Bionimbus: From Big Data to Clouds and Commons

Robert Grossman
University of Chicago
Open Cloud Consortium

June 17, 2014
Open Science Data Cloud PIRE Workshop
Amsterdam

Four questions and
one challenge

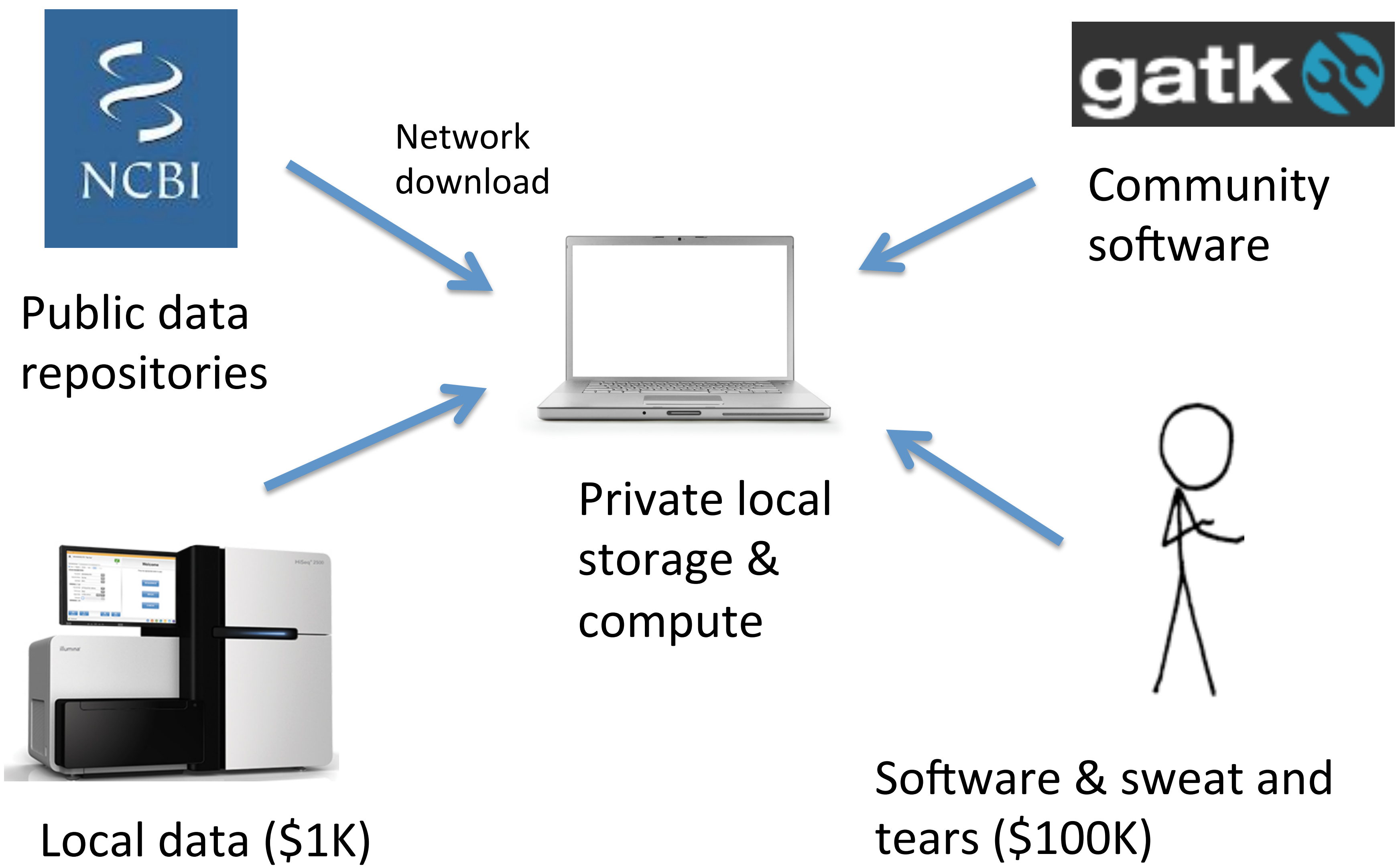
1. What is the same and what is different about big biomedical data vs big science data and vs big commercial data?
2. What instrument should we use to make discoveries over big biomedical data?
3. Do we need new types of mathematical and statistical models for big biomedical data?
4. How do we organize large biomedical datasets to maximize the discoveries we make and their impact on health care?

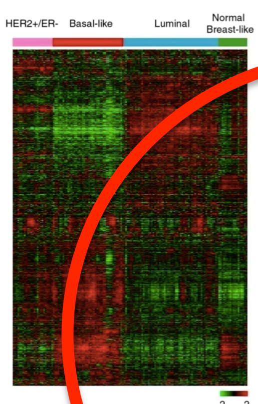
One Million Genome Challenge

- Sequencing a million genomes would likely change the way we understand genomic variation.
- The genomic data for a patient is about 1 TB (including samples from both tumor and normal tissue).
- One million genomes is about 1000 PB or 1 EB
- With compression, it may be about 100 PB
- At $1000/genome, the sequencing would cost about $1B
- Think of this as one hundred studies with 10,000 patients each over three years.

