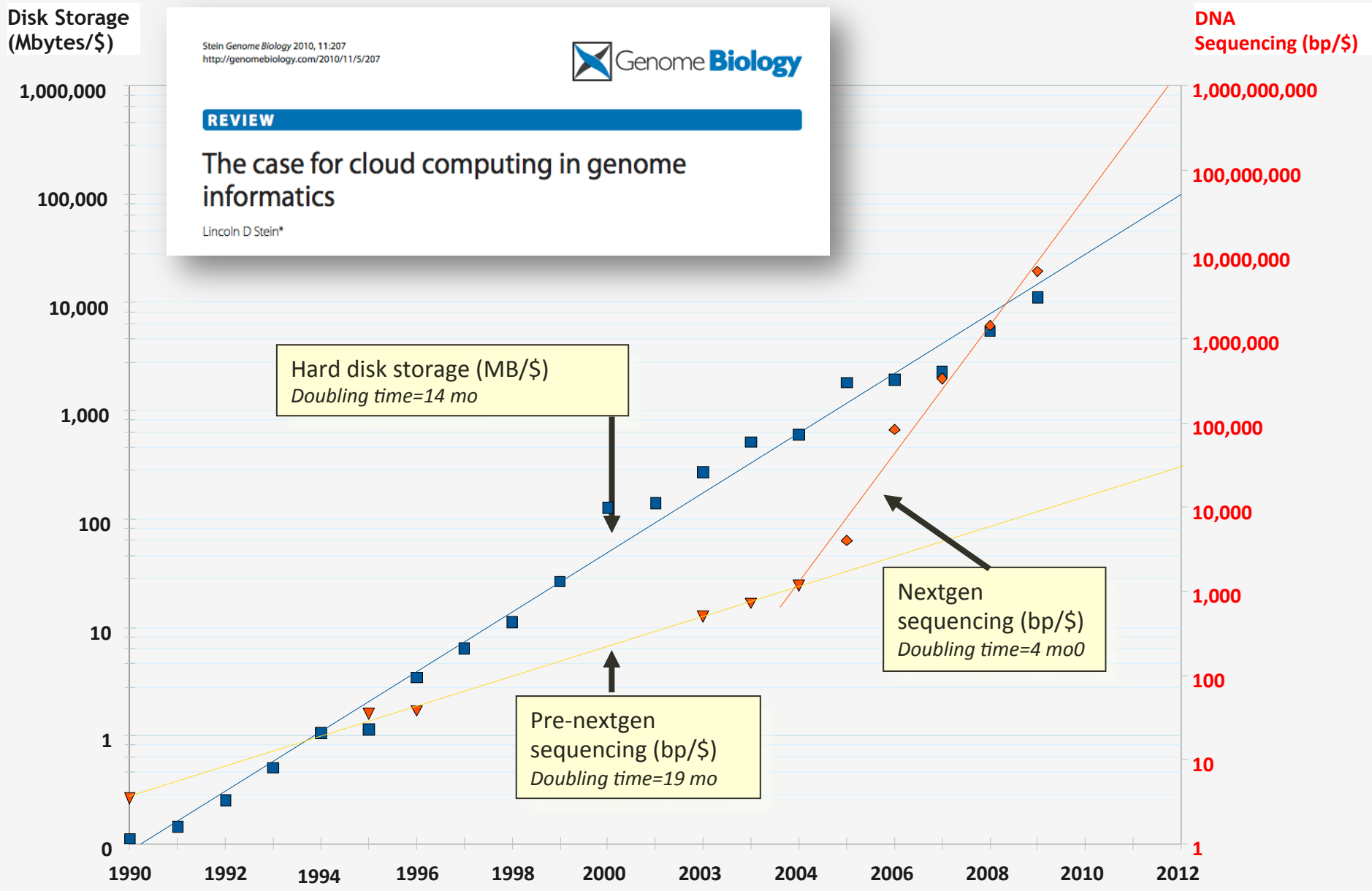
Learning objectives of module 01: Setup

- **Key concepts**: Cloud computing, UNIX/Linux OS, bioinformatics tool installation

- Start a cloud compute instance and log into it
- Become familiar with basic Linux terminal commands (the "command line")
- Install bioinformatics tools that will be used for genomics analysis

- Navigation of the course website and other online learning resources
  - www.pmbio.org
  - Other resources: www.genviz.org, www.rnabio.org, www.biostars.org

Disk Capacity vs Sequencing Capacity, 1990-2012



Stein *Genome Biology* 2010, 11:207
http://genomebiology.com/2010/11/5/207

Genome Biology

REVIEW

The case for cloud computing in genome informatics

Lincoln D Stein*

About DNA and computers

- We hit the $1000 genome* in ~2016
  - Need to think about the $100 genome
- The doubling time of sequencing has been ~5-6 months.
- The doubling time of storage and network bandwidth is ~12 months.
- The doubling time of CPU speed is ~18 months.
- The cost of sequencing a base pair will eventually equal the cost of storing a base pair

What is the general biomedical scientist to do?

- Lots of data
- Poor IT infrastructure in many labs
- Where do they go?
- Write more grants?
- Get bigger hardware?

