Lecture 6: Parallel computing, cloud computing and working on Amazon Web Services

Greg Caporaso

gregcaporaso@gmail.com

Some last thoughts on regular expressions

Robust searches

- Sometimes your queries will fail
  - Won’t produce output (good)
  - Will produce incorrect output (bad)
- Fail loudly! Produce a (useful) error message on failure.

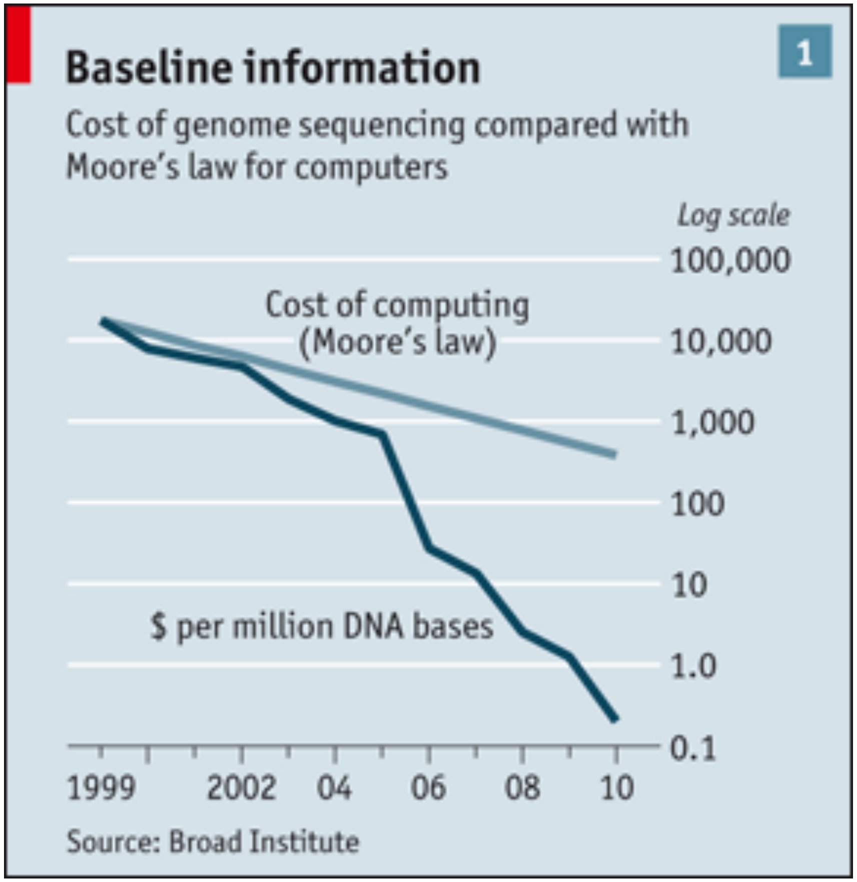
Designing robust searches

- Make assumptions explicit
  - If you're assuming that your records start with '>', search for '^>' to avoid matching '>' characters that show up in other places
- Match full lines by including ^ and $ in your search query
- Check the number of matches that were made: is it reasonable?

Testing of software

- Start thinking about what positive and negative controls for these terms might look like. Software testing is something we'll be discussing regularly through-out the semester.

Why is parallel computing important in bioinformatics?