Part 1:
Biomedical computing is being disrupted by big data

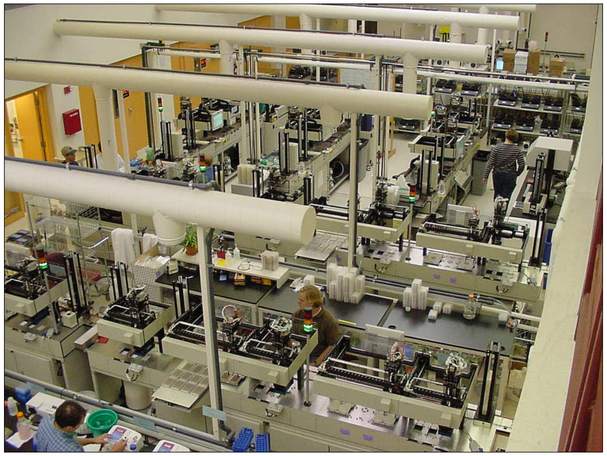
Standard Model of Biomedical Computing



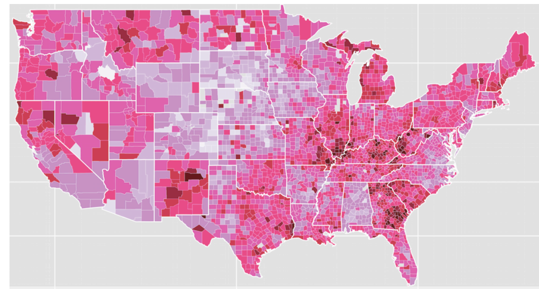


1000 patients

We have a problem …



Growth of data



THE CANCER IMAGING ARCHIVE

New types of data

It takes over three weeks to download the TCGA data at 10 Gbps

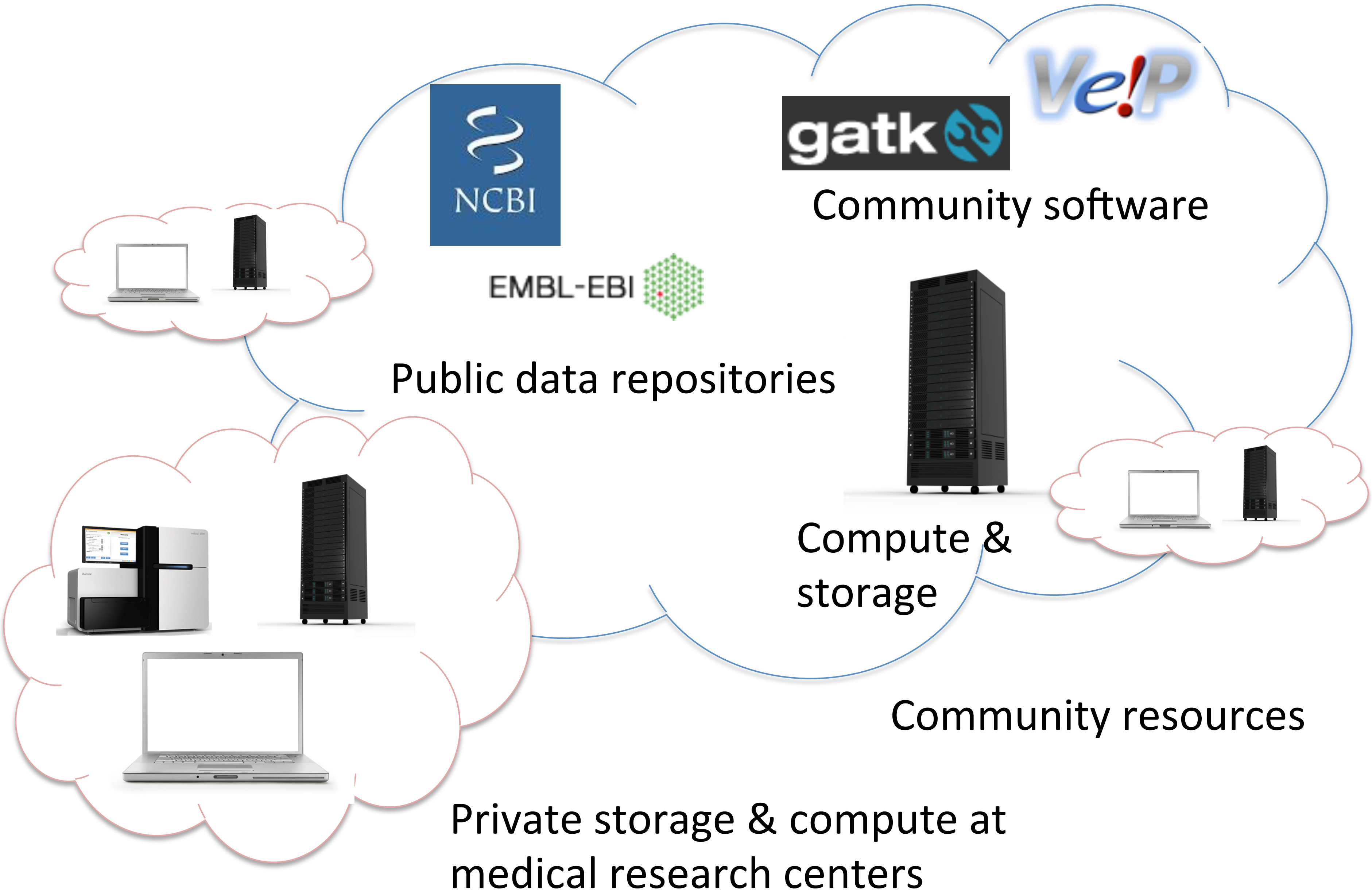
Analyzing the data is more expensive than producing it

Image: A large-scale sequencing center at the Broad Institute of MIT and Harvard.

The Smart Phone is Becoming a Home for Medical & Environmental Sensors

Source: LifeWatch V from LifeWatch AG, www.lifewatchv.com.

New Model of Biomedical Computing



The Tragedy of the Commons

The Tragedy of the Commons

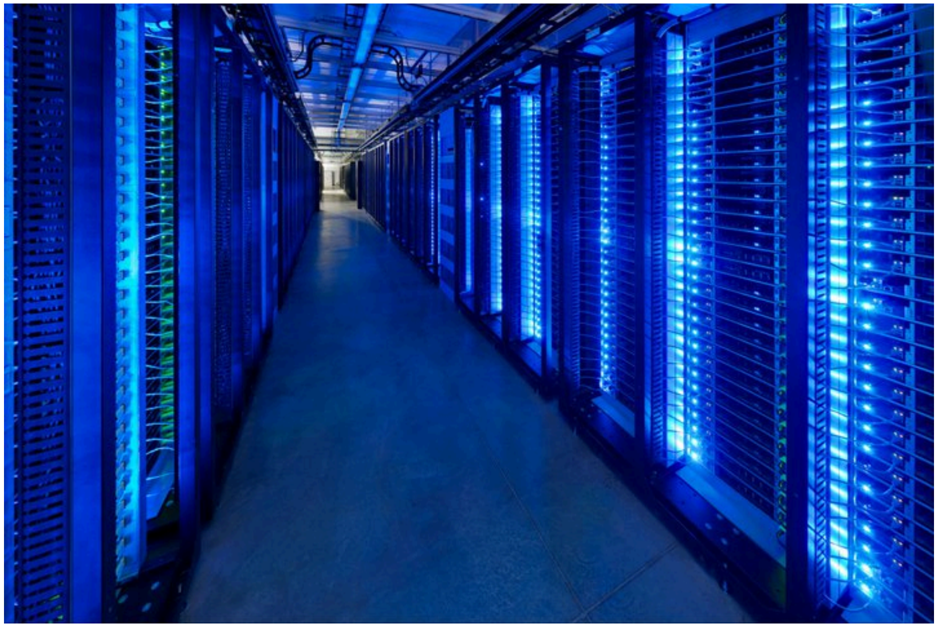
The population problem has no technical solution; it requires a fundamental extension in morality.