Cloud computing providers

- Amazon AWS
  - https://aws.amazon.com/
- Google cloud
  - https://cloud.google.com/
- Digital ocean
  - https://www.digitalocean.com/
- Microsoft Azure
  - https://azure.microsoft.com/en-us/
- More…

Amazon Web Services (AWS)

- Infinite storage (scalable): S3 (simple storage service)
- Compute per hour: EC2 (elastic cloud computing)
- Ready when you are High Performance Computing
- Multiple football fields of HPC throughout the world
- HPC are expanded at one container at a time:

amazon web services™

Some of the challenges of cloud computing:

- Not necessarily cheap!
- Getting files to and from there
- Not the best solution for everybody
- Standardization
- PHI: personal health information & security concerns
- In the USA: HIPAA act, PSQIA act, HITECH act, Patriot act, CLIA and CAP programs, etc.
  - http://www.biostars.org/p/70204/

Some of the advantages of cloud computing:

- We received a grant from Amazon, so supported by ‘AWS in Education grant award’.
- There are better ways of transferring large files, and now AWS makes it free to upload files.
- A number of datasets exist on AWS (e.g. 1000 genome data).
- Many useful bioinformatics AMI’s (Amazon Machine Images) exist on AWS: e.g. cloudbiolinux & CloudMan (Galaxy) – now one for this course!
- Many flavors of cloud available, not just AWS

Key AWS concepts and terminology

- AWS - Amazon Web Services. A collection of cloud computing services provided by Amazon.
- EC2 - Elastic Compute. An AWS service that allows you to configure and rent computers to meet your compute needs on an as needed basis.
- EBS - Elastic Block Storage. A data storage solution that allows you to rent disk storage and associate that storage with your compute resources. EBS volumes are generally backed by SSD devices.

Key AWS concepts and terminology

- S3 - Simple storage service. Cheaper than EBS and allows for storage of larger amounts of data with some drawbacks compared to EBS. S3 volumes store data as objects that are accessed by an API or command line interface or other application designed to work with S3. EBS volumes on the other hand can be mounted as if they were a local disk drive associated with the Instance.
- SSD - Solid state drive. A particular type of storage hardware that is generally faster and more expensive than traditional hard drives.

Key AWS concepts and terminology

- HDD - Hard disk drive. A particular type of storage hardware that is generally cheaper and larger but slower than SSD. HDD drives are traditional hard drives that access data on a spinning magnetic disk.
- Ephemeral storage - Also known as Instance Store storage. Data storage associated with an EC2 instance that is local to the host computer. This storage does not persist when the instance is stopped or terminated. In other words, anything you store in this way will be lost if the system is stopped or terminated. Instance store volumes may be backed by SSD or HDD devices.

What is a Region?

- An AWS Region is set of compute resources that Amazon maintains (like the Data Center image shown before)
- Each Region corresponds to a physical warehouse of compute hardware (computers, storage, networking, etc.).
- At the time of writing there are 14 regions: (US East (N.Virginia), US East (Ohio), US West (Oregon), US West (N. California), Canada (Central), EU (Ireland), EU (Frankfurt), EU (London), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Seoul), Asia Pacific (Tokyo), Asia Pacific (Mumbai) and South America (Sao Paulo).
- When you are logged into the AWS EC2 console you are always operating in one of these regions.

What is a Region?

- Current region shown in the upper right corner of console
- It is important to pay attention to what region you are using for several reasons.
  - When you create an EC2 instance (EBS volume, etc) in one region you won't see it in another region.
  - The cost to use many AWS resources varies by region.
  - The region may influence network performance when you are accessing the instance, especially if you need to transfer large amounts of data in or out.
  - Billing is tracked separately for each region
  - Generally you should choose a region that is close to you or your users. But cost is also a consideration.

What is difference between the 'Start', 'Stop', 'Reboot', and 'Terminate' (Instance States)?

- Start – turn on an EC2 instance that you have previously created
- Stop – turn off an EC2 instance that you have previously created
- Reboot – restart an EC2 instance
- Terminate – permanently stop and destroy an EC2 instance. Any associated EBS volumes may also be destroyed at this time depending on configuration

What is an AMI/snapshot?

- AMI (Amazon Machine Image) – a template that specifies how to launch EC2 instances
  - Root volume with operating system (OS), pre-installed applications, etc
  - Launch permissions the determine who can use the AMI
  - Specification of (data) volumes to attach when launched
- You can create an AMI for any instance you have created/ configured
- AMI can be made public for sharing (region-specific)
- Creating an AMI involves creating a snapshot of the root and any attached volumes. You will be charged to store this snapshot.

I can not log into my EC2 instance, what might have gone wrong?

- Is your instance running?
- Are you providing the correct path to your key file?
- Is it the correct key file?
- Have you set the permissions for your key file correctly?
- Did you specify a valid user for your AMI (e.g., ubuntu)?
- Did you specify the correct IP address?
- Does the Security Group for the instance allow access for your connection protocol (e.g., SSH) and location?

How much does it cost to use AWS EC2 resources?

Linux | RHEL | SLES | Windows | Windows with SQL Standard | Windows with SQL Web | Windows with SQL Enterprise

Region: US West (Oregon)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.nano	1	Variable	0.5	EBS Only	$0.0058 per Hour
t2.micro	1	Variable	1	EBS Only	$0.0116 per Hour
t2.small	1	Variable	2	EBS Only	$0.023 per Hour
t2.medium	2	Variable	4	EBS Only	$0.0464 per Hour
t2.large	2	Variable	8	EBS Only	$0.0928 per Hour
t2.xlarge	4	Variable	16	EBS Only	$0.1856 per Hour
t2.2xlarge	8	Variable	32	EBS Only	$0.3712 per Hour
m4.large	2	6.5	8	EBS Only	$0.1 per Hour
m4.xlarge	4	13	16	EBS Only	$0.2 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.4 per Hour

Data transfer (GB): In: free or $0.01; Out: free, $0.01 or $0.02
EBS storage (GB/Month): $0.10
S3 storage (GB/Month): $0.023 standard, $0.0125 infrequent access, or $0.004 glacier

Why am I still getting a monthly bill?