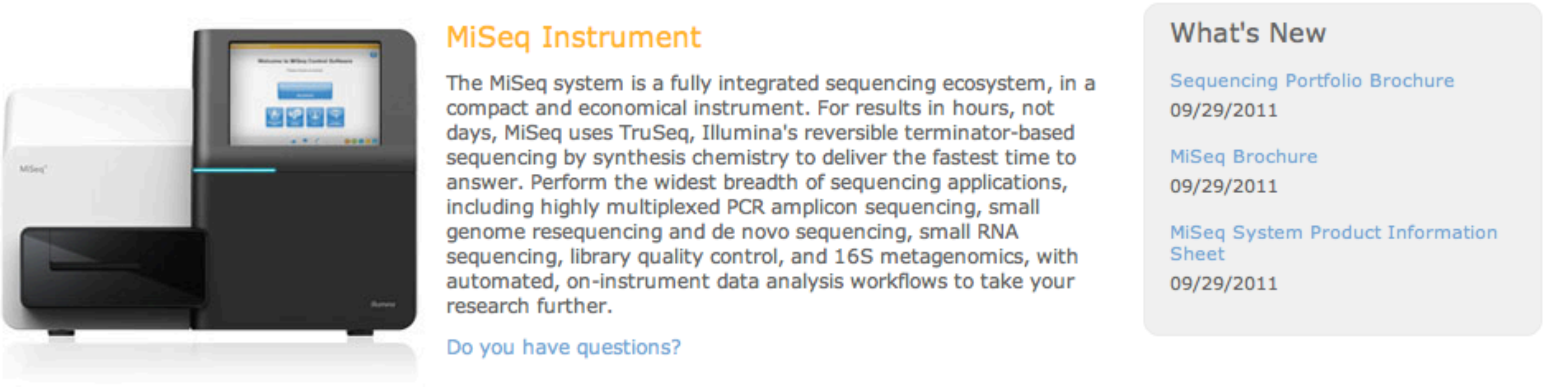
Cluster computing

- Many computers connected to one another to serve as a larger compute resource.
- Compute-intensive jobs can be split over many systems and run in parallel.
- Similar to desktop compute hardware, but different casing, no (or only few) displays/ keyboards directly connected.
- Owned and maintained “in-house”.

Why is parallel computing important in bioinformatics?

Platform	Sanger	454 (Titanium)	Illumina Genome Analyzer II	Illumina HiSeq2000	Illumina MiSeq
Read Length (bases)	~1000	~400	150 (single end)	100 (single end)	150 (single end); 250 soon
Number of reads	96 or 384	~1,000,000	~100,000,000	~1,600,000,000	~10,000,000
Maximum number of samples per run	n/a	1000	12,000 (barcode-limited)	24,000 (barcode-limited)	2500 (barcode-limited)
Sequences per $1 (sequencing costs only)	0.44	100	5000	200,000	12,500

The “benchtop” sequencer



MiSeq Instrument

The MiSeq system is a fully integrated sequencing ecosystem, in a compact and economical instrument. For results in hours, not days, MiSeq uses TruSeq, Illumina's reversible terminator-based sequencing by synthesis chemistry to deliver the fastest time to answer. Perform the widest breadth of sequencing applications, including highly multiplexed PCR amplicon sequencing, small genome resequencing and de novo sequencing, small RNA sequencing, library quality control, and 16S metagenomics, with automated, on-instrument data analysis workflows to take your research further.

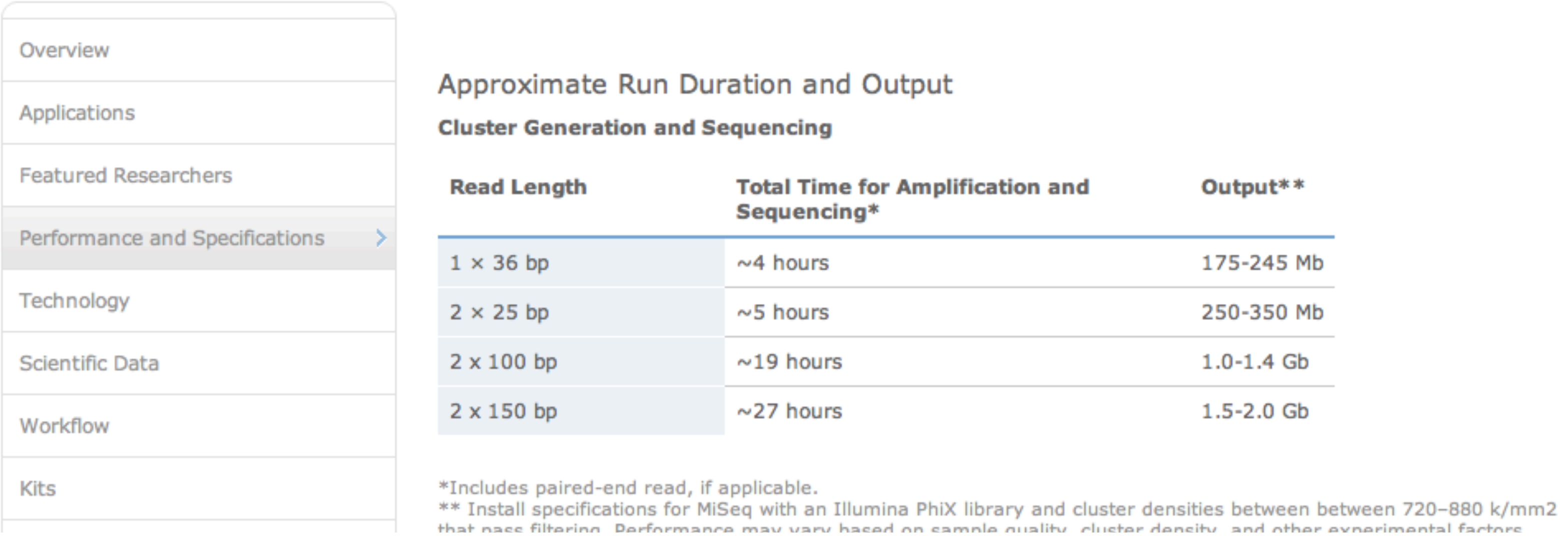
Do you have questions?