Garrett Hardin

Individuals when they act independently following their self interests can deplete a deplete a common resource, contrary to a whole group's long-term best interests.

Source: Garrett Hardin, The Tragedy of the Commons, Science, Volume 162, Number 3859, pages 1243-1248, 13 December 1968.

A Possible Big Data Strategy for Biomedial Data



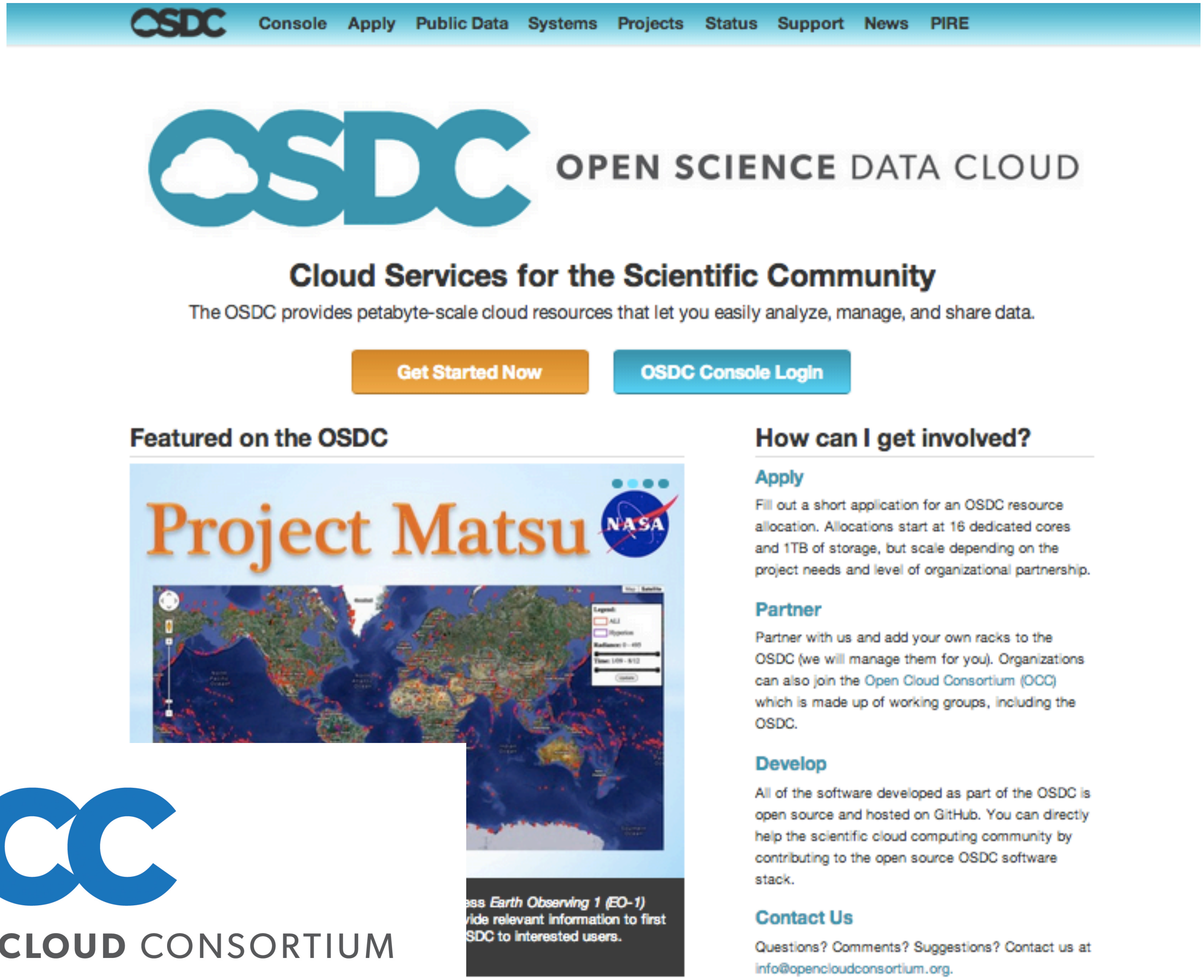
1. Create several community data commons for biomedical data.
2. Develop several secure, compliant cloud computing infrastructures for biomedical data.
3. Interoperate the data commons with the clouds and other computing infrastructure.

2005 - 2015	**Bioinformatics tools & their integration.** Examples: Galaxy, GenomeSpace, workflow systems, portals, etc.
2010 - 2020	**???**
2015 - 2025	**???**

Part 2
Biomedical and Genomic Clouds
(Data Center Scale Computing over Biomedical Data)