- Generally you get an accounting of usage and cost on a 30 day cycle
  - Pricing is per instance-hour (now instance-second!) consumed for each instance type.
  - Also charges for storage, transfers, etc
- Be aware of regions!
- Even when an instance is stopped, storage for root or other EBS volumes persist
- Creating AMIs/snapshots requires storage
- Explore the billing and cost management tools of AWS to track your spending, set warnings, etc

Amazon AWS documentation

http://pmbio.org/module-01-setup/0001/02/01/AWS_Intro/

http://aws.amazon.com/console/

In this workshop:

- Some tools (data) are
  - on your computer
  - on the web
  - on the cloud.
- You will become efficient at traversing these various spaces, and finding resources you need, and using what is best for you.
- There are different ways of using the cloud:
  1. Command line (like your own very powerful Unix box)
  2. With a web-browser (e.g. Galaxy): not in this workshop

Things we have set up:

- Loaded data files to a web server
- We brought up an Ubuntu (Linux) instance, and loaded a whole bunch of software for NGS analysis.
- We will clone this and create separate instances for everybody in the class.
- We've simplified the security: you basically all have the same login and file access, and opened ports. In your own world you would be more secure.

Demonstration of the AWS EC2 console

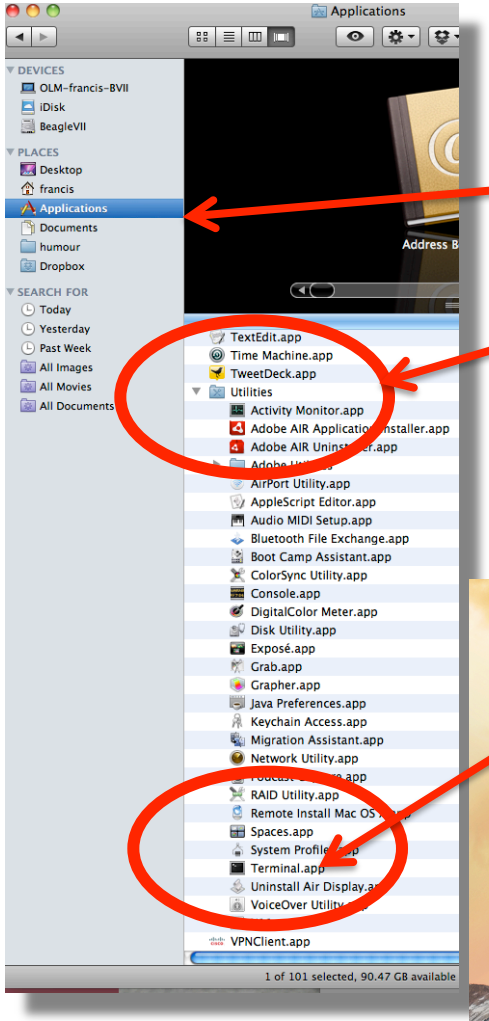
- Walk through the process by which each student instance was created

connect.wustl.edu/awsconsole

Logging into Amazon AWS

(each student will be assigned their own cloud compute instance and unique number)

