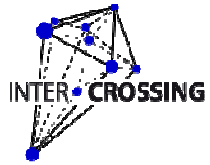


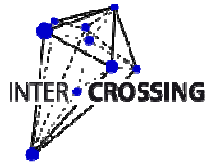
Cloud Computing and NGS data analysis

INTERCROSSING course

August 2013 - Granada

Welcome and Introduction
Eduardo Pareja

Era7 bioinformatics
ohnosequences!

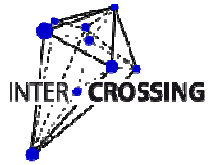


Era7 Bioinformatics activity is based in:

- NGS (Next Generation Sequencing)
- Research
- Focus in Bacterial Genomics
- Cloud Computing

Era7 bioinformatics

ohnosequences!



Era7 Bioinformatics activity is based in:

- NGS (Next Generation Sequencing)
- Research
- Focus in Bacterial Genomics
- Cloud Computing

Era7 bioinformatics

ohnosequences!

Next Generation Sequencing.
DNA sequences
GenBank:

Walter Goad of the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory and others established the Los Alamos Sequence Database in 1979, which culminated in 1982 with the creation of the public GenBank.[4] Funding was provided by the National Institutes of Health, the National Science Foundation, the Department of Energy, and the Department of Defense. LANL collaborated on GenBank with the firm Bolt, Beranek, and Newman, and by the end of 1983 more than 2,000 sequences were stored in it.

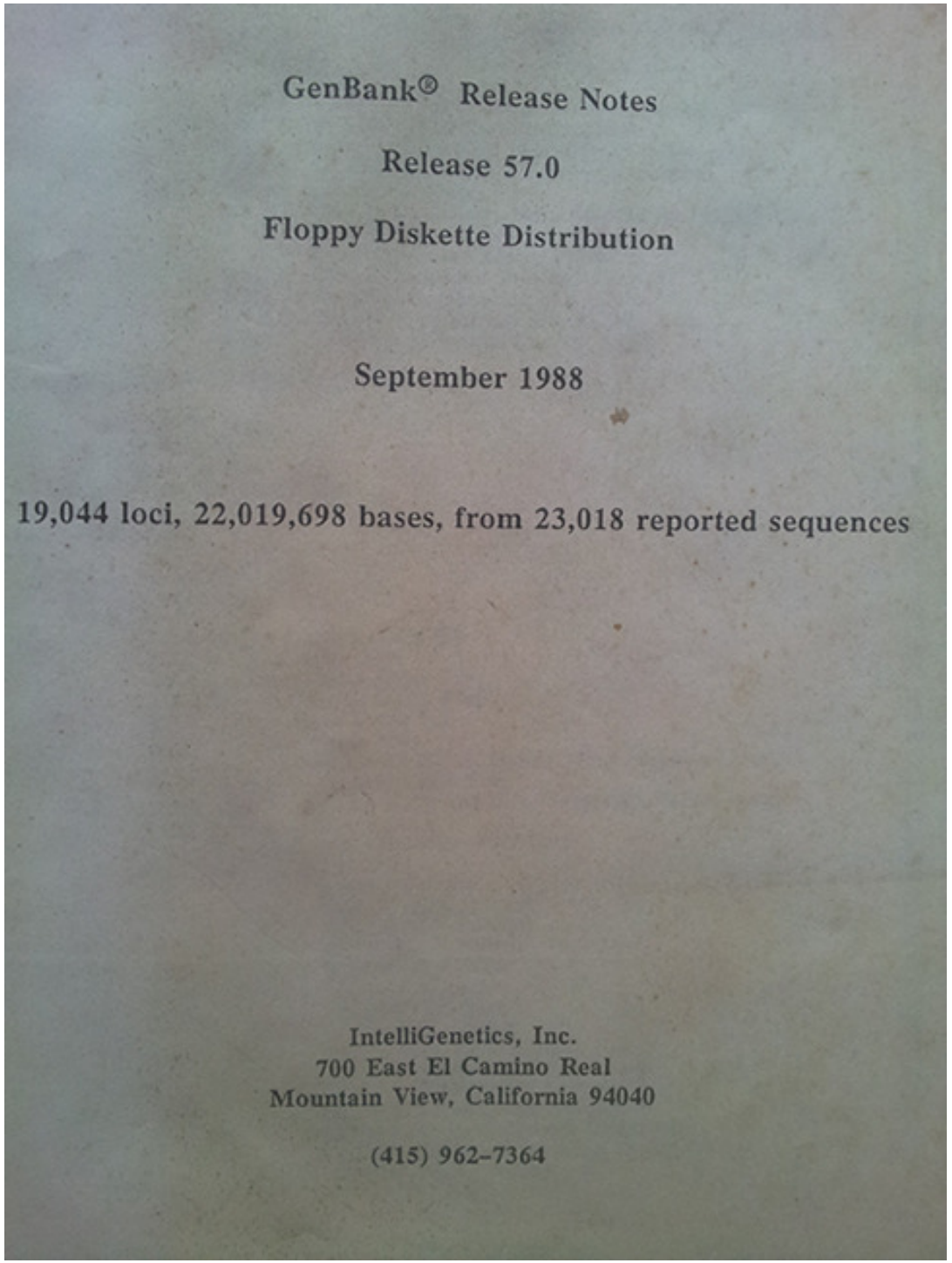
In the mid 1980s, the Intelligenetics bioinformatics company at Stanford University managed the GenBank project in collaboration with LANL.[5]

1988

INTER CROSSING

Era7 bioinformatics

ohnosequences!



1988

INTER CROSSING

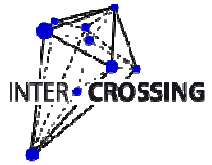
19,044 loci
23,018 sequences
22,019,698 bases

Era7 bioinformatics
ohnosequences!

1988


GenBank ®
Release 57 Disk 1 360KB
GPRI1.ANO GPRI1.SEQ
IntelliGenetics, Inc.

360 Kb

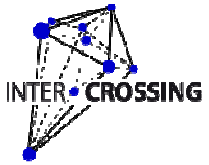


Some numbers related with

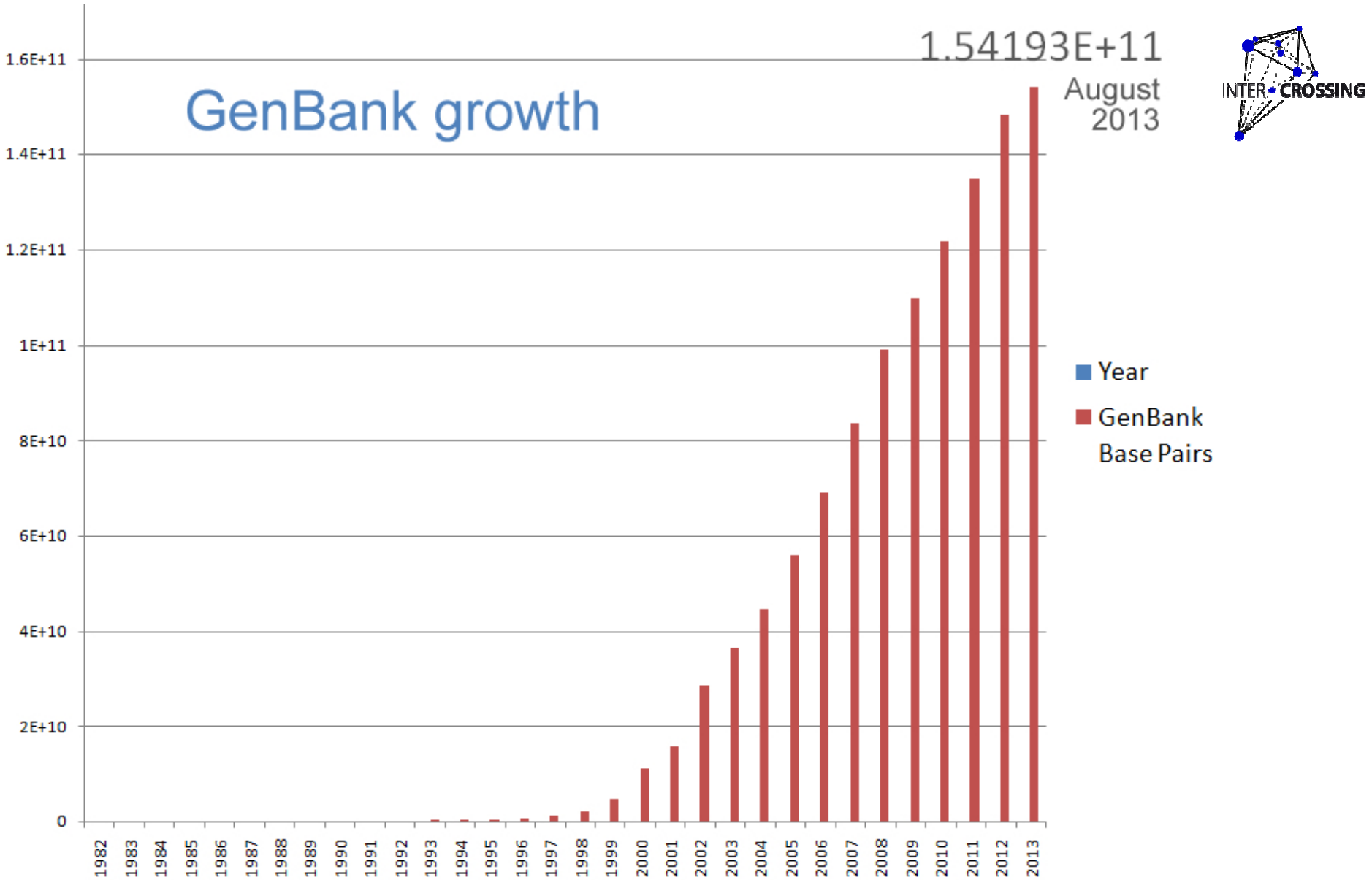
DNA sequencing

Era7 bioinformatics

ohnosequences!



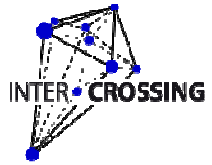
Release	Month	Year	Base Pairs	Entries
3	Dec	1982	680338	606
14	Nov	1983	2274029	2427
20	May	1984	3002088	3665
24	Sep	1984	3323270	4135
25	Oct	1984	3368765	4175
26	Nov	1984	3689752	4393
32	May	1985	4211931	4954
36	Sep	1985	5204420	5700
40	Feb	1986	5925429	6642
42	May	1986	6765476	7416
44	Aug	1986	8442357	8823
46	Nov	1986	9615371	9978
48	Feb	1987	10961380	10913
50	May	1987	13048473	12534
52	Aug	1987	14855145	14020



INTER CROSSING

Era7 bioinformatics

ohnosequences!