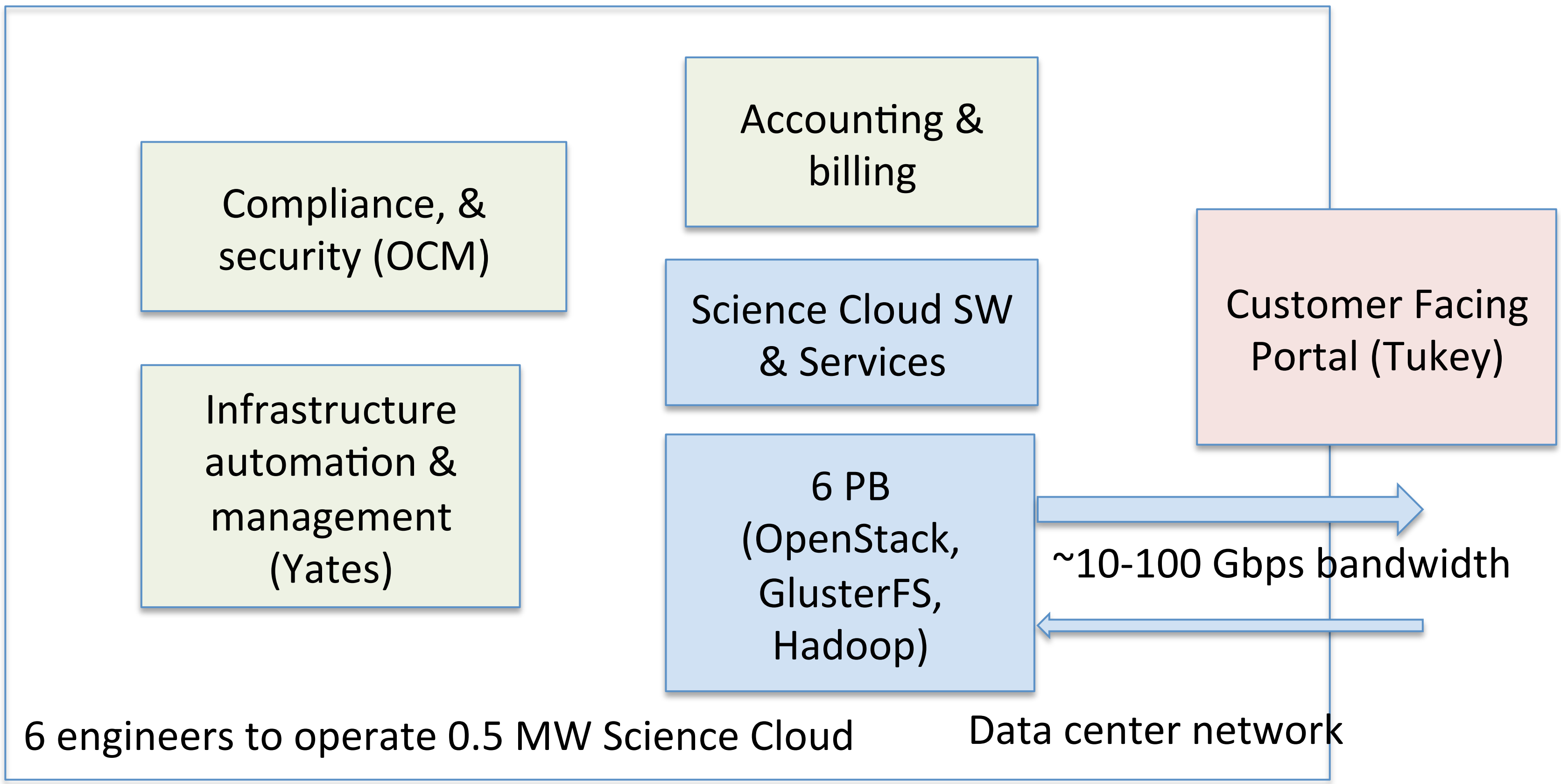
Source: OSDC container. Part of the OSDC Project Matsu and OSDC Testbed (2010-2014).

Open Science Data Cloud

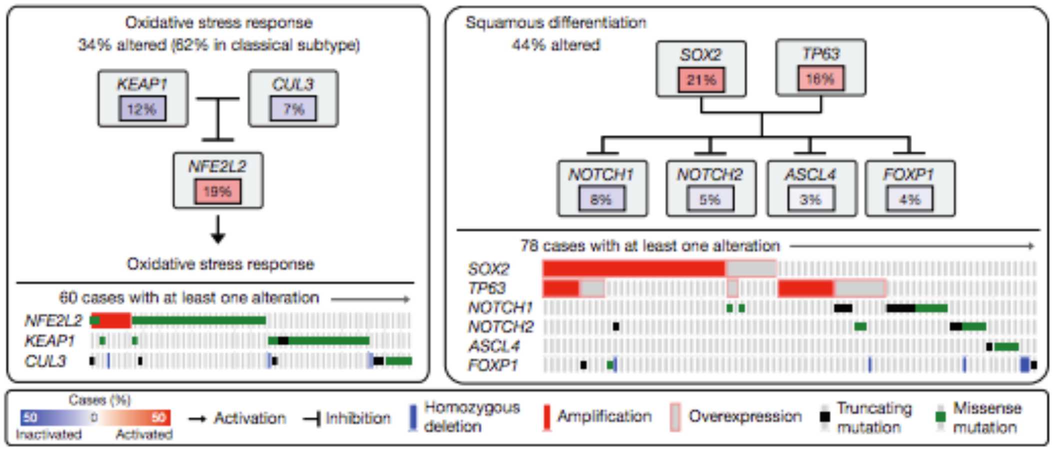


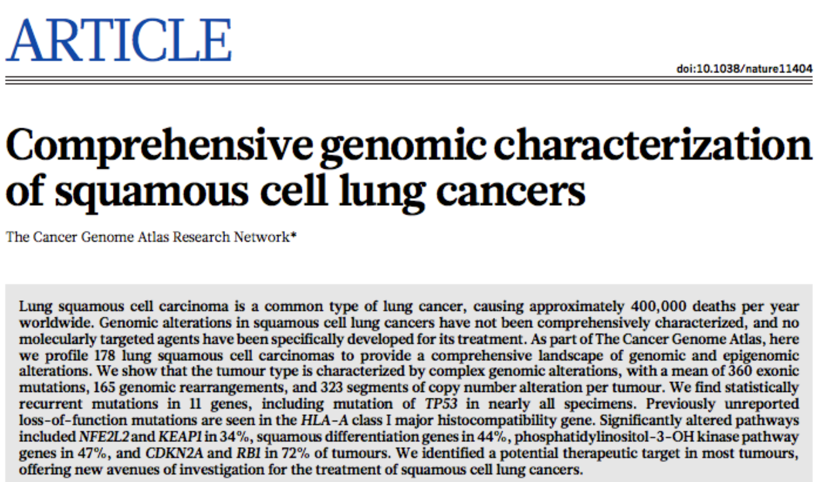
OPEN CLOUD CONSORTIUM

Open Science Data Cloud (Home of Bionimbus)