What's New

Sequencing Portfolio Brochure
09/29/2011

MiSeq Brochure
09/29/2011

MiSeq System Product Information Sheet
09/29/2011



Approximate Run Duration and Output

Cluster Generation and Sequencing

Read Length	Total Time for Amplification and Sequencing*	Output**
1 × 36 bp	~4 hours	175-245 Mb
2 × 25 bp	~5 hours	250-350 Mb
2 x 100 bp	~19 hours	1.0-1.4 Gb
2 x 150 bp	~27 hours	1.5-2.0 Gb

*Includes paired-end read, if applicable.
** Install specifications for MiSeq with an Illumina PhiX library and cluster densities between between 720–880 k/mm2 that pass filtering. Performance may vary based on sample quality, cluster density, and other experimental factors.

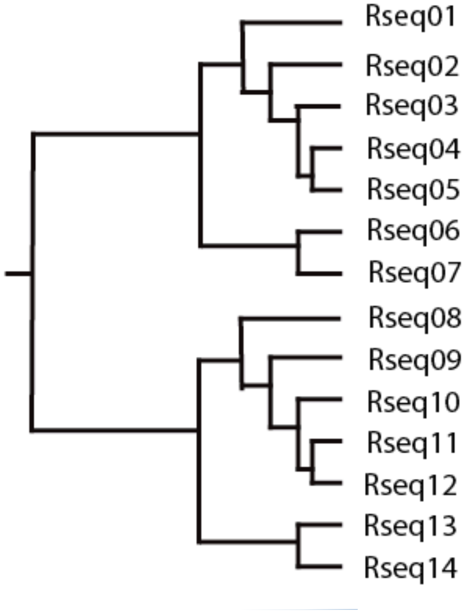
OTU picking: example of a compute intensive process

What taxa are represented in each sample?

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
...
```

BLAST against reference tree

Reference tree of non-redundant full length sequences



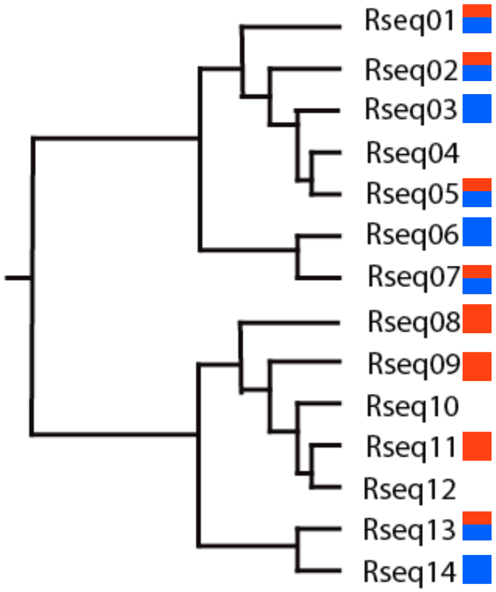
OTU picking: example of a compute intensive process

What taxa are represented in each sample?

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
...
```

BLAST against reference tree

Reference tree of non-redundant full length sequences



Clusters of "Operational Taxonomic Units" (OTUs);
Per sample hits on reference tree;
Taxonomic assignments

OTU Picking

- For 1 billion sequence reads, the initial step ran for ~116 hours on 110 processors requiring 4GB of RAM per job for workers and 64GB of RAM for master.

OTU Picking

- For 1 billion sequence reads, the initial step ran for ~116 hours on 110 processors requiring 4GB of RAM per job for workers and 64GB of RAM for the master job.
- So, on a single processor desktop with 64GB of RAM… 12760 hours or 532 days!