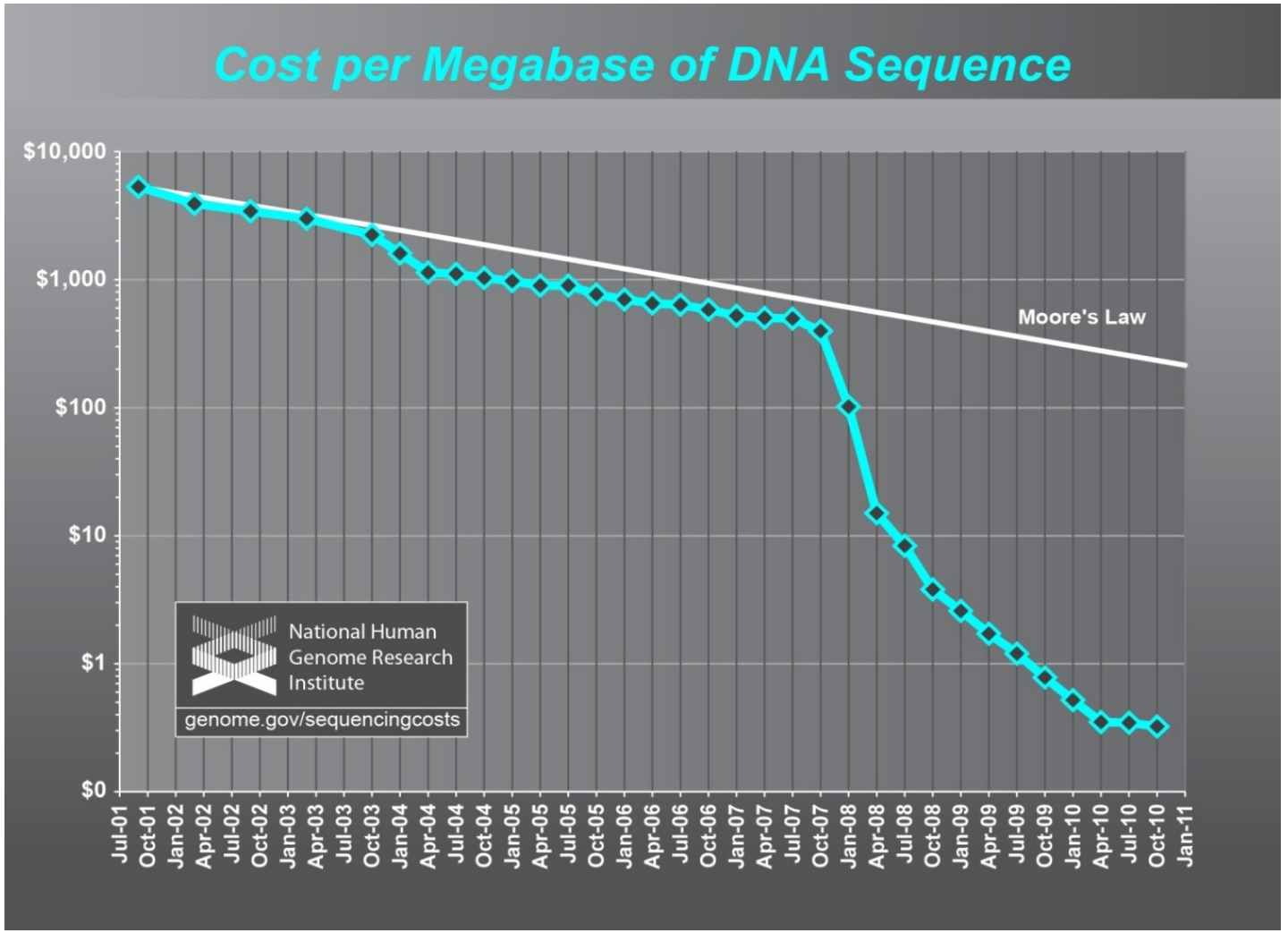


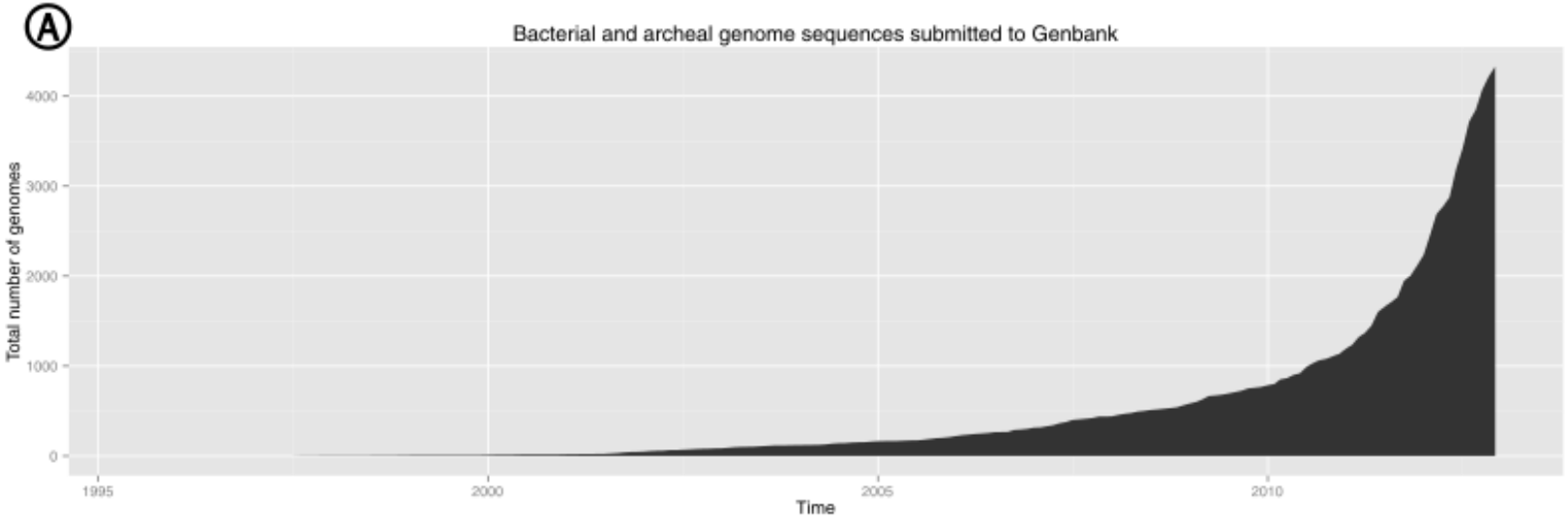
We were interested in DNA sequences and

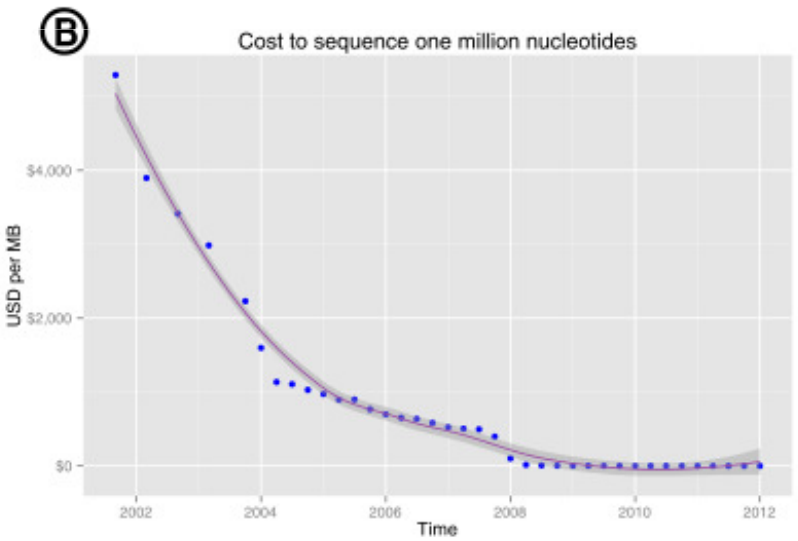
- NGS was introduced in 2005
- Era7 was founded in Sept 2004

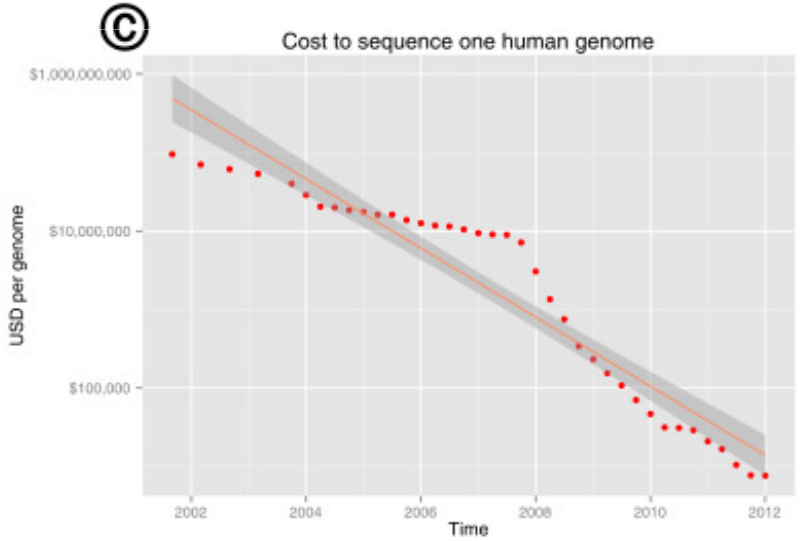
Era7 bioinformatics

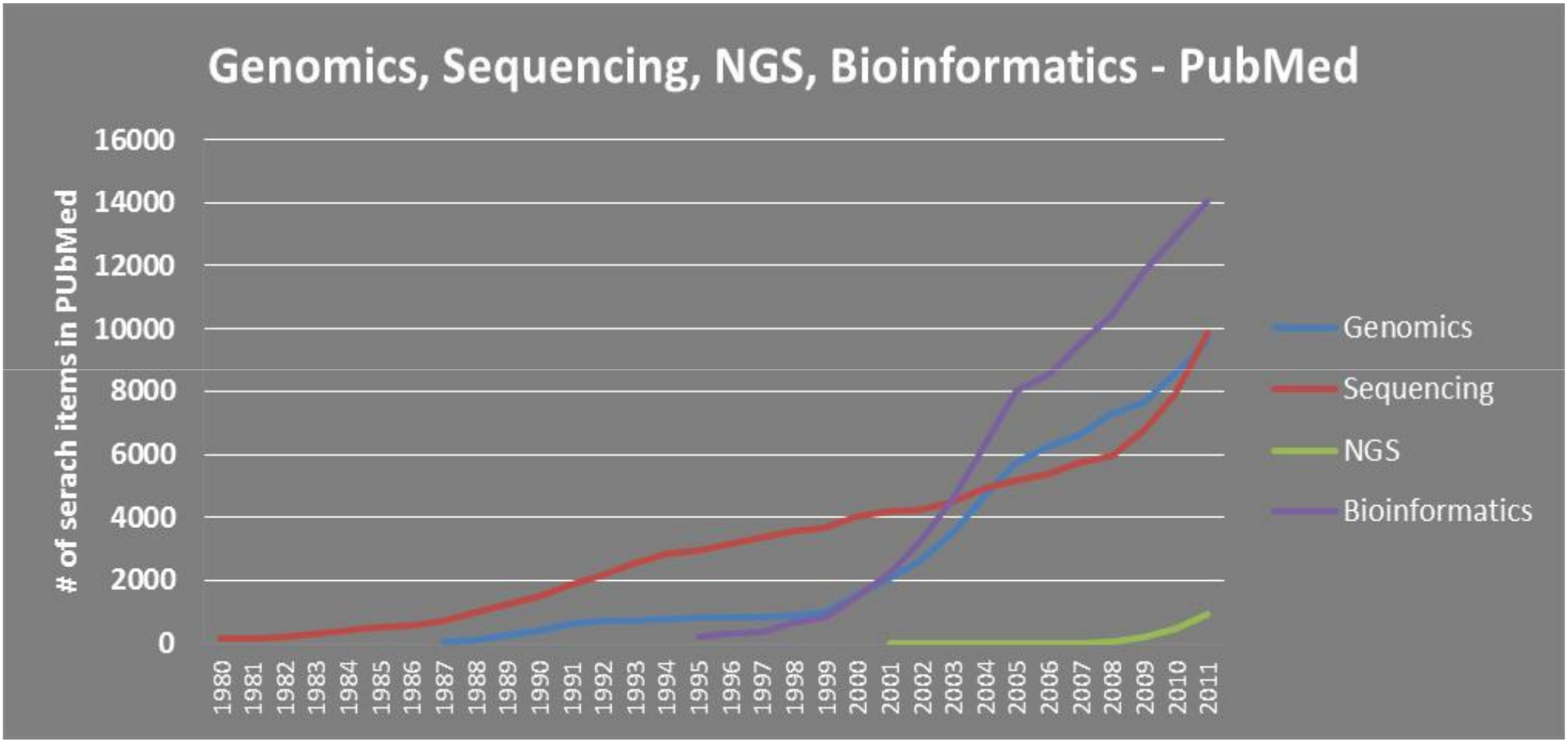
ohnosequences!


