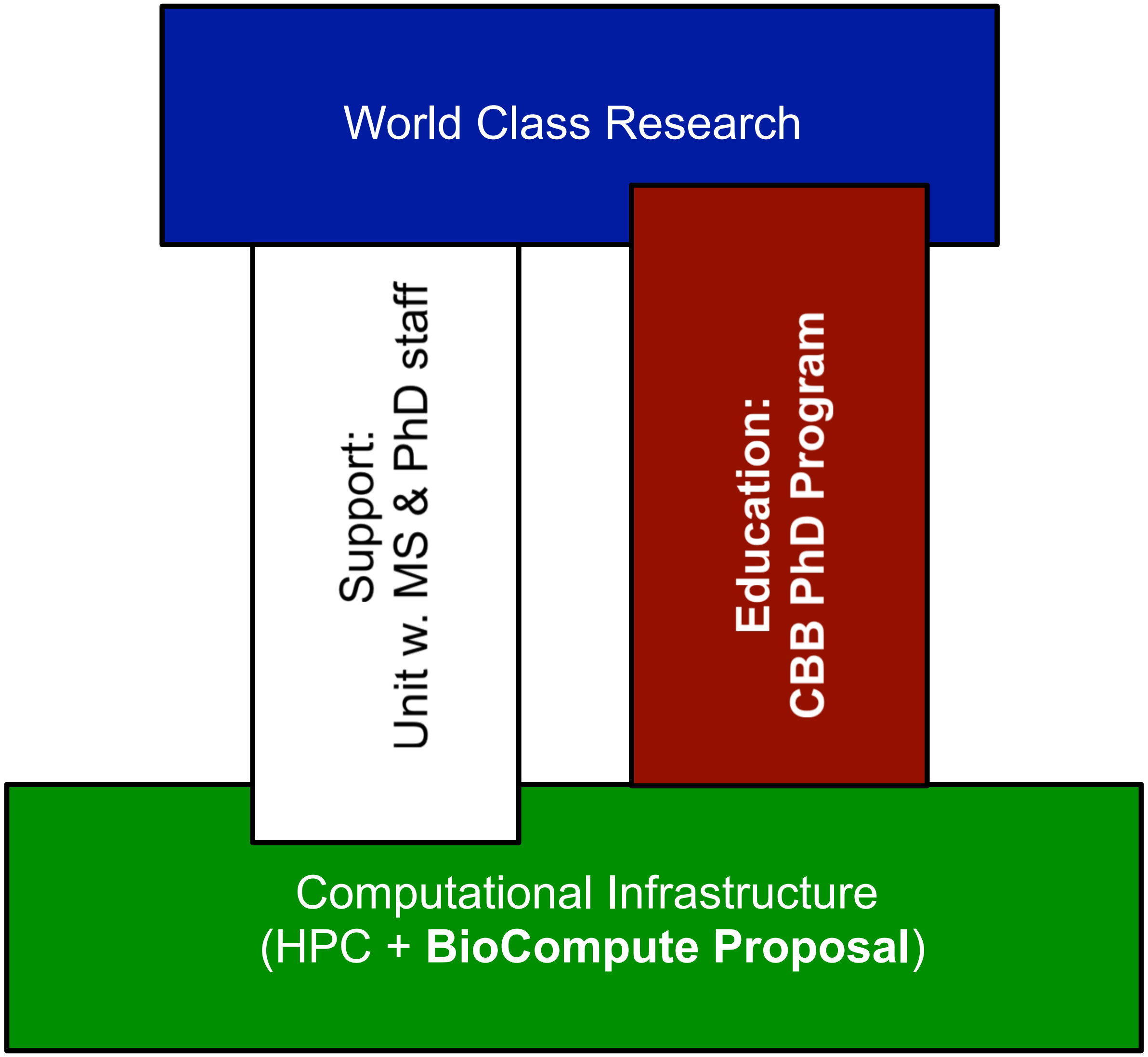
Thoughts on Computational Biology at Yale Related to Research, Education & Infrastructure