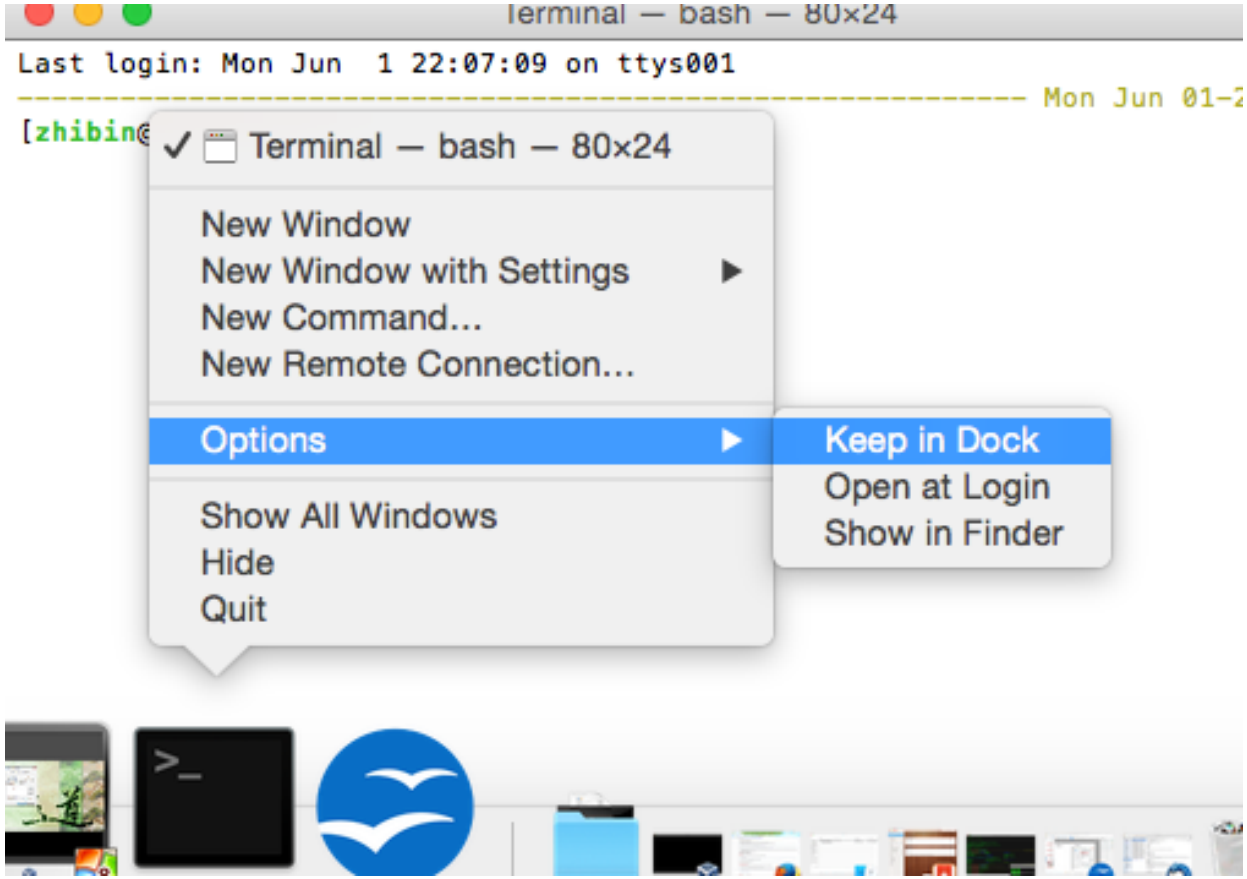
Opening a ‘terminal session’ on a Mac



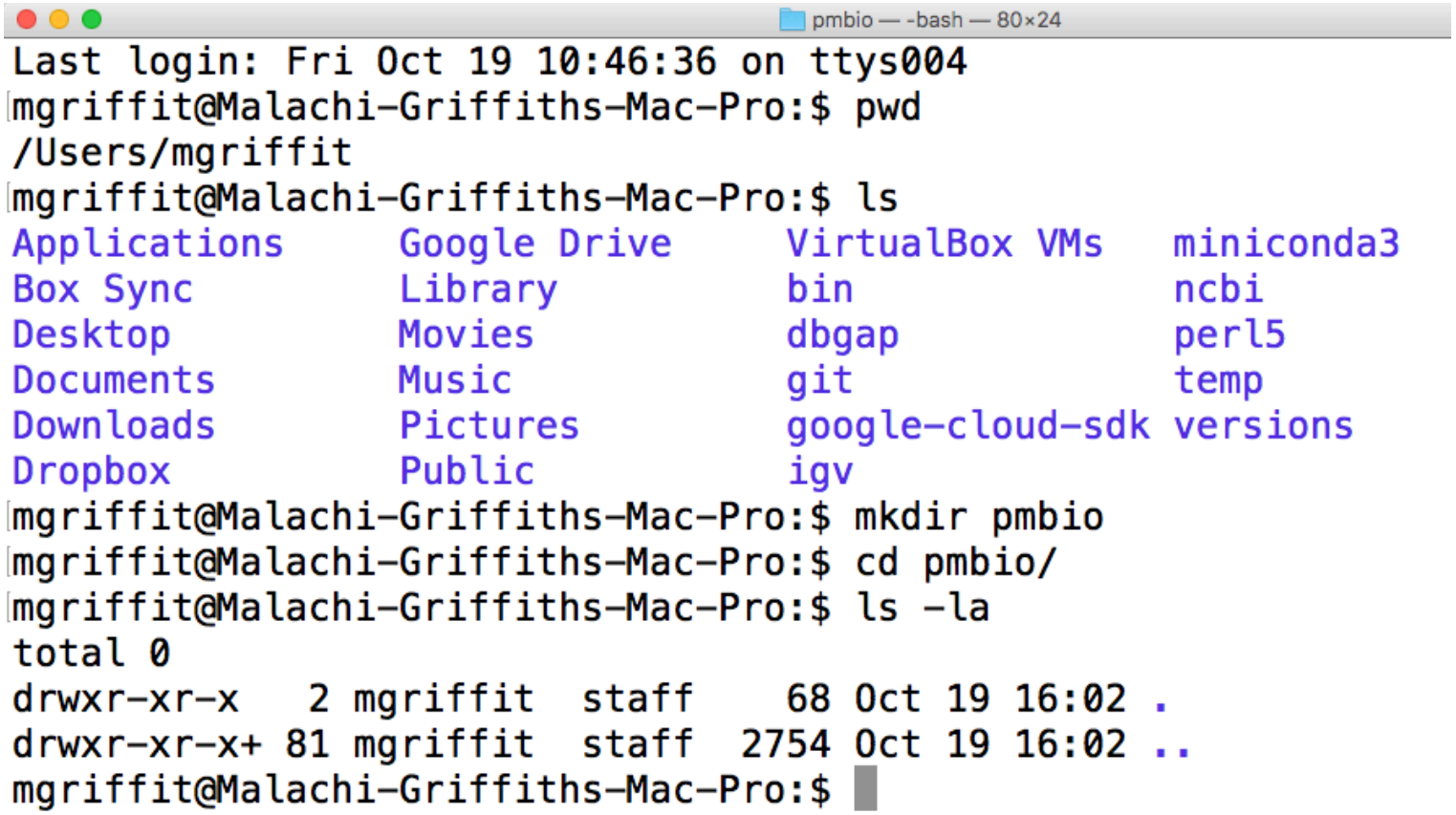
In a Finder window
‘Applications’ -> ‘Utilities’ -> ‘Terminal’