Cost per Megabase of DNA Sequence
$10,000
$1,000
$100
$10
$1
$0
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11

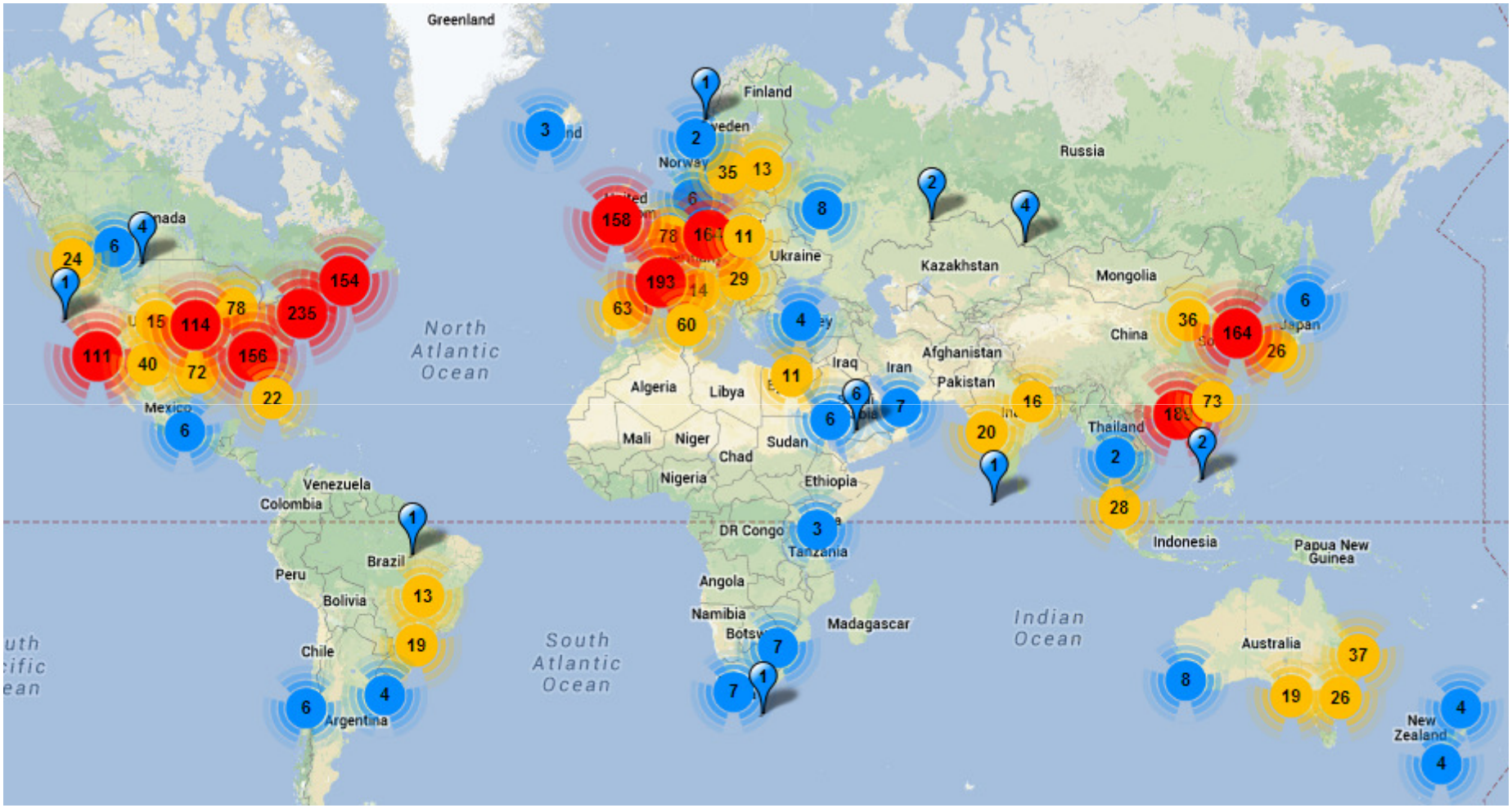

A
Bacterial and archeal genome sequences submitted to Genbank
Total number of genomes
4000
3000
2000
1000
0
1995
2000
2005
2010
Time


B
Cost to sequence one million nucleotides
USD per MB
$4,000
$2,000
$0
2002
2004
2006
2008
2010
2012
Time


C
Cost to sequence one human genome
USD per genome
$1,000,000,000
$10,000,000
$100,000
2002
2004
2006
2008
2010
2012
Time


Genomics, Sequencing, NGS, Bioinformatics - PubMed
# of serach items in PUbMed
16000
14000
12000
10000
8000
6000
4000
2000
0
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Genomics
Sequencing
NGS
Bioinformatics