Mark Gerstein

Computational Biology at Yale



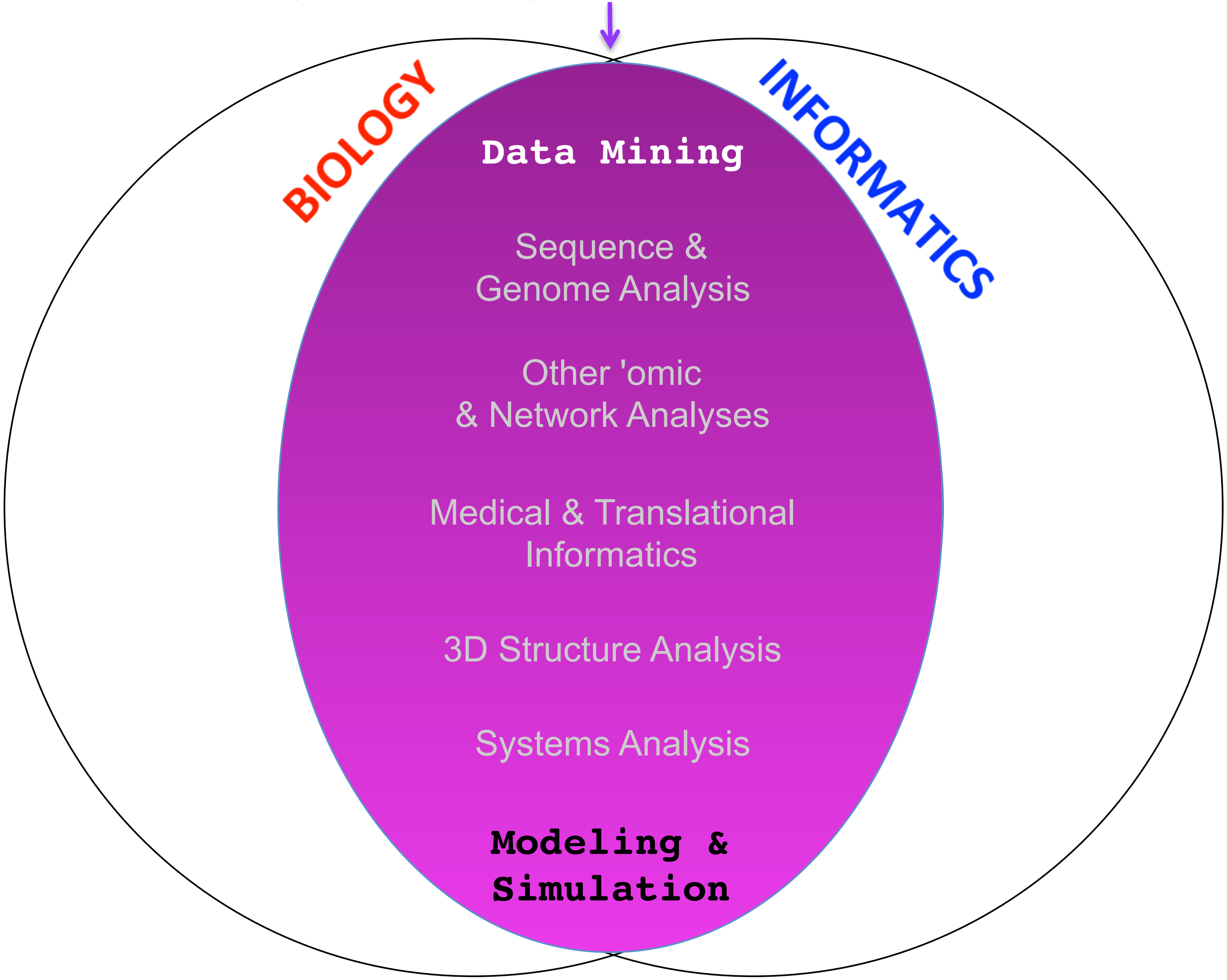
World Class Research

Support:
Unit w. MS & PhD staff

Education:
CBB PhD Program

Computational Infrastructure
(HPC + **BioCompute Proposal**)

(Molecular) BIOINFORMATICS



BIOLOGY

INFORMATICS

Data Mining

Sequence &
Genome Analysis

Other 'omic
& Network Analyses

Medical & Translational
Informatics

3D Structure Analysis

Systems Analysis

**Modeling &
Simulation**

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as CS, stats & physics) to **organize, analyze, model & understand the information associated** with these molecules, **on a large-scale**.
- Bioinformatics is a practical discipline with many **applications**.