TCGA Analysis of Lung Cancer





ARTICLE

doi:10.1038/nature11404

Comprehensive genomic characterization of squamous cell lung cancers

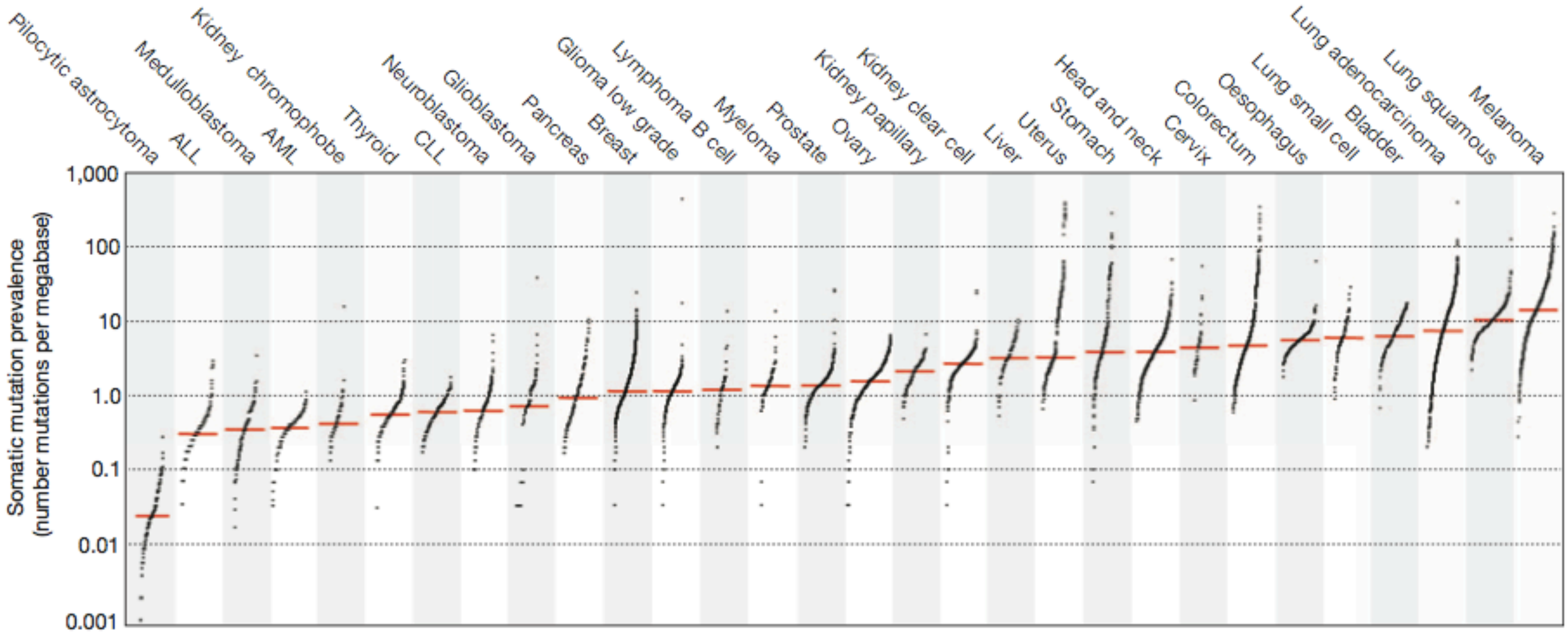
The Cancer Genome Atlas Research Network*

Lung squamous cell carcinoma is a common type of lung cancer, causing approximately 400,000 deaths per year worldwide. Genomic alterations in squamous cell lung cancers have not been comprehensively characterized, and no molecularly targeted agents have been specifically developed for its treatment. As part of The Cancer Genome Atlas, here we profile 178 lung squamous cell carcinomas to provide a comprehensive landscape of genomic and epigenomic alterations. We show that the tumour type is characterized by complex genomic alterations, with a mean of 360 exonic mutations, 165 genomic rearrangements, and 323 segments of copy number alteration per tumour. We find statistically recurrent mutations in 11 genes, including mutation of *TP53* in nearly all specimens. Previously unreported loss-of-function mutations are seen in the *HLA-A* class I major histocompatibility gene. Significantly altered pathways included *NFE2L2* and *KEAP1* in 34%, squamous differentiation genes in 44%, phosphatidylinositol-3-OH kinase pathway genes in 47%, and *CDKN2A* and *RB1* in 72% of tumours. We identified a potential therapeutic target in most tumours, offering new avenues of investigation for the treatment of squamous cell lung cancers.

- 178 cases of SQCC (lung cancer)
- Matched tumor & normal
- Mean of 360 exonic mutations, 323 CNV, & 165 rearrangements per tumor
- Tumors also vary spatially and temporally.

Source: The Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers, Nature, 2012, doi:10.1038/nature11404.

Number of Mutations by Cancer Type



Source: Michael S. Lawrence, Petar Stojanov, Paz Polak, et. al., Mutational heterogeneity in cancer and the search for new cancer-associated genes, Nature 449, pages 214-218, 2013.

The Cancer Genome Atlas (TCGA)

- Targeting 20 cancers x 500+ patients
- 1.2PB of data today, growing to 2.5 PB

Analyzing Data From
The Cancer Genome Atlas (TCGA)

Current Practice

1. Apply to dbGaP for access to data.
2. Hire staff, set up and operate secure compliant computing environment to mange 10 – 100+ TB of data.
3. Get environment approved by your research center.
4. Setup analysis pipelines.
5. Download data from CG-Hub (takes days to weeks).
6. Begin analysis.

PDC Console Apply Status

BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

The Bionimbus Protected Data Cloud (PDC) is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB,) the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The PDC allows users authorized by NIH to compute over human genomic data from dbGaP in a secure compliant fashion. Currently, selected datasets from the The Cancer Genome Atlas (TCGA) are available in the PDC.

How can I get involved?

- Apply for an Bionimbus PDC account and use the Bionimbus PDC to manage, analyze and share your data.
- Partner with us and add your own racks to the Bionimbus PDC (we will manage them for you).
- Help us develop the open source Bionimbus PDC software stack.

You can contact us at info@opencloudconsortium.org.

How do I get started?

First, apply for an account. Once your account is approved, you can login to the console and get started. Support questions can be directed to support@opencloudconsortium.org.

Apply for the PDC Now

Login to the PDC Console

OPEN CLOUD CONSORTIUM | OPEN SCIENCE DATA CLOUD | Institute for Genomics & Systems Biology | THE UNIVERSITY OF CHICAGO and our sponsors.

bionimbus.opensciencedatacloud.org

Analyzing Data From The Cancer Genome Atlas (TCGA)

Current Practice

1. Apply to dbGaP for access to data.
2. Hire staff, set up and operate secure compliant computing environment to mange 10 – 100+ TB of data.
3. Get environment approved by your research center.
4. Setup analysis pipelines.
5. Download data from CG-Hub (takes days to weeks).
6. Begin analysis.

With Bionimbus Protected Data Cloud (PDC)

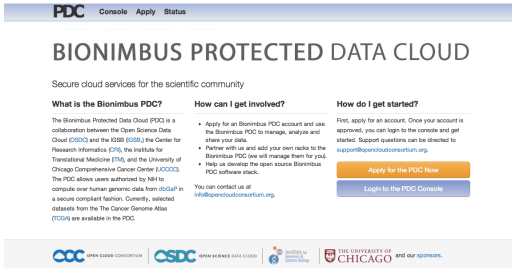
1. Apply to dbGaP for access to data.
2. Use your existing NIH grant credentials to login to the PDC, select the data that you want to analyze, and the pipelines that you want to use.
3. Begin analysis.