OTU Picking

- For 1 billion sequence reads, the initial step ran for ~116 hours on 110 processors requiring 4GB of RAM per job for workers and 64GB of RAM for the master job.
- So, on a single processor desktop with 64GB of RAM… 12760 hours or 532 days!
- One HiSeq2000 generates this data in a week!

Cloud computing

- Implemented on a cluster (or grid), but compute power is rented as a service to support arbitrary applications.

Maintaining hardware is expensive

- Temperature (redundant cooling systems)
- Redundant network connections
- Hardware maintenance (e.g., replacing hard drives)
- Fire suppression
- Back-up power
- System administrator ($$)

Pay-as-you-go compute power

- Public clouds (e.g., Amazon) rent compute resources
- Log in, boot virtual machine image, run analyses, and terminate instance.
- Cheaper for many tasks than buying, maintaining, and supporting a compute cluster.

Types of cloud offerings

- Applications/SaaS (e.g., Google Docs, gmail, Dropbox, iCloud)
- Computing platform/PaaS (e.g., Google App Engine)
- Raw compute resources/IaaS (e.g., Amazon Elastic Compute Cloud (EC2))

Cloud computing options

- Amazon Elastic Compute Cloud (EC2)
- Magellan – Argonne's DOE Cloud Computing
- Data Intensive Academic Grid (DIAG) – Institute for Genome Sciences (IGS), University of Maryland School of Medicine (UMSOM)

Interacting with the Amazon Cloud

- Boot virtual machine image via web interface (or a third-party tool like StarCluster).
- Log in and work via terminal (or via web interface with IPython Notebook)
- Move data back and forth via sftp/scp or a graphical sftp client (e.g., Cyberduck [free/ cross-platform])

Virtual machines

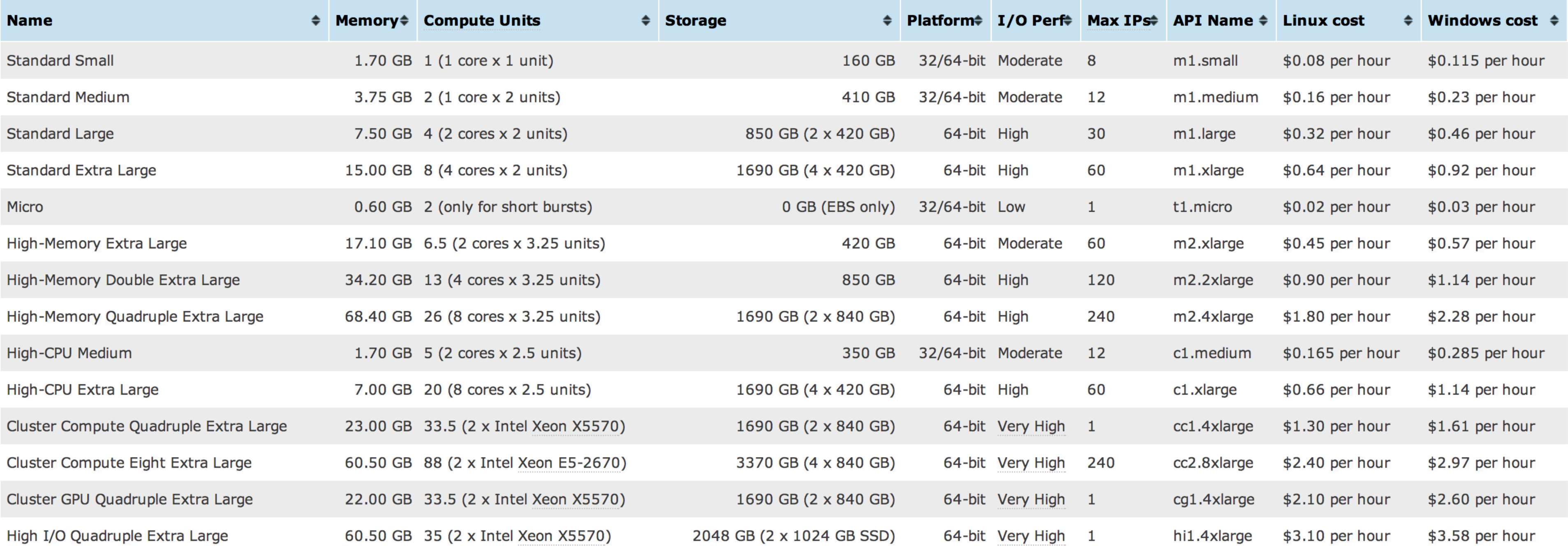
- A “guest” operating system running within a “host” operating system
- A software implementation of a computer, that operates like a physical computer.
- A developer can create a virtual machine *image* which contains their tools pre-installed. Users can then *instantiate* that image to work with those tools.