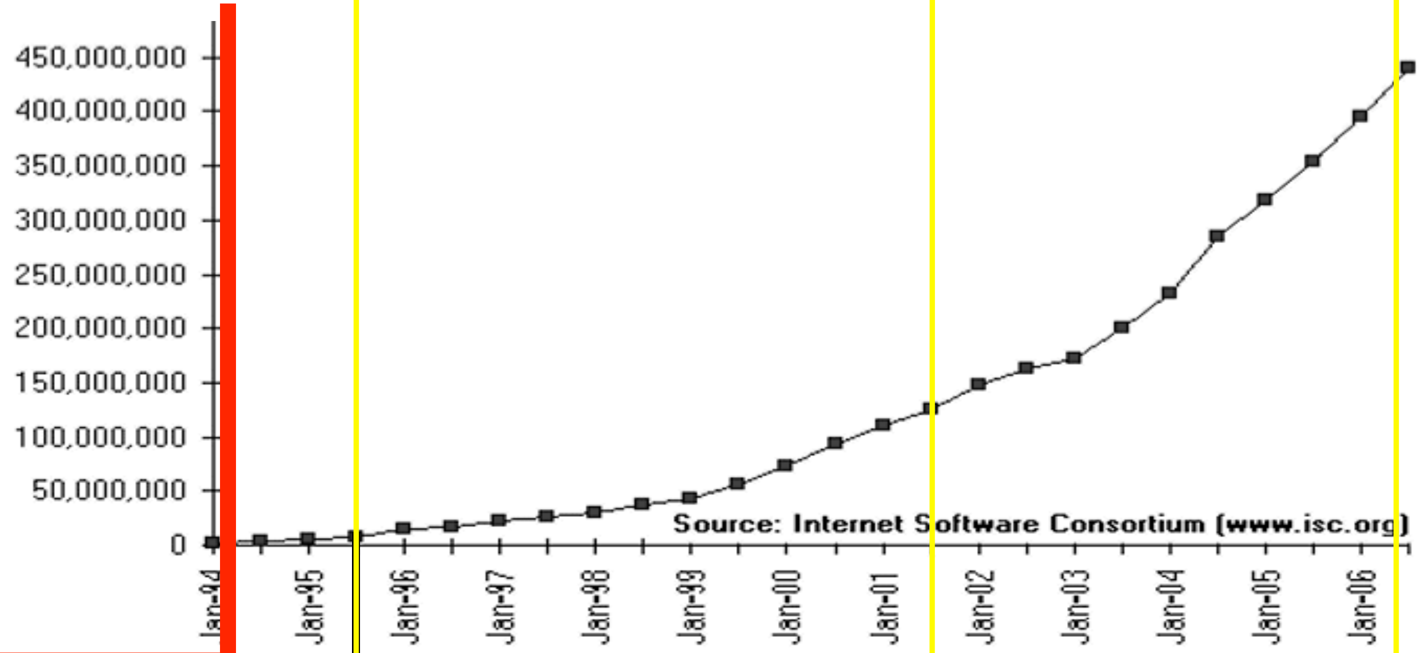
[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

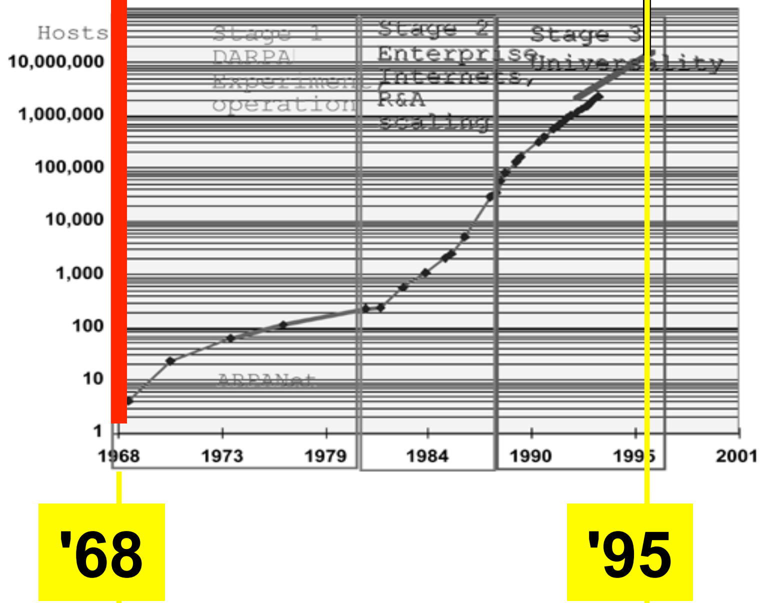
What **Information** to Organize?