Genomic Data Commons (GDC)

- Will host human genomic data from several large NIH/NCI-funded projects, including:
  - The Cancer Genome Atlas (TCGA)
  - TARGET (pediatric cancer genomic data)
  - Selected other current and planned NCI genomics projects

A Key Question

- Is biomedical computing at the scale of a data center important enough for the research community to do or do we *only* outsource to commercial cloud service providers (certainly we will interoperate with commercial cloud service providers)?



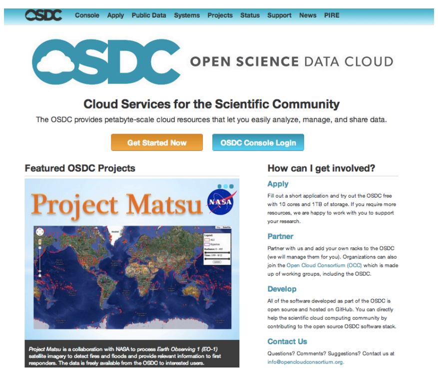
Biomedical Commons Cloud

Bionimbus Protected Data Cloud

- University of Chicago
- 1 PB (TCGA)
- 2013

Part of / interoperates with

- Not for Profit Open Cloud Consortium
- Involves multiple organizations and cancer centers
- 2014 / 2015 launch.



Open Science Data Cloud

- Open Cloud Consortium
- 5 PB (pan science)
- 2009

Same open source software stack.

Working with the Bionimbus Commons Community - Clouds

- If you have a biomedical cloud, consider interoperating it with the OCC Biomedical Commons Cloud.
- Design and test standards so that Biomedical Commons Clouds can interoperate:
  - Data synchronization between two biomedical clouds
  - APIs to access data (e.g. NCBI SRA Toolkit)
  - Restful queries (e.g. Genomespace)
  - Scattering queries, gathering the results
  - Coordinated analysis

Genomic Clouds

Bionimbus Protected Data Cloud (Chicago)

OSDC OPEN SCIENCE DATA CLOUD

Embassy (EBI)

GenomeBridge

(Cambridge)

Cancer Genomics Hub (Santa Cruz)

DNAnexus

(Mountain View)

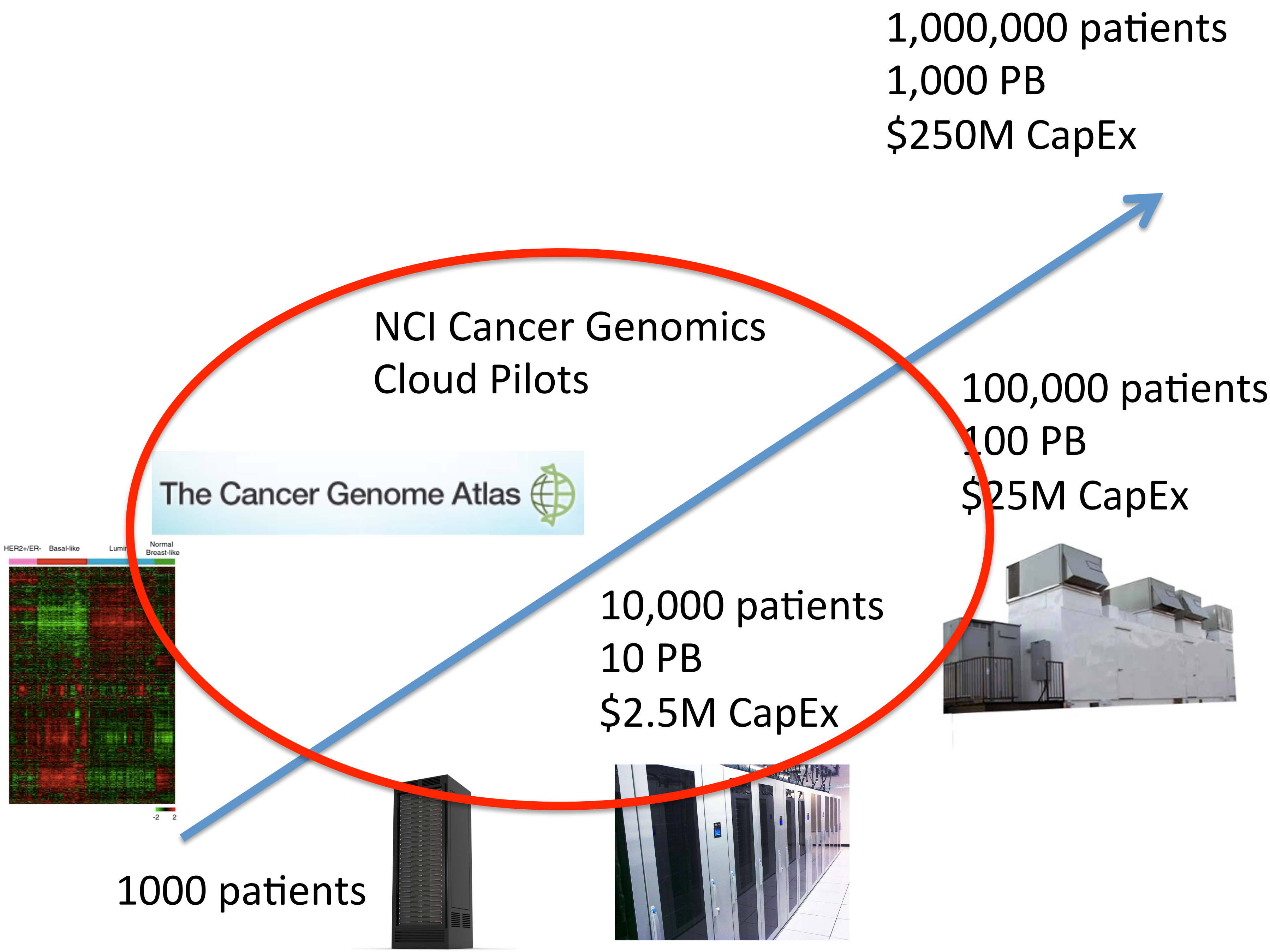
Cancer Genome Collaboratory (Toronto)

1. What scale is required for biomedical clouds?
2. What is the design for biomedical clouds?
3. What tools and applications do users need to make discoveries in large amounts of biomedical data?
4. How do different biomedical clouds interoperate?