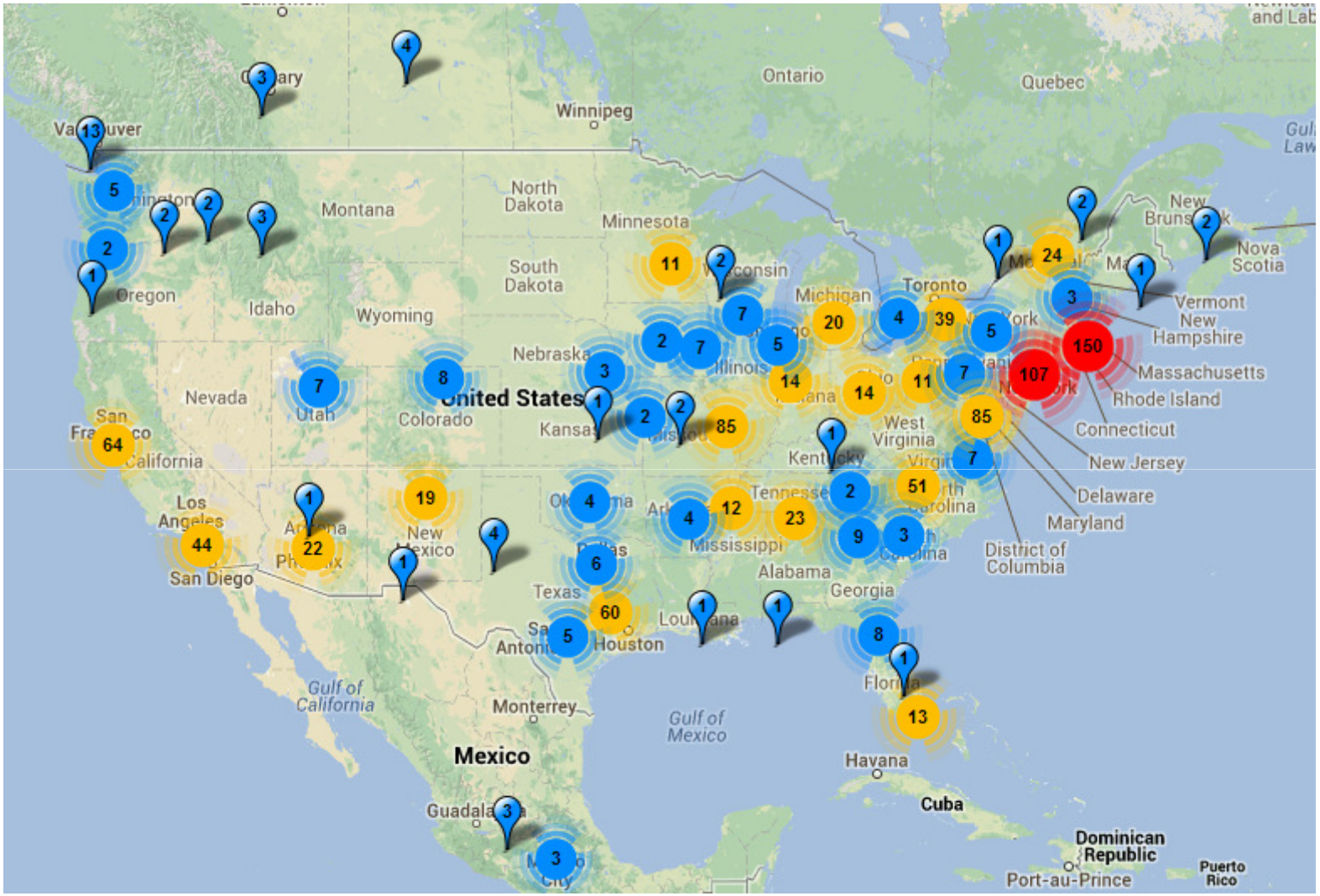
http://omicsmaps.com/



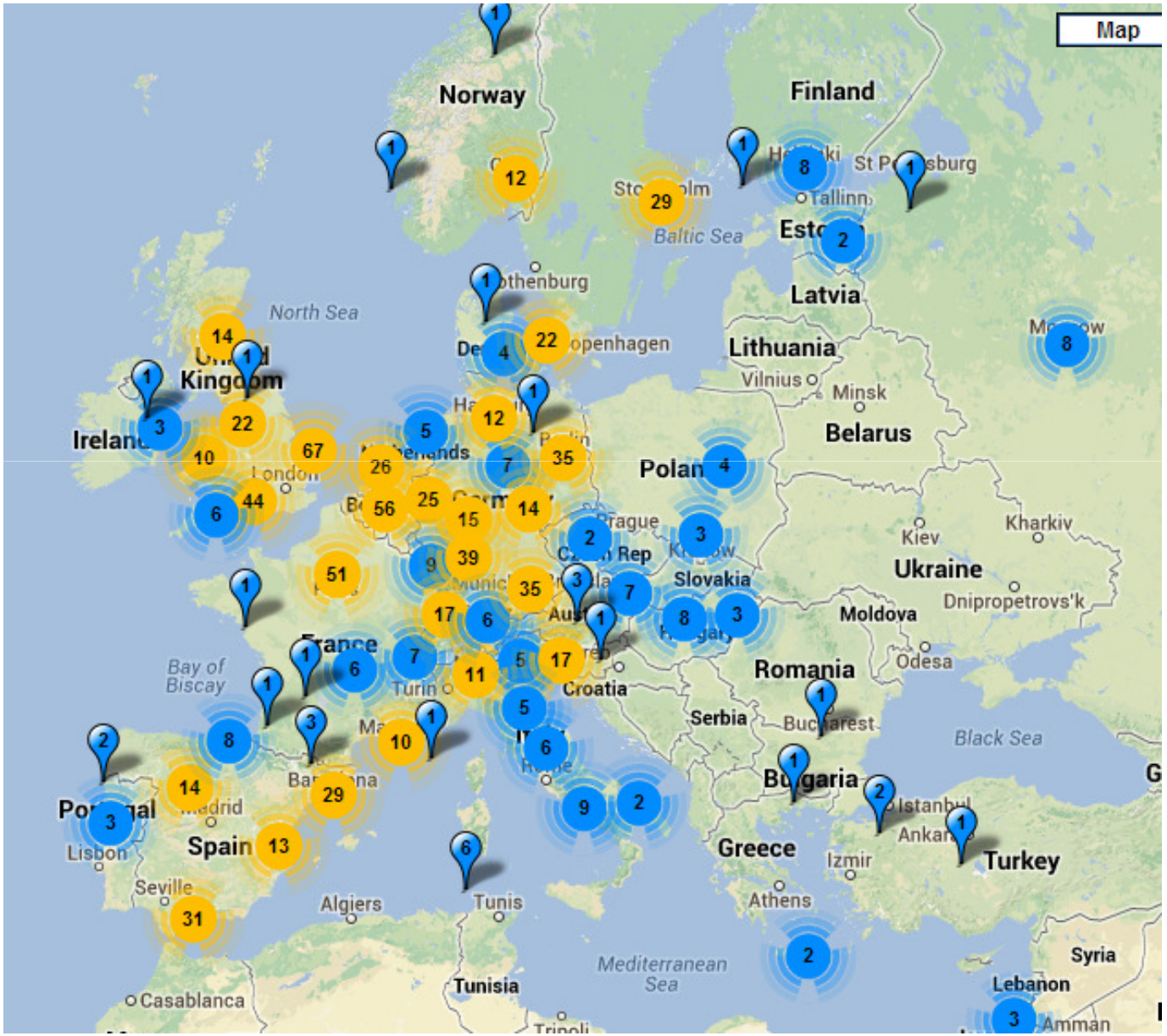
Era7 bioinformatics

ohnosequences!

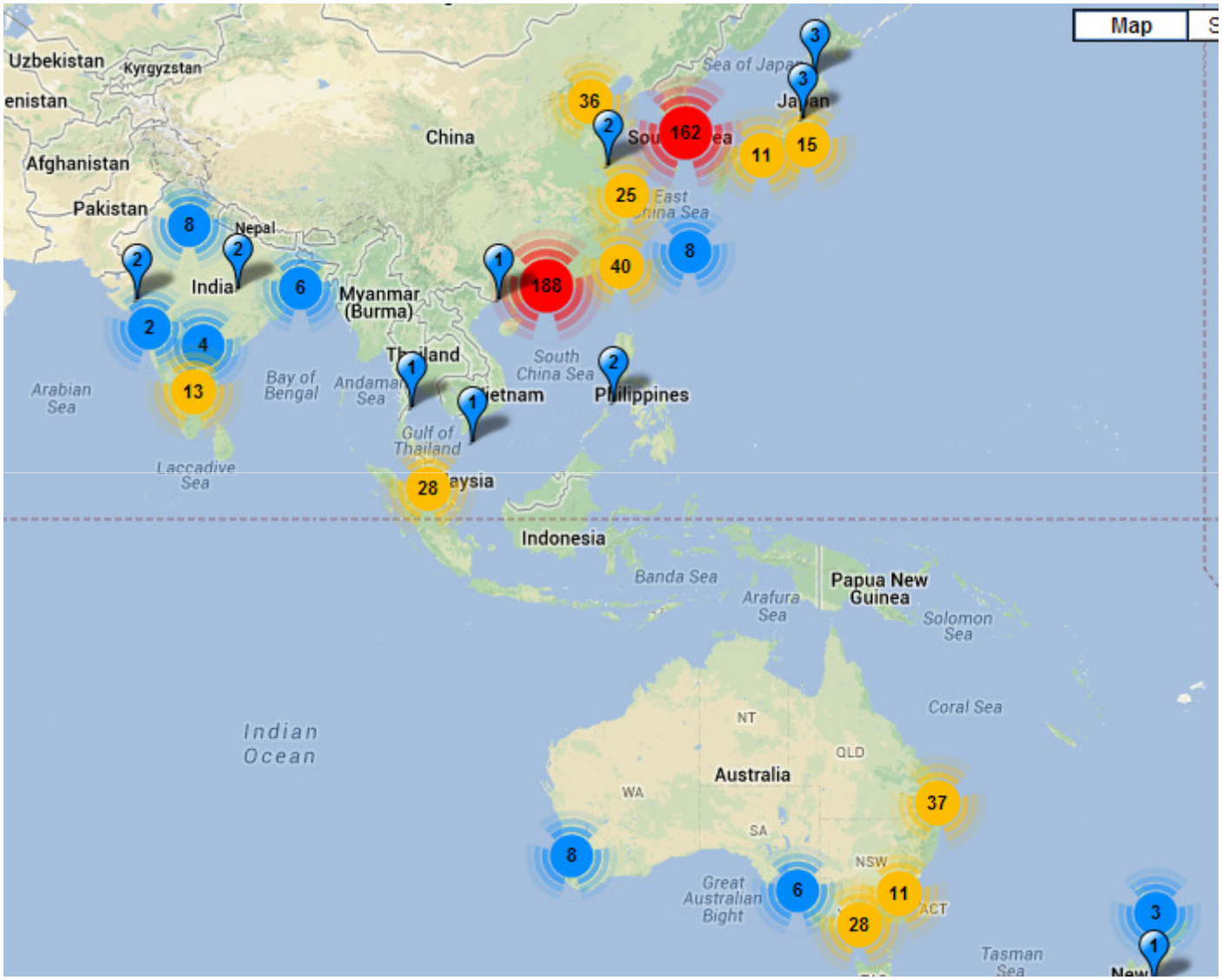
http://omicsmaps.com/



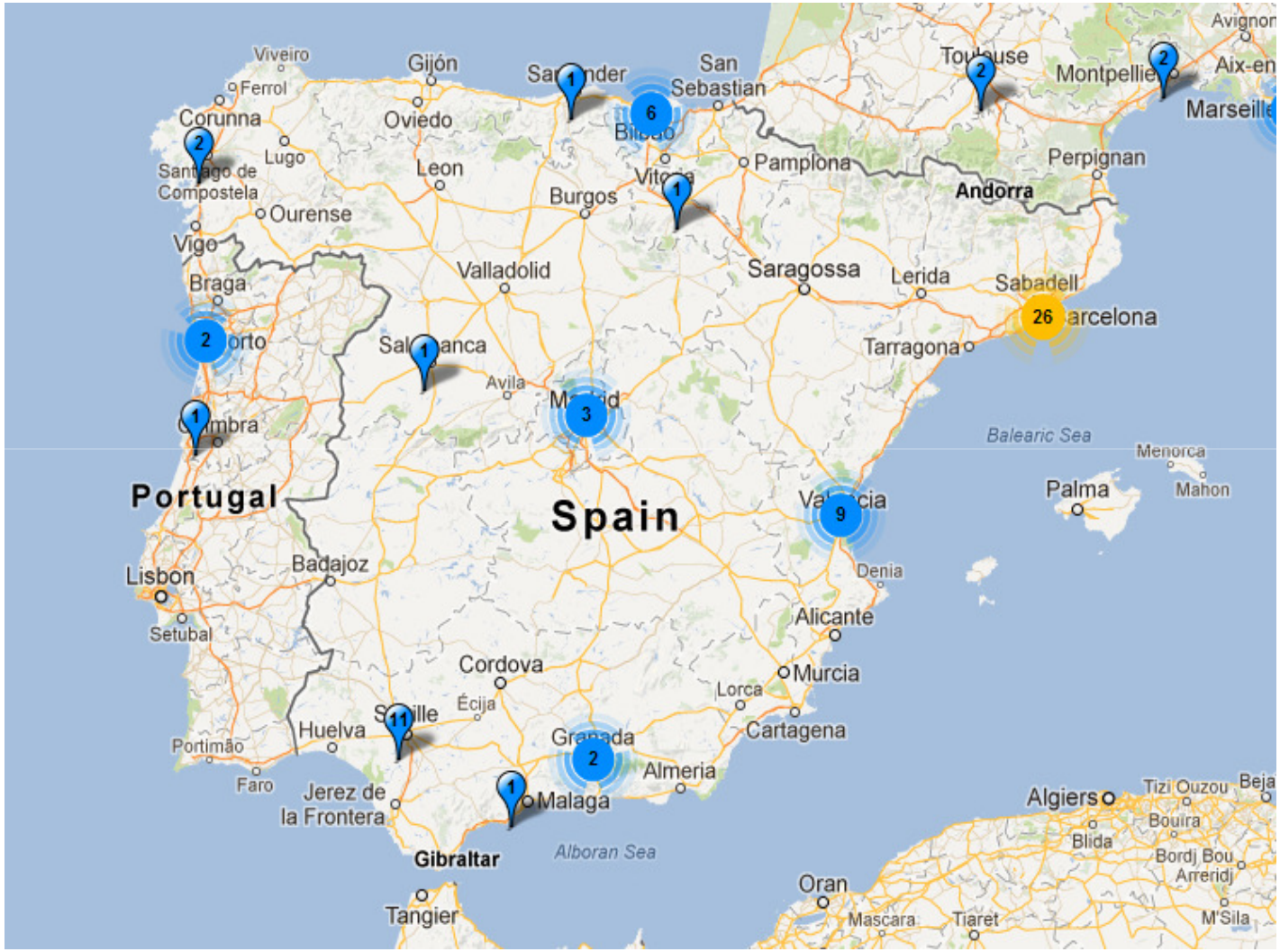
http://omicsmaps.com/

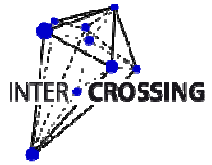


http://omicsmaps.com/



http://omicsmaps.com/





What is the situation of

NGS today ?

Era7 bioinformatics

ohnosequences!

illumina

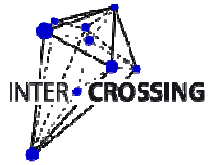
PacBio

Ion Torrent

Roche 454

Era7 bioinformatics

ohnosequences!