- **Sequences** (DNA & Protein)
- 3D Structures
- Network & Pathway Connectivity
- Phylogenetic tree relationships
- Large-scale gene expression & functional genomics data
- Phenotypic data & medical records....

Internet Hosts

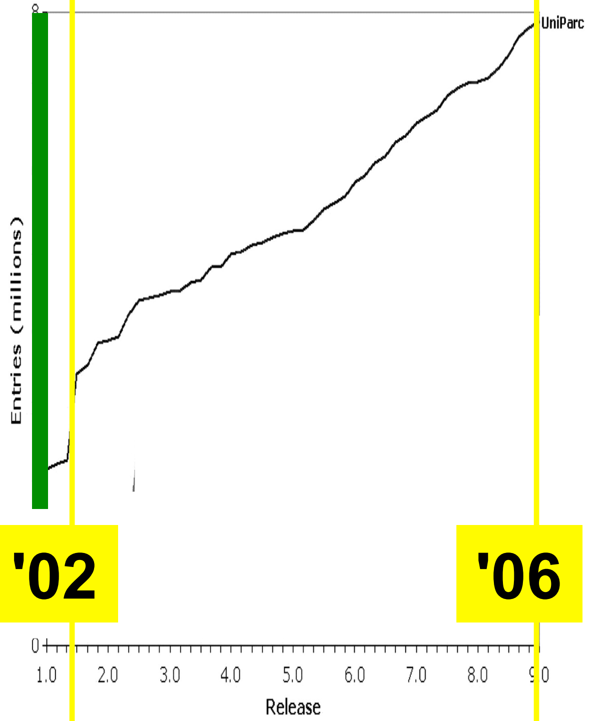
(adapted from D Brutlag, Stanford & http://navigators.com/stats.html)