Or on your dock

Add the terminal App to your dock



Creating a working directory on your laptop called ‘pmbio’



mkdir pmbio
cd pmbio

Obtain the course SSH key file

- NOTE for Mac users. You will need to use a ".pem" file
- **NOTE for Windows Users.** You will need to use a ".ppk" file instead.
  - This is created from the ".pem" file.
  - https://aws.amazon.com/premiumsupport/knowledge-center/convert-pem-file-into-ppk/

- Both of these key files will be emailed to you
  - "pmbio.pem" and "pmbio.ppk"

- The SSH key file will be used to securely login to your student instance on the cloud

Viewing the ‘key’ file once downloaded

cat pmbio.pem

```
obis-air:cshl ogriffit$ cd ~/cshl/
obis-air:cshl ogriffit$ ls -la
total 8
drwxr-xr-x   3 ogriffit  staff   102 Nov 13 22:21 .
drwxr-xr-x+ 58 ogriffit  staff  1972 Nov 13 22:18
-rw-r-----@  1 ogriffit  staff  1696 Nov 13 22:21 pmbio.pem
obis-air:cshl ogriffit$ cat pmbio.pem
-----BEGIN RSA PRIVATE KEY-----
MIIEpgIBAAKCAQEAvJ5gwmtby9QZ2Idz+ugiEQQHW6PsOZAZFvr+mWDnM4pKpccaVmDh7XjcEOLF
OkJzaP9+jj0kSFOyNinitoB32DgrmVhgNhyheEqH5XMn28szxUj1EuoNXAogNuY7mWMo6MoWssSW
Rqy+rj19vMGQn5rsnMLjCM1smebPoqYOL8EPa1ccRbdGXG1dMTlCC1ho/Hk9bZweamGiZLaAWVmf
zOK/L0zxgY3K4cwaL48HV6oGuMh5lTDpnobxXghQ4oC5Mej+DpCRF8C+EG2uNDuyuLzRJfQmFBV2
GKDWDwhdgGmKmX9IpMT9ubvNoQPy0vYLvM80eG3cMbz2IZpaNryihwIDAQABAoIBAQCZYT0TvFO4
a3DdCEEC/rN9HMaS+bjFkm0kp9RTi15XJhTPvBmptjzibA6gWJfDaXgKIQGbzxJrEkxwCR2IBO3v
0LV7jEcomZ2ggRMDPeJitFoUCuDnkZZtivppSk2azOzeaD+O/ZeqPx0L+Yr+7HSbpVLVoxEV/l5a
xDuCawBMSY2cnGWKfEBlSPnB6fGZj8luGzv0aP/CETx/K78TIS56m4yrTIQIeEPfFt/PQr/EUqoL
7co5oy9K3sD1noPLDhk3vJa1VNrMjHkMZLkbZuaoHPzgSQHninm80Ca25WWTGsSZ8vQsBIUTlGI1
W7lzXH3wD1jJNd+03QK4bnKaZ+DZAoGBAPVpisa49JY/6K2f9B8naqtX/ljzVWTl3Q7r6t6uh21Y
oexmC8eJ2wQwd0qNjZWVxSMVksIwdM6xcsBIJRMmltWTVdmD0fkDvOfjd8CM4nctH76tvSvZz02e
qI9wSshHY1fh+09CoLZeefFSURxqWbkJfREjoZ4UGUWMi3k1rxC9AoGBAMTBlBBOWQ+5ojzQYu0L
Q4YrsIPg1/ni0WmJ+05vcTCJ2aeI88VhK5c2PoXPWWiJ9CdD2VFZDiCm2XuJA5iwJmnhuwGGHHEn
BFBqEF/ueJrW+r43pRcYRuRIXjiH4mQQlK4Zemecym5fAHvxZxq4fs2kWfMPySFaVufcP0VC7X6T
AoGBAMhro0xbrFQwaUOyh9oRhMneGPhn8WtvVjNjc/LcMfmZEtRPGnuhF965/hJCvEhXgiH+8lXo
4NwUixSBVtXnA/P0WX5Ea2ykIth2Kkx0Qlb14SEGHqH7RZOsaRiLqmcZ9gXFpkm6rimByrDMezVr
nU7CcwNWSB0jaOgluZoJv6k5AoGBAJJuFsmD5ZhkaS+lTtpnlZtXDIk5XsMkYQGQpS0clzqufQPI
UtPEm3Jv9lwTktDQSpqmTifShUcbpaPgtoJ+JjiKvGhH7QbxKK7IIOOkULG76OSD+SOU972Rdj3Q
M1aRWHWxlH1kHOvDXFLhuAAU6poVBLR2PRPLbf4k1hmv05xtAoGBAJVQy1GF8uVNwkOCNzLIqmkY
uk9M24hfqn3N2GY3Zgqf43bD4kdYgL4rvsgpO8QzotPf+19kVlCv0ciolSjEHLyUdlyPGzj4CTTH
1f1RoGHmYzVn9VuFTu4hJ17J+uwgXgIr9Sx/UTjwkmCjPf7CEyIuGxaThG/ZoR9stufZB5db
-----END RSA PRIVATE KEY-----obis-air:cshl ogriffit$
```