>90 % of the DNA ever sequenced

has been sequenced with

illumina machines

Era7 bioinformatics

ohnosequences!

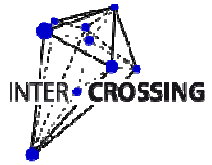
The Next Big Thing could be

Personal Sequencers

Similar, perhaps, to the PCR in the 90s

Era7 bioinformatics

ohnosequences!



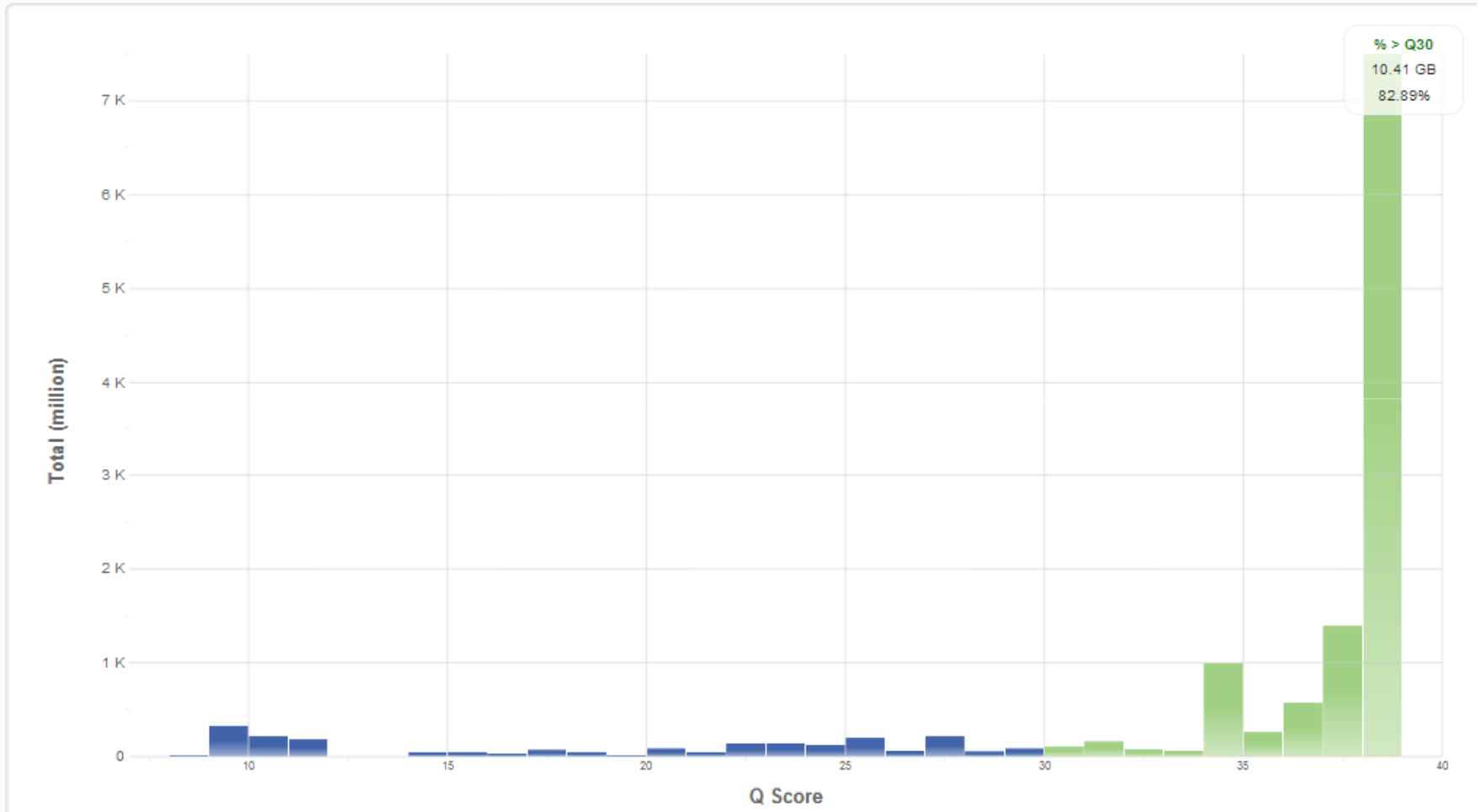
MiSeq from illumina:

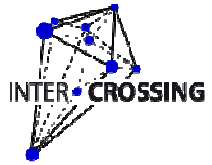
Up to 15 Gb and

2 × 300 bp runs—with the highest data quality.

Era7 bioinformatics
ohnosequences!

E coli 2x300 MiSeq



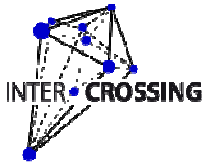


Era7 Bioinformatics activity is based in:

- NGS (Next Generation Sequencing)
- **Research**
- Focus in Bacterial Genomics
- Cloud Computing

Era7 bioinformatics

ohnosequences!

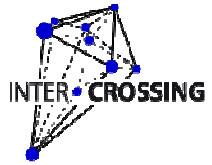


Research at Era7:

ohnosequences!

Era7 bioinformatics

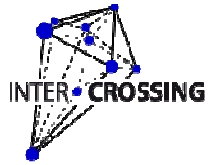
ohnosequences!



Some Projects:

- INTERCROSSING
- bio4j
- BIOGRAPHIKA (bio4j related)
- NEXTMICRO

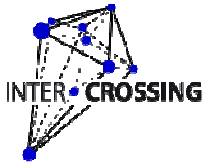
Era7 bioinformatics
ohnosequences!



Some Projects:

- INTERCROSSING
- bio4j
- BIOGRAPHIKA (bio4j related)
- NEXTMICRO

Era7 bioinformatics
ohnosequences!



NEXTMICRO

- AG7 Assembling Genomes: illumina and PacBio
- BG7 Bacterial Genome Annotation (PLOS ONE Nov 2012)
- CG7 Comparative Genomics

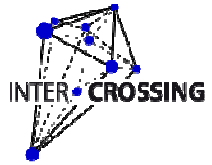
Era7 bioinformatics
ohnosequences!

NEXTMICRO

- Outbreaks
- Different Steps in the Management
- Managing Information about Clones

NEXTMICRO

- Era7 Bioinformatics
- Hospital Ramon y Cajal Madrid
- Funded by CDTI



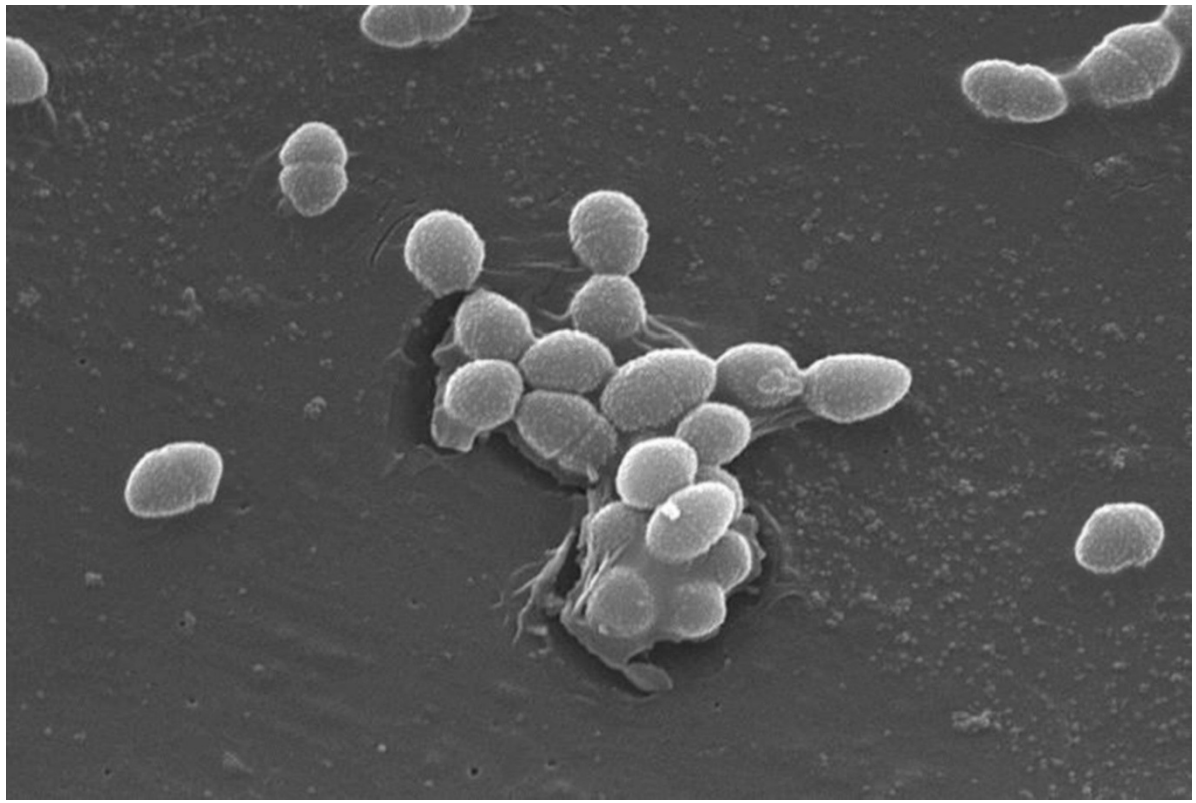
Era7 Bioinformatics activity is based in:

- NGS (Next Generation Sequencing)
- Research
- **Focus in Bacterial Genomics**
- Cloud Computing

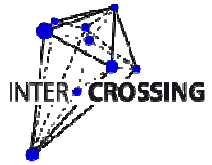
Era7 bioinformatics

ohnosequences!

Bacteria are all over the world

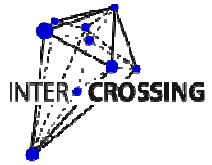


Focus in Bacterial Genomics:



- Bacteria
- Microbiome
- Host-Pathogen relationships: Dual RNA-seq
- Human and animal models
- Biofuels
- Food
- Environmental
- .....

Era7 bioinformatics
ohnosequences!

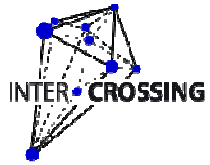


Era7 Bioinformatics activity is based in:

- NGS (Next Generation Sequencing)
- Research
- Focus in Bacterial Genomics
- **Cloud Computing**

Era7 bioinformatics

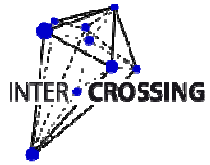
ohnosequences!



Cloud Computing and NGS data analysis
INTERCROSSING course

Era7 bioinformatics

ohnosequences!



Objectives of the course:

To understand Cloud Computing meaning and importance for data analysis in NGS and science in general

To be able to design and use (basic) Cloud Solutions for not to be tied to current solutions

Era7 bioinformatics

ohnosequences!