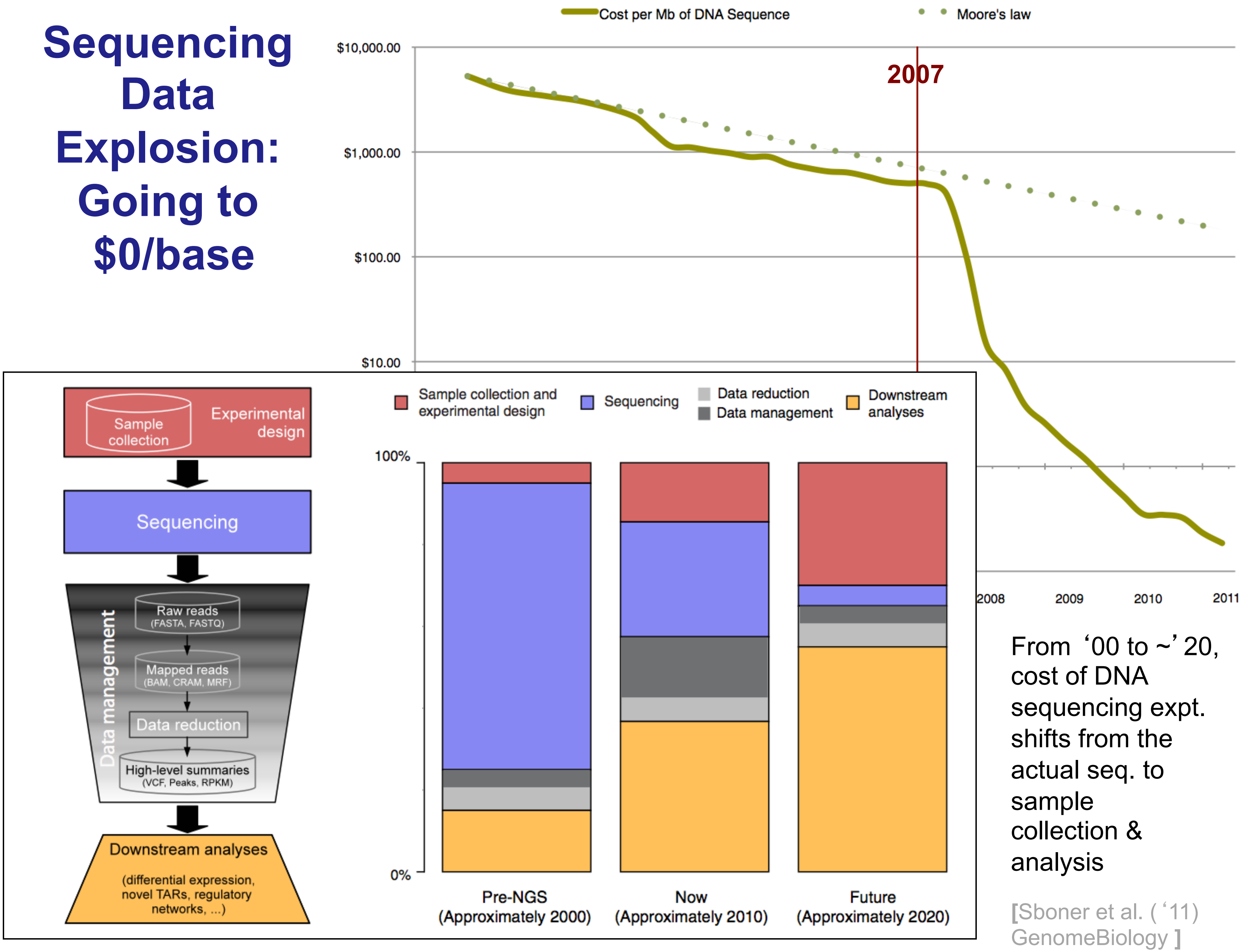


Proteins

Suzek, B. E. et al. Bioinformatics 2007 23:1282-1288; doi:10.1093/bioinformatics/btm098



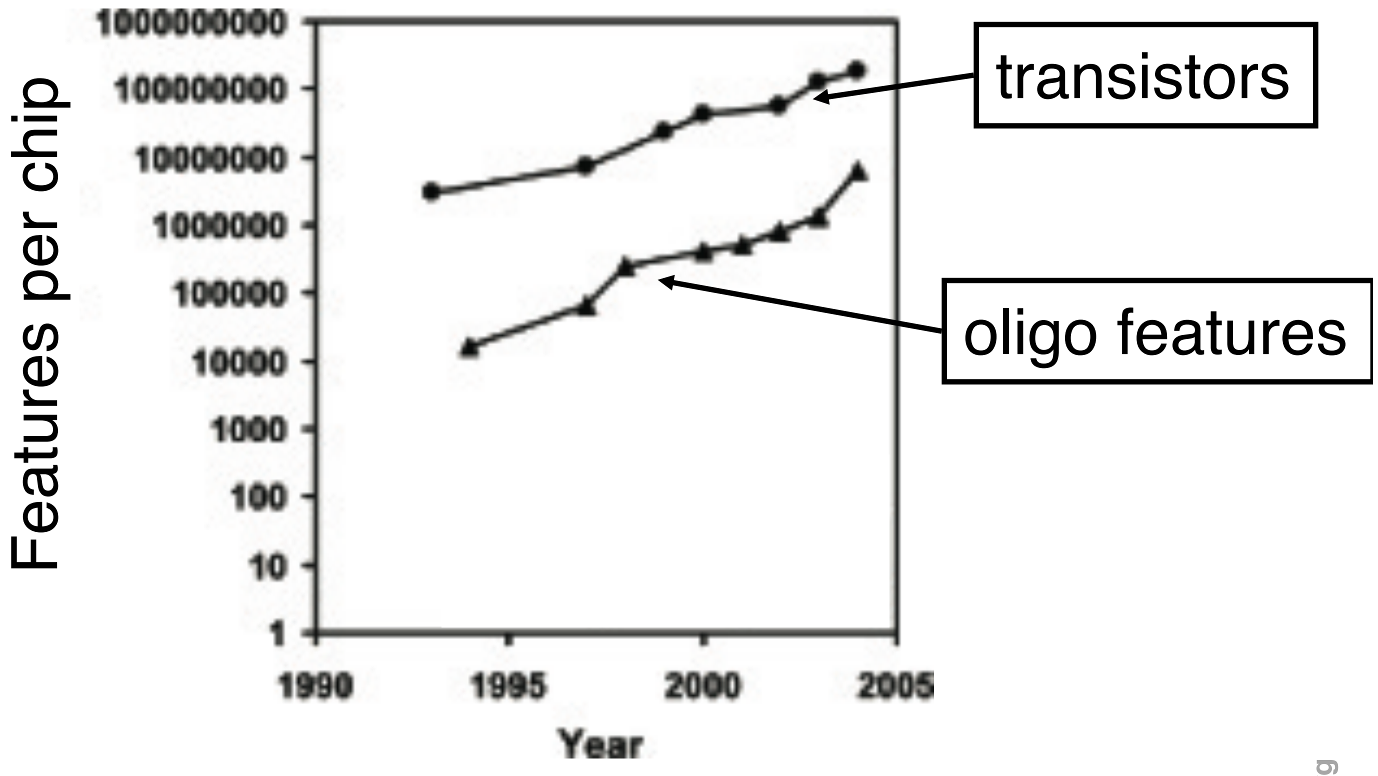
Sequencing Data Explosion: Going to $0/base