Browse this page: http://en.wikipedia.org/wiki/Virtual_machine

Benefits that virtual machines offer bioinformatics

- Reproducibility: can publish protocols with a virtual machine instance id.
- Updates are burden of developer, not user.
- Coupled with cloud computing, it's the perfect model for users with sporadic compute needs.

EC2 costs: www.ec2instances.info



Name	Memory	Compute Units	Storage	Platform	I/O Perf	Max IPs	API Name	Linux cost	Windows cost
Standard Small	1.70 GB	1 (1 core x 1 unit)	160 GB	32/64-bit	Moderate	8	m1.small	$0.08 per hour	$0.115 per hour
Standard Medium	3.75 GB	2 (1 core x 2 units)	410 GB	32/64-bit	Moderate	12	m1.medium	$0.16 per hour	$0.23 per hour
Standard Large	7.50 GB	4 (2 cores x 2 units)	850 GB (2 x 420 GB)	64-bit	High	30	m1.large	$0.32 per hour	$0.46 per hour
Standard Extra Large	15.00 GB	8 (4 cores x 2 units)	1690 GB (4 x 420 GB)	64-bit	High	60	m1.xlarge	$0.64 per hour	$0.92 per hour
Micro	0.60 GB	2 (only for short bursts)	0 GB (EBS only)	32/64-bit	Low	1	t1.micro	$0.02 per hour	$0.03 per hour
High-Memory Extra Large	17.10 GB	6.5 (2 cores x 3.25 units)	420 GB	64-bit	Moderate	60	m2.xlarge	$0.45 per hour	$0.57 per hour
High-Memory Double Extra Large	34.20 GB	13 (4 cores x 3.25 units)	850 GB	64-bit	High	120	m2.2xlarge	$0.90 per hour	$1.14 per hour
High-Memory Quadruple Extra Large	68.40 GB	26 (8 cores x 3.25 units)	1690 GB (2 x 840 GB)	64-bit	High	240	m2.4xlarge	$1.80 per hour	$2.28 per hour
High-CPU Medium	1.70 GB	5 (2 cores x 2.5 units)	350 GB	32/64-bit	Moderate	12	c1.medium	$0.165 per hour	$0.285 per hour
High-CPU Extra Large	7.00 GB	20 (8 cores x 2.5 units)	1690 GB (4 x 420 GB)	64-bit	High	60	c1.xlarge	$0.66 per hour	$1.14 per hour
Cluster Compute Quadruple Extra Large	23.00 GB	33.5 (2 x Intel Xeon X5570)	1690 GB (2 x 840 GB)	64-bit	Very High	1	cc1.4xlarge	$1.30 per hour	$1.61 per hour
Cluster Compute Eight Extra Large	60.50 GB	88 (2 x Intel Xeon E5-2670)	3370 GB (4 x 840 GB)	64-bit	Very High	240	cc2.8xlarge	$2.40 per hour	$2.97 per hour
Cluster GPU Quadruple Extra Large	22.00 GB	33.5 (2 x Intel Xeon X5570)	1690 GB (2 x 840 GB)	64-bit	Very High	1	cg1.4xlarge	$2.10 per hour	$2.60 per hour
High I/O Quadruple Extra Large	60.50 GB	35 (2 x Intel Xeon X5570)	2048 GB (2 x 1024 GB SSD)	64-bit	Very High	1	hi1.4xlarge	$3.10 per hour	$3.58 per hour