To reach these goals:

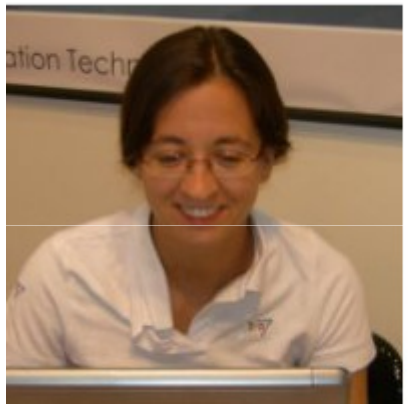
1. We will give an overview of what is the cloud, how it affects research in general and data analysis (NGS) in particular

2. introduce some of the work that we're doing within intercrossing, giving other partners the opportunity to find possible uses and collaboration through these developments

3. hands-on approach: we want you to do something, and to do it by yourselves (with our help of course). Don't hide real, practical issues under the rug of thoroughly prepared artificial examples

	Monday 26	Tuesday 27	Wednesday 28	Thursday 29	Friday 30
10:00 - 11:00	T Welcome	T/P Problem	T Architechture	P Q&A III	P Presentations
11:00 - 11:30	break	break	break	break	break
11:30 - 12:30	T Introduction	T NGS	P nispero	P TW III	P Presentations
12:30 - 14:00	lunch	lunch	lunch	lunch	lunch
14:00 - 15:30	T Cloud What?	P statika	P bio4j	P TW IV	Conclusions
15:30 - 15:45	break	break	break	break	
15:45 - 16:45	P AWS I	P Q&A I	P Q&A II	P Q&A IV	
16:45 - 17:15	break	break	break	break	
17:15: - END	P AWS II	P TW I	P TW II	P TW V	

People:



Marina Manrique

Bioinformatician at Era7 Information Technologies SLU

Granada Area, Spain | Biotechnology

Current	Era7 Information Technologies SLU
Education	Universidad de Granada

People:

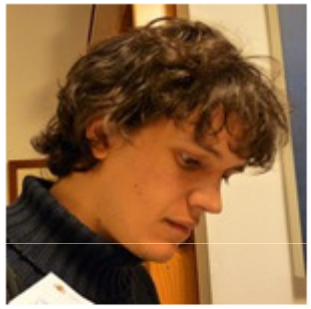
Pablo Pareja Tobes

Bioinformatics IT consultant en Era7 Information Technologies SLU

Granada Area, Spain | Biotechnology

Current	Era7 Information Technologies SLU
Previous	The Ohio State University, Models of Decision and Optimization (MODO) Research Group (Universidad de Granada)
Education	Universidad de Granada

People:



Eduardo Pareja-Tobes

math/cs freak at era7 bioinformatics

Granada Area, Spain | Research

People:

Raquel Tobes

Research Director at Era7 Information Technologies SLU

Granada Area, Spain | Biotechnology

People:

Eduardo Pareja

CEO at Era7 bioinformatics

Granada y alrededores, España | Biotechnology

Granada:

Granada:

Granada:

How did we get the idea of using

Cloud Computing ?

Era7 bioinformatics

ohnosequences!

Amazon puts network power online

GETTY IMAGES

Plug in: industry supercomputer power on the desktop PC could have a big impact on scientific research.

When Dutch computer scientist Rudi Cilibrasi needed hundreds of hours worth of computing time to test a data-mining algorithm earlier this month, he went not to his IT department but to Amazon.com. He paid $60 with his credit card, and in minutes had the equivalent of ten servers installed, which crunched through his job in a couple of days — ten times faster than his desktop PC would have managed.

Large web companies such as Google, eBay and Amazon have far more computing power at their disposal than any academic networks, and have become leaders in massive-scale distributed computing. Many of their innovations can help scientists, and Amazon's computing-on-demand service, which has been running since August, is no exception. It enables customers to create multiple virtual computers on Amazon's massive computing infrastructure for $0.10 per computing hour, and to store data for $0.15 per gigabyte per month.

The service is still in a test phase, so few scientists have even heard of it yet, let alone tried it. But it is a movement that experts believe could revolutionize how researchers use computers. In future, they will export computing jobs to industry networks rather than trying to run them in-house, says Alberto Pace, head of Internet services at CERN, the European particle-physics laboratory near Geneva. CERN has built the world's largest scientific computing grid, bringing together 10,000 computers in 31 countries to handle the 1.5 gigabytes of data that its new accelerator, the Large Hadron Collider, will churn out every second once it is switched on next year.

"I see no reason why the Amazon service wouldn't take off," Pace says. "For a lab that wants to go fast and cheaply, this is a huge advantage over buying material and hiring IT staff. You spend a few dollars, you have a computer farm and you get results."

Cilibrasi, a researcher at the National Institute for Mathematics and Computer Science in Amsterdam, was using Amazon's service to test an algorithm aimed at predicting how much someone will like a movie based on their current preferences. He says he is a convert: "It's substantially more reliable, cheaper and easier to use [than academic computing networks]. It opens up powerful computing-on-demand to the masses."

"You spend a few dollars, you have a computer farm and you get results."

It's not just computer scientists who could use cheap computing power. Climate researchers can run global-warming models, cosmologists might simulate the inside of a supernova or biologists could scale up the workings of a liver cell to the whole organ.

The South African National Bioinformatics Institute at the University of Western Cape, Belleville, has already been testing Amazon's system to power large-scale genome comparisons. The pay-as-you-go system offers computing power and bandwidth that the institute could not afford to maintain itself, says Inus Scheepers, a systems administrator there.

Cost is certainly one reason observers are excited about Amazon's system. Other companies, including Sun Microsystems, offer computing-on-demand, but Amazon's service costs a tenth as much. But the main attraction is Amazon's use of 'virtualization' technologies, which many predict will change not just research but computing itself.

Virtualization uses a layer of software to allow multiple operating systems to run together. This means that different computers can be recreated on the same machine. So one machine can host say ten 'virtual' computers, each with a different operating system.

That's a big deal. Running multiple virtual computers on a single server uses available resources much more efficiently. But it also means that instead of having to physically install a machine with a particular operating system, a virtual version can be created in seconds. Such virtual computers can be copied just like a file, and will run on any machine irrespective of the hardware it is using. "In the past, we had to install hardware and software for each machine," says Pace.

Virtualization software is becoming widely available from companies such as Microsoft and VMware. Amazon uses Xen, an open-source system developed at the University of Cambridge, UK, which is fast becoming popular with researchers. As well as using Xen to enable virtual computers to run across a grid or cluster, despite different operating systems on the individual machines, researchers can also use applications developed on their lab computer.

At present, to run an application on a large scale, they often need to rewrite it. With virtualization, researchers can create a copy of their own machine and use it to run large-scale simulations or searches, and it should work exactly as it does in the lab.

Amazon's service combines the joys of virtualization with the huge computing power it has at its disposal. It's an approach that looks set to catch on. CERN has started an internal service similar to Amazon's, in which users can create or delete virtual machines on the fly. "Virtualization is revolutionary," says Pace. "It's clear that this is one way to do scientific research in the future."

Declan Butler

528

Nature
News
Nov.
2006

From the news article in Nature:

"You spend a few dollars, you have a computer farm and you get results"

Era7 bioinformatics
ohnosequences!

From the news article in Nature:

The South African National Bioinformatics Institute at the University of Westerns Cape, Belleville, has already been testing Amazon's system to power large-scale genome comparisons.

"The pay-as-you-go system offers computing power and bandwith that the Institute could not afford to maintain itself."

From the news article in Nature:

Runing since August 2006, Amazon's service enables customers to create multiple virtual computers for $0.10 per computing hour and to store data for $0.15 per gigabyte per month

Today is even cheaper !!

From the news article in Nature:

Industry supercomputer power on the desktop PC could have a big impact on scientific research.

The main attraction is Amazon's use of virtualization technologies, which many predict will change not just research but computing itself

Granada’s local provider:

So, It seemed that we could have Computing and Storage:

- On-demand
- Scalable
- Pay-per-use

Era7 bioinformatics

ohnosequences!

We discussed the news, and

we started to work in AWS at

Era7 Bioinformatics from 2007

Era7 bioinformatics

ohnosequences!

Aws Services Today

Compute & Networking

- AWS Direct Connect »
- Amazon EC2 »
- Elastic Load Balancing »
- Auto Scaling »
- Amazon EMR »
- Amazon Route 53 »
- Amazon VPC »

Storage & Content Delivery

- Amazon S3 »
- Amazon Glacier »
- Amazon CloudFront »
- AWS Storage Gateway »
- AWS Import/Export »

App Services

- Amazon Elastic Transcoder »
- Amazon SQS »
- Amazon SNS »
- Amazon SES »
- Amazon SWF »
- Amazon CloudSearch »

Database

- Amazon DynamoDB »
- Amazon RDS »
- Amazon Redshift »
- Amazon ElastiCache »
- Amazon SimpleDB »

Deployment & Management

- AWS Elastic Beanstalk »
- AWS CloudFormation »
- Amazon CloudWatch »
- AWS Data Pipeline »
- AWS Identity and Access Management »
- AWS OpsWorks »

Era7 bioinformatics

ohnosequences!

Use cases:

The New York Times. The New York Times Archives + Amazon Web Services = TimesMachine.

TimesMachine is a collection of full-page image scans of the newspaper from 1851–1922

The New York Times

Use cases: 2008

The New York Times


The New York Times.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

The Lost Titanic Sailing Turned Out of Belfast Harbor.

PARTIAL LIST OF THE SAVED.

Use cases:

Telefonica (Spanish global telephone operator) uses AWS for elaborating the bills once a month

Use cases:

The Force.com Toolkit for Amazon Web Services makes it easy for developers to combine the functionality of Force.com—salesforce.com's platform for building software-as-a-service applications—with Amazon Web Services to create innovative business applications in the cloud.