Changing file permissions of your ‘key’ file (Mac/Linux)

ls -l (long listing)

drwx------+ 67 ogriffit staff 2278 22 May 21:25 ../

-rw-r--r--@ 1 ogriffit staff 1696 22 May 21:31 pmbio.pem

rwx : owner

rwx : group

rwx: world

r read (4)

w write (2)

x execute (1)

Which ever way you add these 3 numbers, you know which integers were used (6 is always 4+2, 5 is 4+1, 4 is by itself, 0 is none of them etc …)

So, when you have:

chmod 600 <file name>

It is “r” for the the file owner **only**

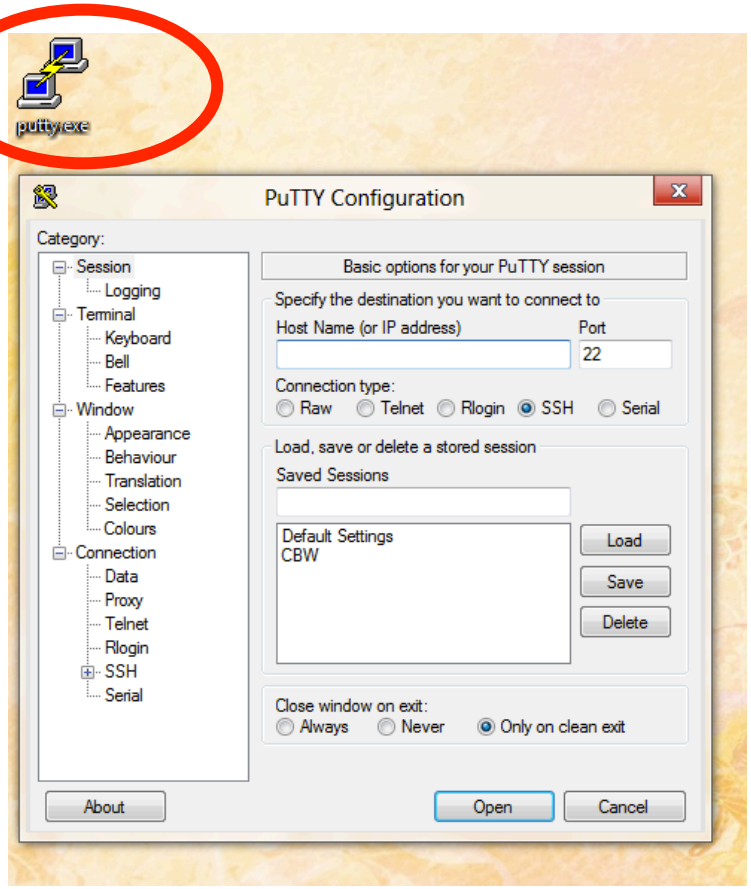
Logging into your instance (Mac/Linux)

Mac/Linux

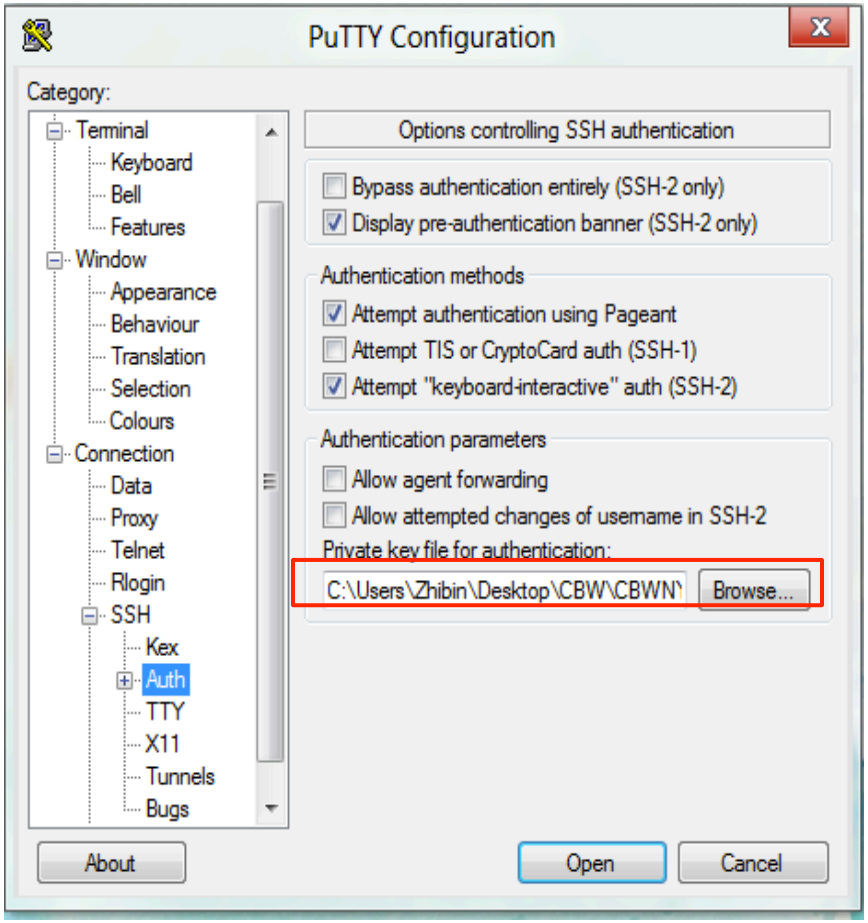
```
cd ~/pmbio
chmod 600 pmbio.pem
ssh −i pmbio.pem ubuntu@[YOUR PUBLIC IP]
```