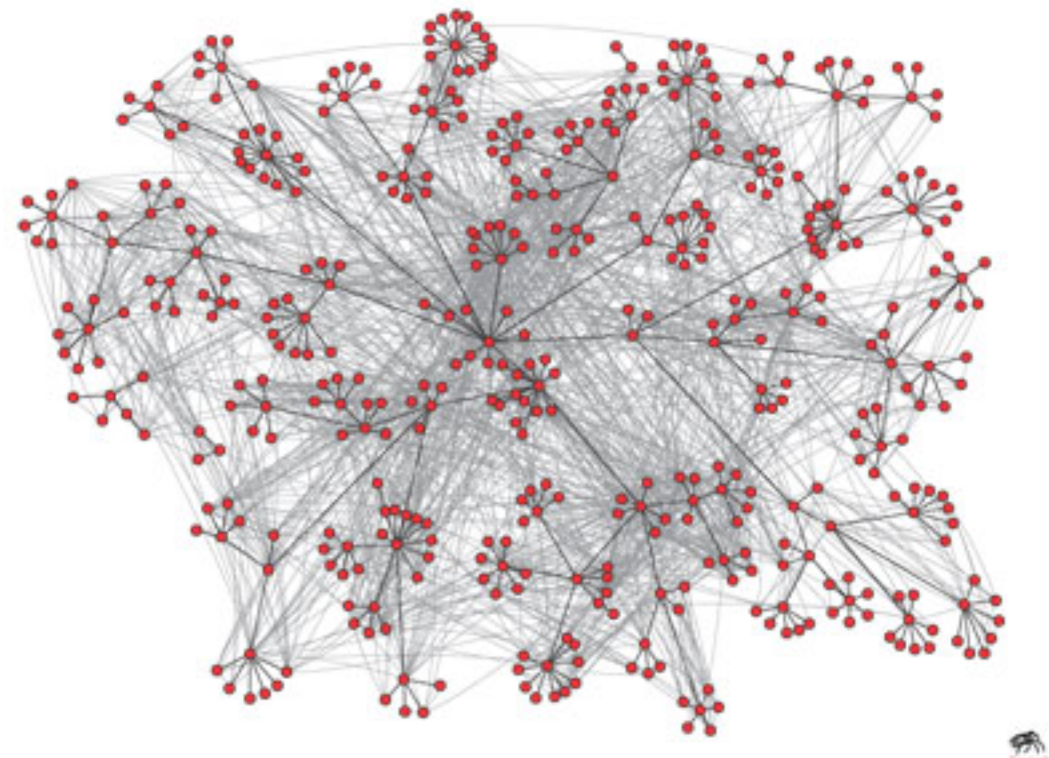
Some Biomedical Data Commons Guidelines for the Next Five Years

- There is a societal benefit when biomedical data is also available in data commons operated by the research community (vs sold exclusively as data products by commercial entities or only offered for download by the USG).
- Large data commons providers should peer.
- Data commons providers should develop standards for interoperating.
- Standards should not be developed ahead of open source reference implementations.
- We need a period of experimentation as we develop the best technology and practices.
- The details are hard (consent, scalable APIs, open vs controlled access, sustainability, security, etc.)

2005 - 2015	**Bioinformatics tools & their integration.** Examples: Galaxy, GenomeSpace, workflow systems, portals, etc.
2010 - 2020	**Data center scale science.** Examples: Bionimbus/OSDC, CG Hub, Cancer Collaboratory, GenomeBridge, etc.
2015 - 2025	**???**