Use cases:

DICOM Grid, Arizona, uses AWS to store, distribute and share medical images

Powering Medical Image Exchange

Use cases:

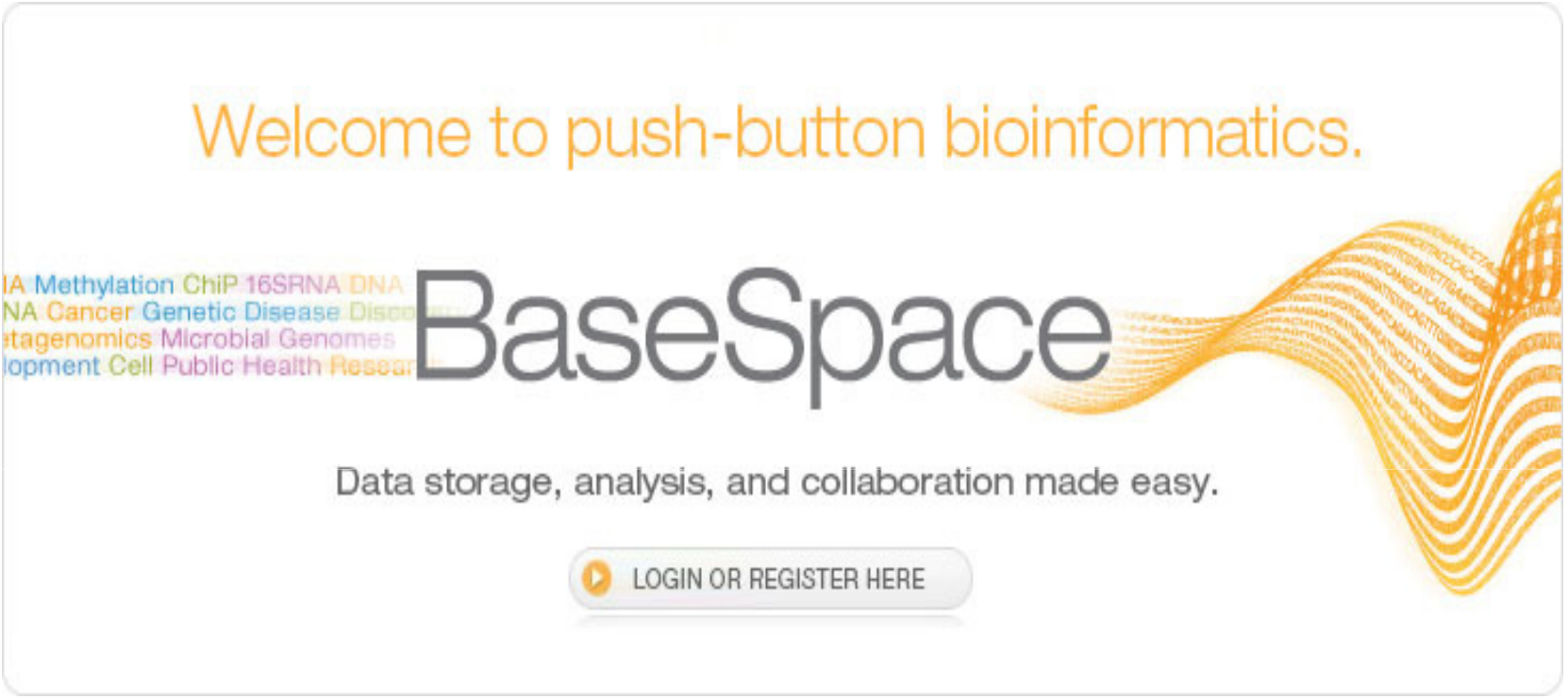
DNAnexus relies on Amazon Simple Storage Service (Amazon S3) to meet the company's extensive storage demand, which will grow from terabytes into petabytes of data

Use cases:

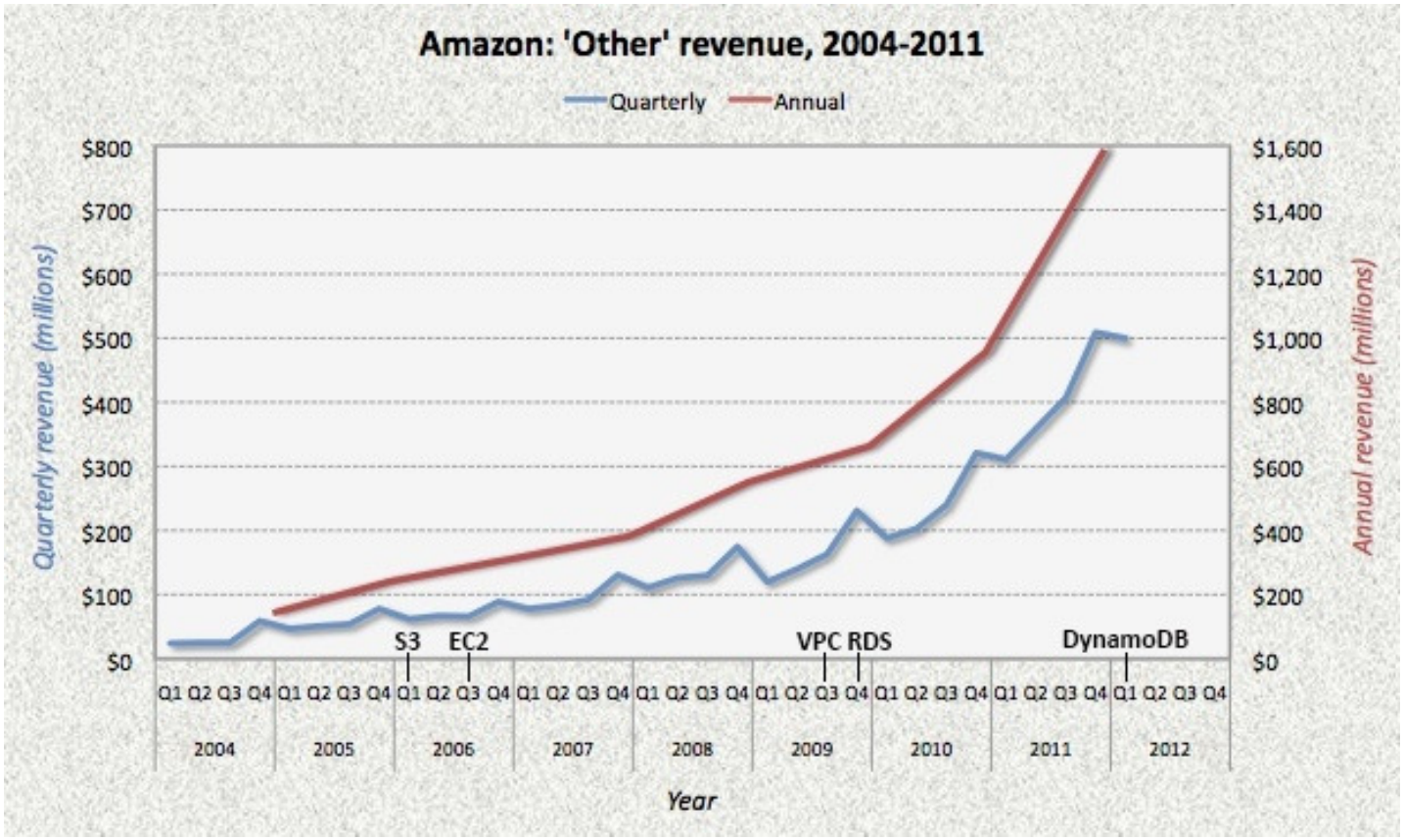
Era7 Bioinformatics uses S3, EC2, ...
To assemble, annotate and compare Bacterial Genomes and performs Metagenomics studies

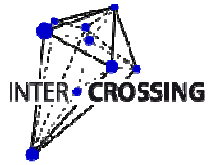
Era7 bioinfo

Era7 bioinformatics
ohnosequences!



This is a very interesting use case based in AWS because the data is uploaded from the machines in real time before the run has finished


Amazon: 'Other' revenue, 2004-2011
Quarterly
Annual
Quarterly revenue (millions)
Annual revenue (millions)
$800
$700
$600
$500
$400
$300
$200
$100
$0
$1,600
$1,400
$1,200
$1,000
$800
$600
$400
$200
$0
S3
EC2
VPC
RDS
DynamoDB
Q1 Q2 Q3 Q4
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year



Is there any reason for not using AWS?

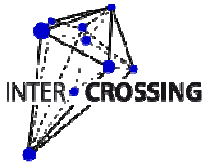
Probably there could be a few. What I have found many times:

Security and Privacy Concerns

Era7 bioinformatics

ohnosequences!

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy, Security and Breach Notification Rules



AWS Assurance Programs

The AWS cloud infrastructure has been designed and managed in alignment with regulations, standards, and best-practices including:

- HIPAA
- SOC 1/SSAE 16/ISAE 3402 (formerly SAS70)
- SOC 2
- SOC 3
- PCI DSS Level 1
- ISO 27001
- FedRAMPSM
- DIACAP and FISMA
- ITAR
- FIPS 140-2
- CSA
- MPAA

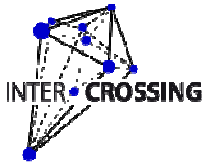
Era7 bioinformatics
ohnosequences!

But the privacy and security problem is not

a specific problem of the Cloud:

A lot of laptop thefts in the USA with patient's data from medical records, clinical trials, etc.

This would not happen in the Cloud

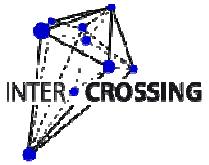


Third-Party Attestations, Reports and Certifications

HIPAA

AWS enables covered entities and their business associates subject to the U.S. Health Insurance Portability and Accountability Act (HIPAA) to leverage the secure AWS environment to process, maintain, and store protected health information and AWS will be signing business associate agreements with such customers. AWS also offers a HIPAA-focused whitepaper for customers interested in learning more about how they can leverage AWS for the processing and storage of health information. The Creating HIPAA-Compliant Medical Data Applications with AWS whitepaper outlines how companies can use AWS to process systems that facilitate HIPAA and HITECH compliance. For more information on the AWS HIPAA compliance program please contact AWS Sales and Business Development.

2013 New version of HIPAA

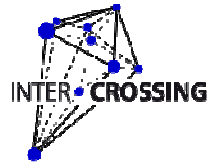


HIPAA Support Widens In Cloud Vendor Community

Earlier this year, the Department of Health and Human Services issued their "Final Rule" to modify HIPAA – the 1996 legislation that's the beating heart of healthcare security, privacy and data portability. The lack of data portability continues to plague the industry – but the security and privacy are getting more serious with each iteration. The fun reading (courtesy of the Federal Register) is online here. One fast fact from a quick search of the document finds no less than 1,358 occurrences of the phrase "Business Associate."

HHS estimates that the cost of compliance with the final rule will be somewhere between $114 and $225 million (first year) and about $14 million for each year thereafter. For those who track this sort of thing, that means the

Era7 bioinformatics
ohnosequences!