From ‘00 to ~’ 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11) GenomeBiology]

Features per Slide



Chip Technology

General Types of “Informatics” techniques in Computational Biology

- Databases
  - Building, Querying
  - Representing Complex data
- Data mining
  - Machine Learning techniques
  - Clustering & Tree construction
  - Rapid Text String Comparison & textmining
  - Detailed statistics of significance & association
- Network Analysis
  - Analysis of Topology (eg Hubs)
  - Predicting Connectivity
- Structure Analysis & Geometry
  - Graphics (Surfaces, Volumes)
  - Comparison & 3D Matching (Vision, recognition, docking)
- Physical Modeling
  - Newtonian Mechanics
  - Electrostatics
  - Numerical Algorithms
  - Simulation
  - Modeling Chemical Reactions & Cellular Processes

Defining the Boundaries of the Field

(Determining the "Support Vectors")

Are They or Aren't They Comp. Bio.? (#1)

- (Digital Libraries & Medical Record Analysis
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- (Motif Discovery Using Gibb's Sampling
- (Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - (Distance Geometry
- (Metabolic Pathway Simulation
- (The DNA Computer

Are They or Aren't They Comp. Bio.? (#1, `Answers`)

- **(YES?)** Digital Libraries & Medical Record Analysis
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- **(YES)** Motif Discovery Using Gibb's Sampling
- **(NO?)** Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - **(YES)** Distance Geometry
- **(YES)** Metabolic Pathway Simulation
- **(NO)** The DNA Computer

Are They or Aren't They Comp. Bio.? (#2)

- (Gene identification by sequence characteristics
  - ◊ Prediction of splice sites
- (DNA methods in forensics
- (Modeling of Populations of Organisms
  - ◊ Ecological Modeling (predator & prey)
- (Modeling the nervous system
  - ◊ Computational neuroscience
  - ◊ Understanding how brains think & using this to make a better computer
- (Molecular phenotype discovery – looking for gene expression signatures of cancer
  - ◊ What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#2, Answers)

- **(YES)** Gene identification by sequence characteristics
  - ◊ Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  - ◊ Ecological Modeling (predator & prey)
- **(NO?)** Modeling the nervous system
  - ◊ Computational neuroscience
  - ◊ Understanding how brains think & using this to make a better computer
- **(YES)** Molecular phenotype discovery – looking for gene expression signatures of cancer
  - ◊ What if it included non-molecular data such as age ?