QIIME virtual machine

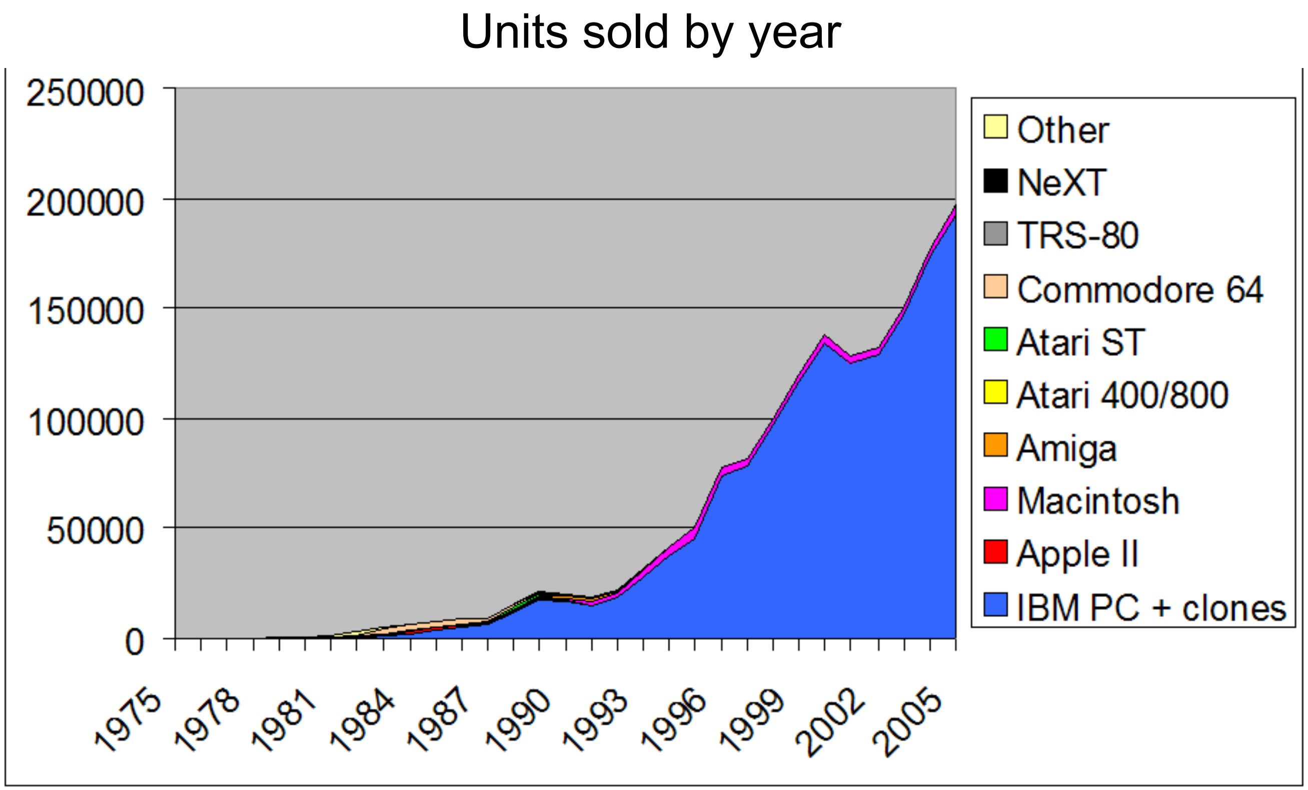
- The QIIME package distributes an EC2 virtual machine with QIIME and its (many) dependencies pre-installed.
- Dependencies include commonly used tools like BLAST, muscle, FastTree, uclust, IPython, and a lot more. A partial list is available here: http://qiime.org/install/install.html
- Latest machine identifier can always be found at: http://qiime.org/home_static/dataFiles.html

I think there is a world market for maybe five computers.

- Thomas Watson, IBM Founder, 1943

I think there is a world market for maybe five computers.

- Thomas Watson, IBM Founder, 1943



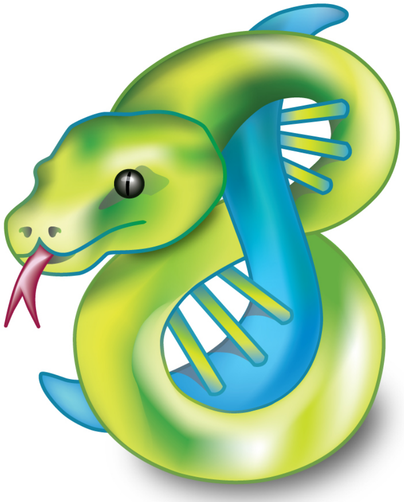
All figures are in units of 1000.
http://jeremyreimer.com/postman/node/329

The democratization of DNA sequencing

©2011, Illumina Inc. All rights reserved.

+

+



Affordable sequencing

Cloud computing

Open-source software