Creating Healthcare Data Applications to Promote HIPAA and HITECH Compliance

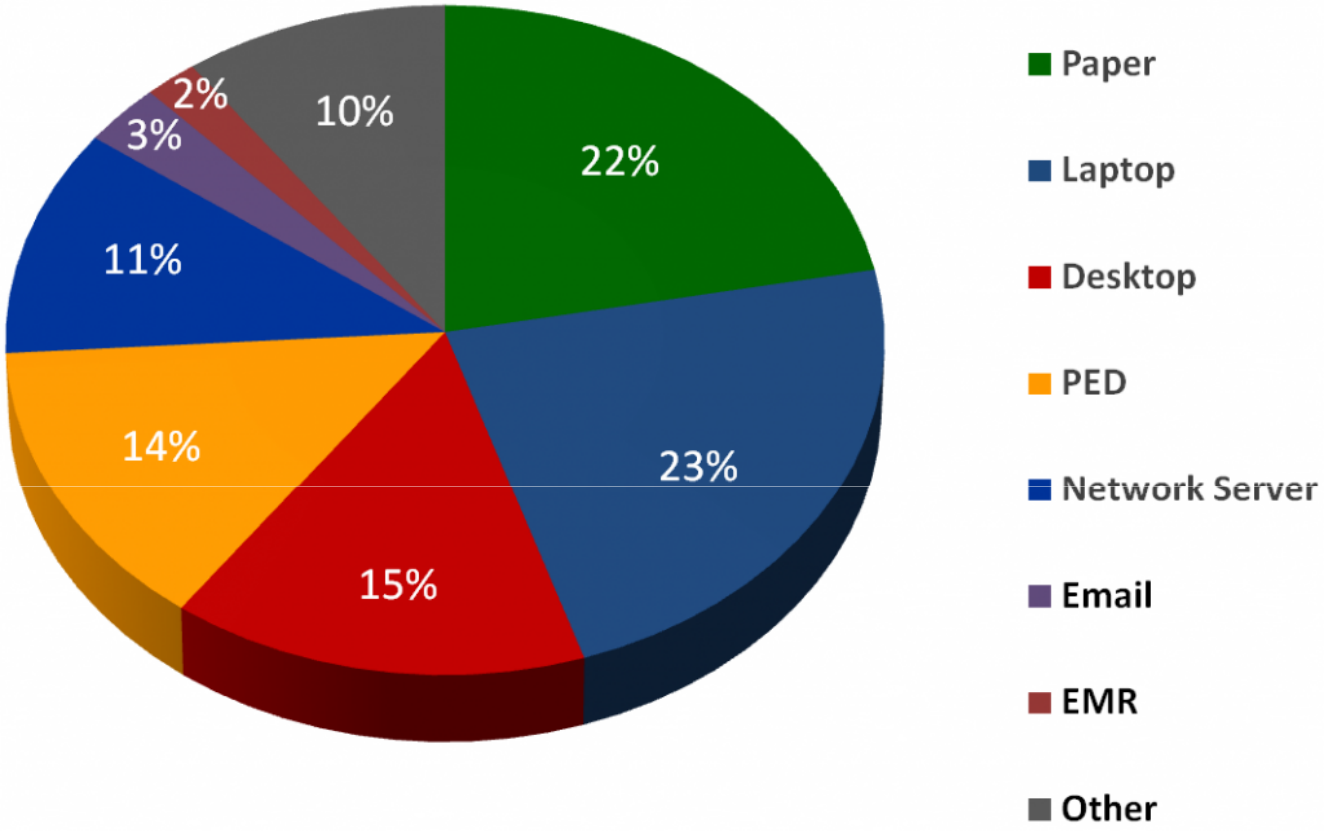
August 2012

Some discussion now from January 2013

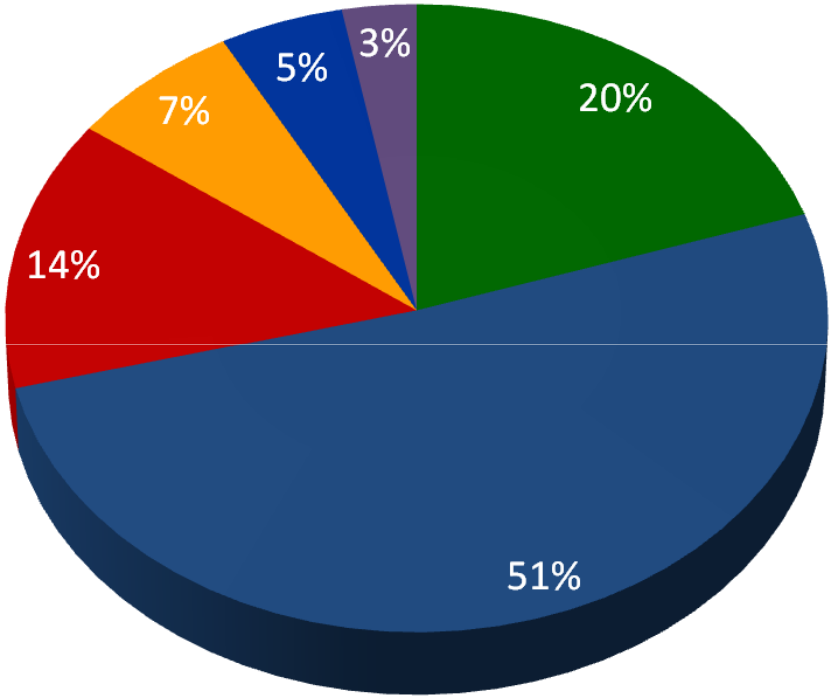
Era7 bioinformatics

ohnosequences!

Breach Notification: 500+ by Location



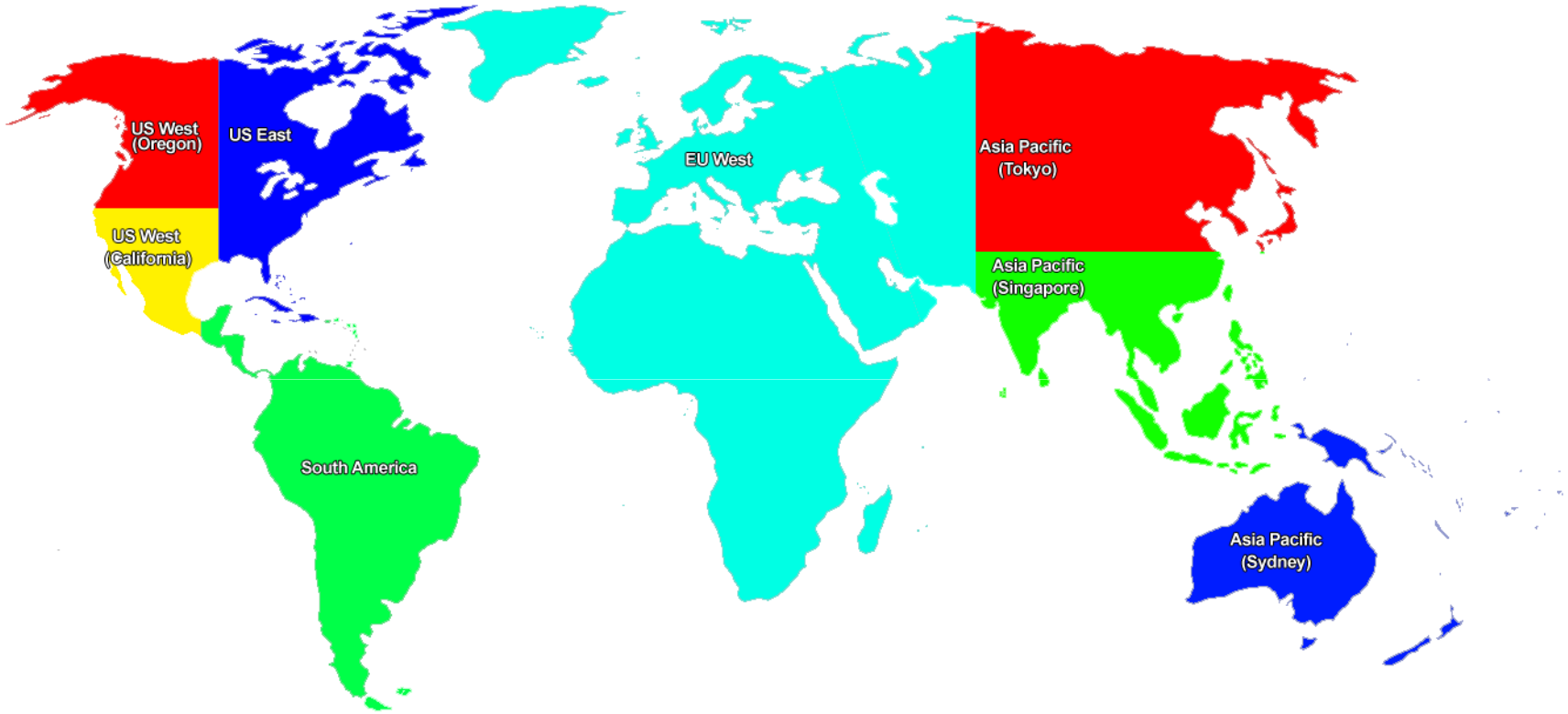
Breach Notification: 500+ by Type

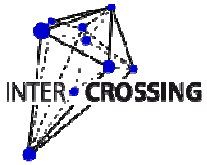


- Unauthorized Access/Disclosure
- Theft
- Loss
- Hacking / IT Incident
- Improper Disposal
- Unknown

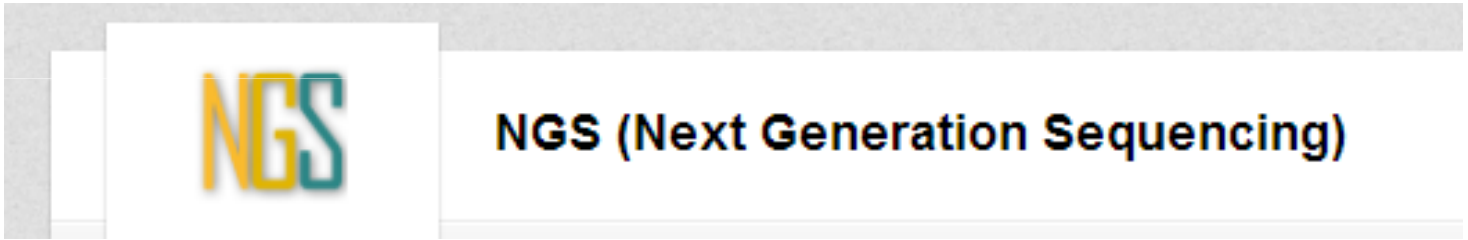
“Foreign clouds in the European sky”

Welcome and Introduction
Eduardo Pareja


US West
(Oregon)
US East
US West
(California)
EU West
Asia Pacific
(Tokyo)
Asia Pacific
(Singapore)
South America
Asia Pacific
(Sydney)



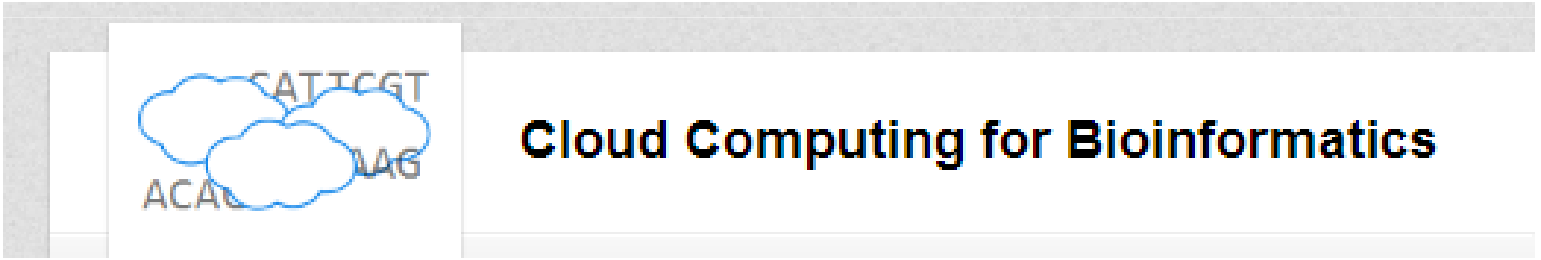
There is a LinkedIn group for NGS:



Era7 bioinformatics
ohnosequences!

INTER·CROSSING

There is also a LinkedIn group for this:



Era7 bioinformatics
ohnosequences!

In summary:

Welcome to Granada !!

and we will do our best

to helping you in your way to the Cloud

INTER·CROSSING

Era7 bioinformatics

ohnosequences!