Logging into your instance (Windows)

Open PuTTY

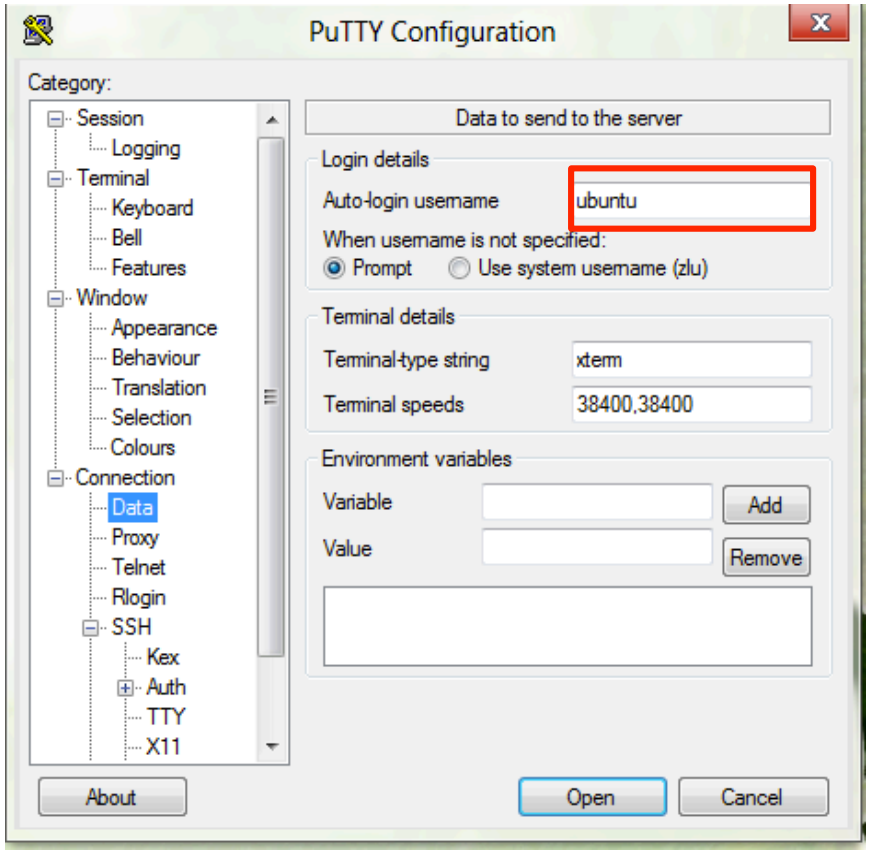


Browse to the pmbio.ppk file

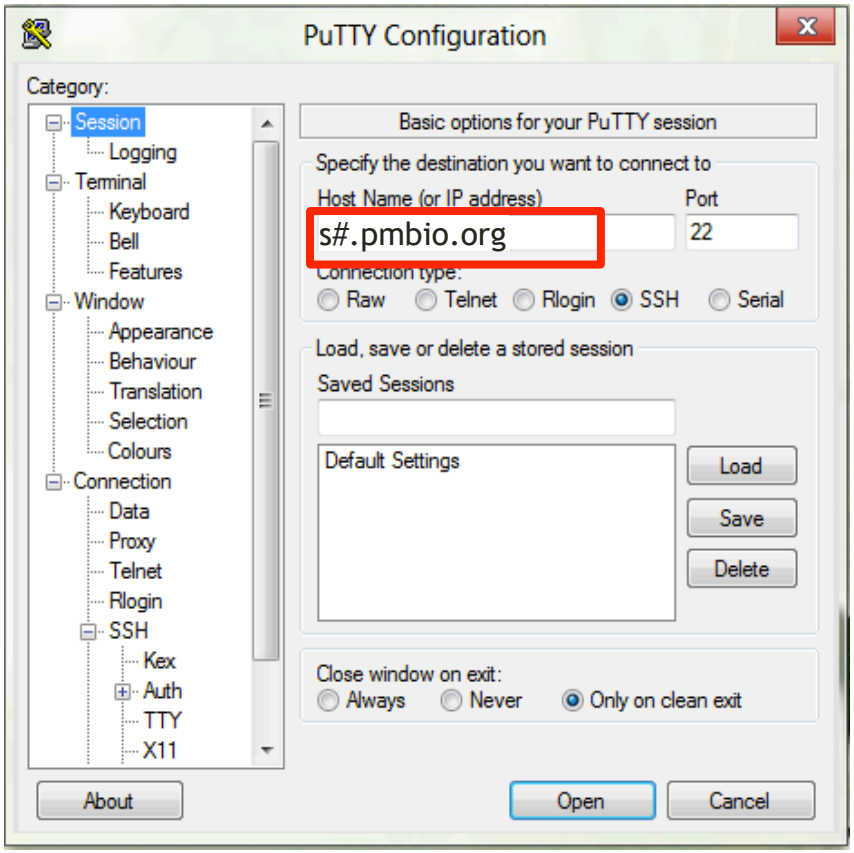


Logging into your instance (Windows)