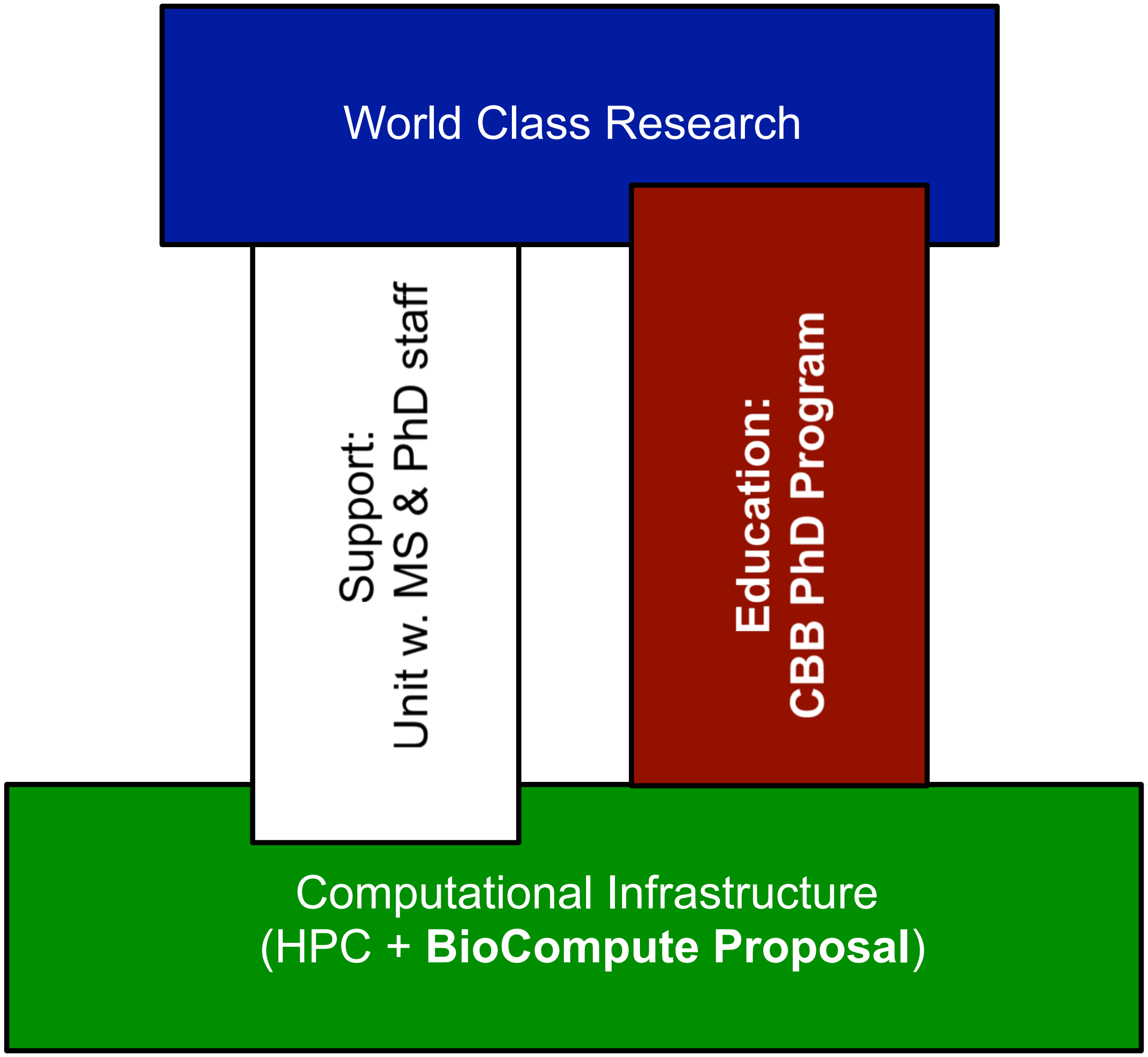
Are They or Aren’t They Comp. Bio.? (#3)

- (RNA structure prediction
- (Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- (Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ (Genetic Algorithms in molecular biology
- (Homology Modeling & Drug Docking
- (Char. drugs & other small molecules (QSAR)
- (Computerized Diagnosis based on Pedigrees
- (Processing of NextGen sequencing image files
- (Module finding in protein networks

Are They or Aren't They Comp. Bio.? (#3, Answers)

- (YES) RNA structure prediction
- (NO) Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- (NO) Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ (NO?) Genetic Algorithms in molecular biology
- (YES) Homology Modeling & Drug Docking
- (YES) Char. drugs & other small molecules (QSAR)
- (NO) Computerized Diagnosis based on Pedigrees
- (NO) Processing of NextGen sequencing image files
- (YES) Module finding in protein networks

Computational Biology at Yale



World Class Research

Support:
Unit w. MS & PhD staff

Education:
CBB PhD Program

Computational Infrastructure
(HPC + **BioCompute Proposal**)