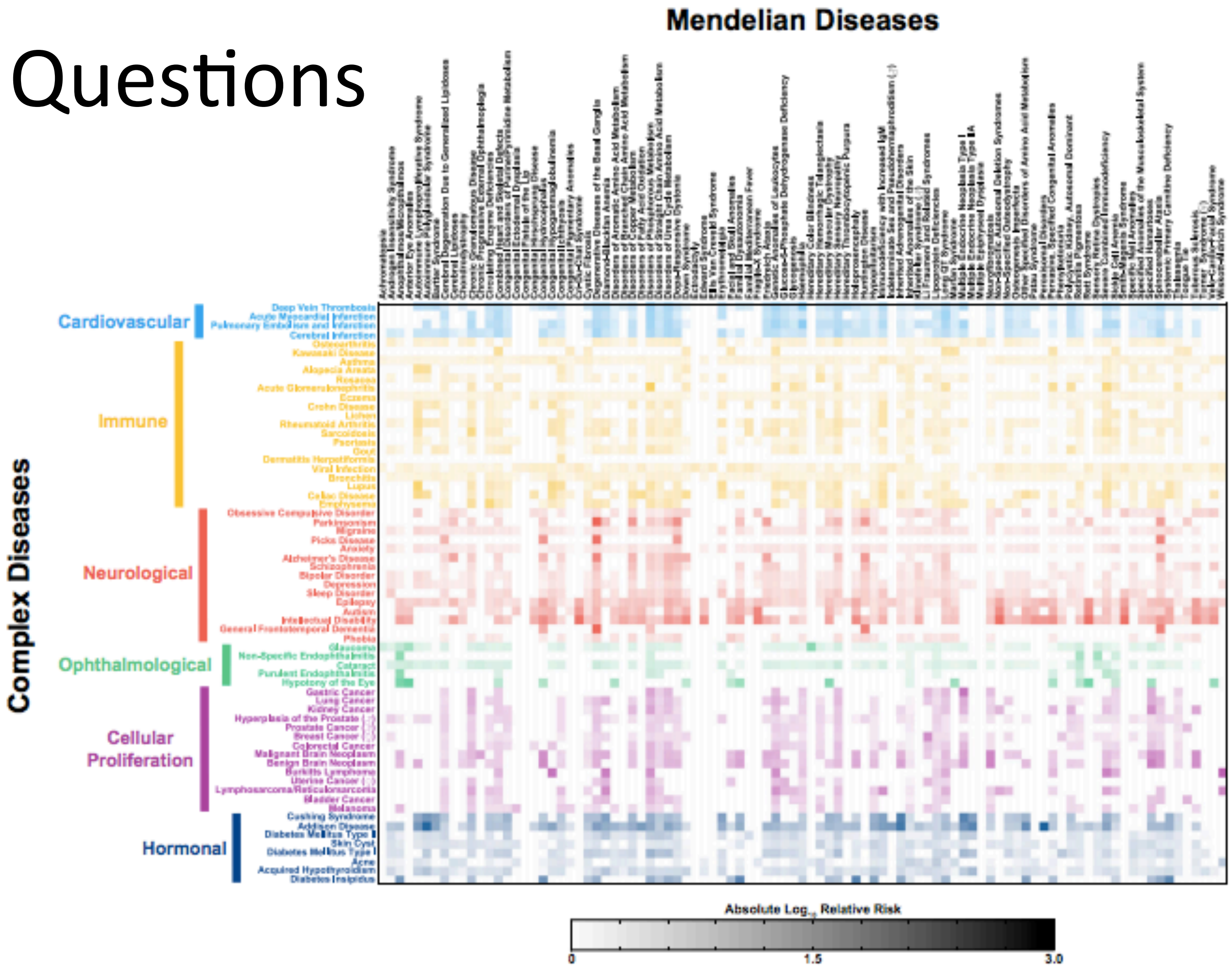

1,000,000 patients
1,000 PB
$250M CapEx
NCI Cancer Genomics
Cloud Pilots
100,000 patients
100 PB
$25M CapEx
The Cancer Genome Atlas
HER2+/ER-
Basal-like
Normal
Breast-like
10,000 patients
10 PB
$2.5M CapEx
1000 patients

Part 3:
Analyzing Biomedical Data at the Scale of a Data Center



Source: Jon Kleinberg, Cornell University, www.cs.cornell.edu/home/kleinber/networks-book/

Research Questions



Source: David R. Blair, Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V. Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H. Shah, Robert L. Grossman, Nancy J. Cox, Kevin P. White, Andrey Rzhetsky, A Non-Degenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk, Cell, September, 2013

Building Models over Big Data

- We know about the "unreasonable effectiveness of ensemble models." Building ensembles of models over computer clusters works well …
- … but, how do machine learning algorithms scale to data center scale science?
- Ensembles of random trees built from templates appear to work better than traditional ensembles of classifiers
- The challenge is often decomposing large heterogeneous datasets into homogeneous components that can be modeled.

Source: Wenxuan Gao, Robert Grossman, Philip Yu, and Yunhong Gu, Why Naive Ensembles Do Not Work in Cloud Computing, Proceedings of the The First Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2009), 2009

New Questions

- How would research be impacted if we could analyze *all of the data* each evening?
- How would health care be impacted if we could *analyze of the data* each evening?

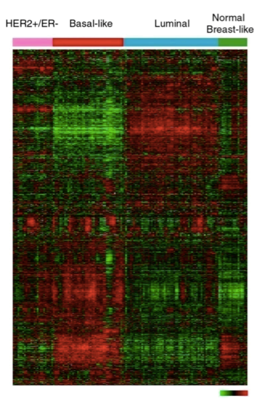
- What are the key common services & APIs?
- How do the biomedical commons clouds interoperate?
- What is the governance structure?
- What is the sustainability model?

1,000,000 patients
1,000 PB

100,000 patients
100 PB

International Cancer Genome Consortium

The Cancer Genome Atlas



10,000 patients
10 PB

1000 patients

2005 - 2015	**Bioinformatics tools & their integration.** Examples: Galaxy, GenomeSpace, workflow systems, portals, etc.
2010 - 2020	**Data center scale science.** Interoperability and preservation/peering/portability of large biomedical datasets. Examples: Bionimbus/OSDC, CG Hub, Cancer Collaboratory, GenomeBridge, etc.
2015 - 2025	**New modeling techniques.** The discovery of new & emergent behavior at scale. Examples: What are the foundations? Is more different?

Thanks To My Colleagues & Collaborators

- Kevin White
- Nancy Cox
- Andrey Rzhetsky
- Lincoln Stein
- Barbara Stranger

Thanks to My Lab

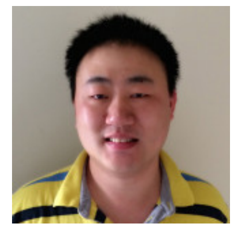
Allison Heath

Maria Patterson

Sean Sullivan

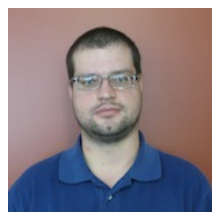
Rafael Suarez

Jonathan Spring



Zhenyu Zhang

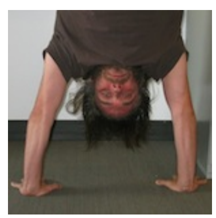
Matt Greenway



Ray Powell

Stuti Agrawal

Renuka Ayra



David Hanley

Thanks to the White Lab

Megan
McNerney

Chaitanya
Bandlamidi

Jason
Grundstad

Jason Pitt

Questions?

For more information

- www.opensciencedatacloud.org
- For more information and background, see Robert L. Grossman and Kevin P. White, A Vision for a Biomedical Cloud, Journal of Internal Medicine, Volume 271, Number 2, pages 122-130, 2012.
- You can find some more information on my blog:

  rgrossman.com.
- My email address is robert.grossman at uchicago dot edu.

OPEN CLOUD CONSORTIUM

Institute for Genomics & Systems Biology

Computation Institute

THE UNIVERSITY OF CHICAGO Center for Research Informatics

Major funding and support for the Open Science Data Cloud (OSDC) is provided by the Gordon and Betty Moore Foundation. This funding is used to support the OSDC-Adler, Sullivan and Root facilities.

Additional funding for the OSDC has been provided by the following sponsors:

- The Bionimbus Protected Data Cloud is supported in by part by NIH/NCI through NIH/SAIC Contract 13XS021 / HHSN261200800001E.
- The OCC-Y Hadoop Cluster (approximately 1000 cores and 1 PB of storage) was donated by Yahoo! in 2011.
- Cisco provides the OSDC access to the Cisco C-Wave, which connects OSDC data centers with 10 Gbps wide area networks.
- The OSDC is supported by a 5-year (2010-2016) PIRE award (OISE – 1129076) to train scientists to use the OSDC and to further develop the underlying technology.
- OSDC technology for high performance data transport is support in part by NSF Award 1127316.
- The StarLight Facility in Chicago enables the OSDC to connect to over 30 high performance research networks around the world at 10 Gbps or higher.
- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NIH or other funders of this research.

The OSDC is managed by the Open Cloud Consortium, a 501(c)(3) not-for-profit corporation. If you are interested in providing funding or donating equipment or services, please contact us at info@opensciencedatacloud.org.