Enter the user name ‘ubuntu’

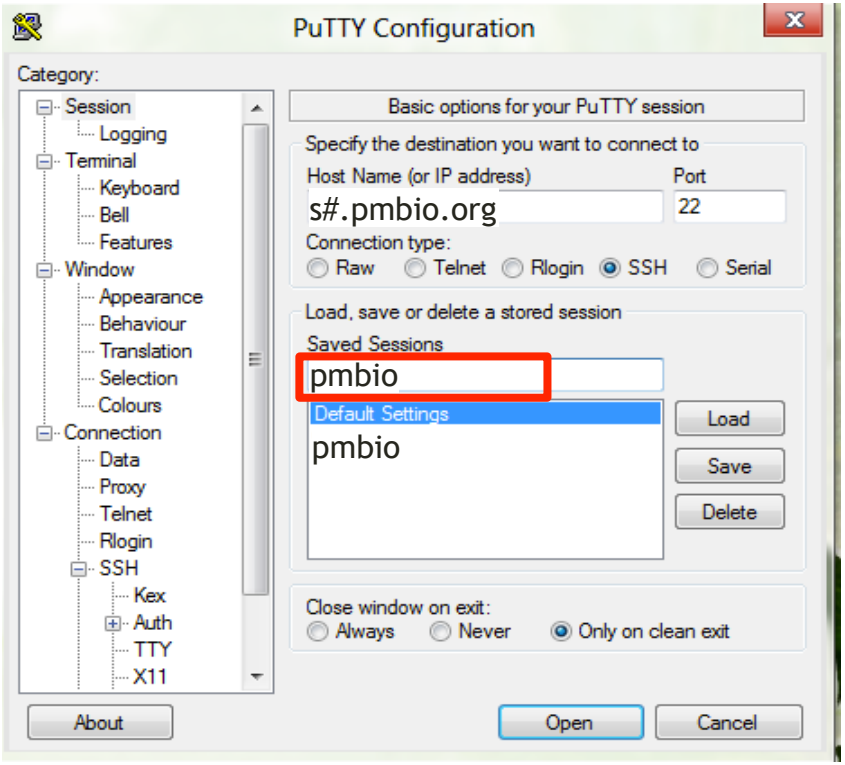


Enter the host name
(where # is your unique number)

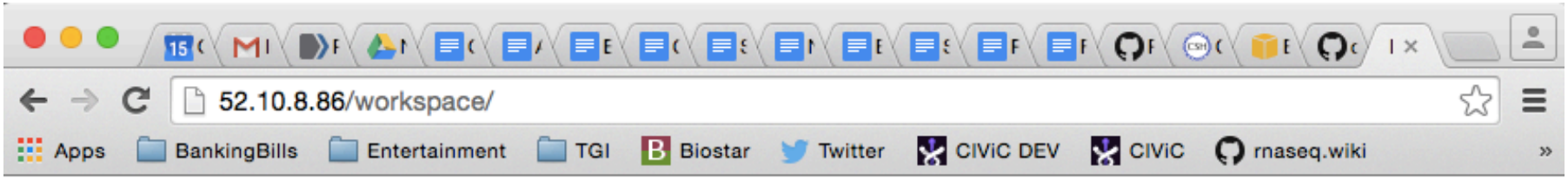


Logging into your instance (Windows)

Open PuTTY



Copying files from AWS to your computer (using a web browser)



Index of /workspace

Name	Last modified	Size	Description
Parent Directory		-	
Homo_sapiens/	2015-11-13 06:45	-	
README.txt	2014-06-17 23:53	5.3K	
bam-demo/	2015-11-14 21:03	-	
data/	2015-11-13 01:39	-	
scratch/	2015-11-13 19:43	-	
tools/	2015-11-13 01:54	-	

Apache/2.4.7 (Ubuntu) Server at 52.10.8.86 Port 80

http://[YOUR PUBLIC DNS OR IP]/

Logging out of your instance

Mac/Linux – simply type exit

```
exit
```

Note, this disconnects the terminal session (ssh connection) to your cloud instance. But, your cloud instance is still running!

So, at this point:

- Your laptop is ready for the workshop
- If it is not, you know where to get the information you need
- You know how to login to AWS

- The next step is to login to your linux machine on AWS and learn the basics of a linux command line

Introduction to linux

- Become familiar with basic Linux terminal commands (the “command line”)

- One of the first sections in the course is a short crash course on command line linux

Installation of bioinformatics tools

- Install bioinformatics tools that will be used for genomics analysis

- We will use many broadly useful bioinformatics tools in this course
  - Installation of most of these tools will be performed by the students
  - e.g. samtools, picard, GATK, BWA, VEP, etc.

The most common problems encountered while working on the tutorials

- Type short commands carefully if you like, but in order to get through all the steps smoothly, it is safer to copy and paste long commands
- Copy/Paste errors
  - Learn the short cuts for copying/pasting on your system and use them (e.g. <command><c> & <command><v> on Mac)
  - Make sure you copy the entire command. Watch out for commands that span across multiple lines
- Being in the wrong directory at the wrong time
  - The simplest way to avoid this is only change directories as instructed
  - If you do change directories to look around, make sure you go back before continuing with commands

Course website

www.pmbio.org