History & Current Structure of PhD Program

- History
  - Started in '02 1st as BBS track & in '03 then as a PhD granting program
  - by M Gerstein & P Miller
  - split betw. **Med School** & **Sci Hill**
- Curr. Structure
  - co-DGSes M Gerstein [MB&B & CS] & H Zhao [Public Health, Genetics & Stats]
  - DGAs (M Krauthammer & C O'Hern)

- Key Numbers
  - 77 matriculated, 34 graduated so far
  - 3 in PEB
  - ~7 students/yr (~40% non-US)

Inputs

- CBB Graduates – Undergrad Majors

Biology	**Bioinformatics**	**Informatics**	**Other**
19	3	15	5

- CBB Current Students – Undergrad Majors

Biology	**Bioinformatics**	**Informatics**	**Other**
18	8	8	1

- Admissions
  - '14 numbers
    XXX131162 % US accepted,
    XXX131162 % foreign accepted,
    XXX131162 % of the accepts come
- XXXXXXX – See Shadow

Curriculum: Courses & Competency in **Core CBB**, **Biological Sciences** & **Informatics**

- 10 Courses in
  Three Core Areas of Competency
  - Computational Biology & Bioinformatics (3 grad courses)
    - CBB 752b Bioinformatics: Practical Applications of Simulation & Data Mining **[18yrs!]**
    - CBB 740a Clinical and Translational Informatics
    - CBB 562a Dynamical Systems in Biology
  - Biological sciences (2 grad courses)
  - Informatics - e.g., CS, stats, app. math (2 grad courses)
  - Electives (2 undergrad or grad courses, in any of the above)

- Competency of incoming students (need to take courses to get to this level)
  - Biology & Natural Science: introductory biology, biochemistry, chemistry
  - CS: introduction to CS, data structures & programming techniques
  - Math & Stat: introduction to probability and statistical inference, multivariate calculus and linear algebra

[More detail in Gerstein et al. ('07) *J Biomed. Inf.*]

Students studying over whole campus

Labs of CBB students (incl. rotations) (*=PhD advisor, incl. jt.)

Location	Faculty
Science Hill	L Regan*, T Emonet*, A Pyle*, M Gerstein*, J Chang, C O'Hern*, W Jorgensen*, A Silberschatz, R Coifman, S Zucker*, F Isaacs, K Miller-Jensen, S Mochrie, S Dellaporta*, J Townsend, J Zhang, G Brudvig, V Batista, A Schepartz, E Yan, A Phillips*, J Peccia*, C Wilson, F Slack*, M Snyder*, A Miranker
West Campus/ VA	M Acar*, A Justice*, G Wagner*, J Gelernter*, A Levchenko, C Jacobs-Wagner
Med. School	M Krauthammer*, S Kleinstein*, Y Kluger*, H Zhao*, F Crawford*, D Stern*, J Noonan*, K Kidd*, V Reinke, M Günel*, H Lin*, K Cheung*, L Pusztai*, C Brandt, C Cotsapas, M Crair, D Hafler, R Lifton, S Ma, S Weissman, M Bosenberg*, J Lu*, M State*, J Cho*, TH Kim*, D Tuck*, R Flavell, P Lizardi*, P Miller*, A Molinaro*, M White*, W Shlomchik

Program is doing well from Grad. Sch. Surveys & Rankings

XXXXXXX – See Shadow

Program is doing well from Grad. Sch. Surveys & Rankings

Outputs

- Over last 7 yrs
- Some faculty; many in industry, split betw. **traditional bioinfo. route in biotech/ pharma** & more general **"data-science" business** positions

Fac.	2003-2007	Assoc Professor, ASU
	2002-2007	Asst Professor, UT
	2005-2010	UCLA Lecturer
	2009-2014	Asst Professor, UNC
	2006-2012	Assoc Bioinformatics Scientist , Children's Hospital of Philadelphia
Postdoc	2002-2008	Postdoc, Stanford University
	2002-2009	Postdoc, Dana Farber Institute
	2004-2010	Resident in General Surgery, Yale
	2007-2012	Computational Biologist, Broad Institute, MA
	2007-2012	Postdoc, Stanford University
	2008-2013	Postdoc, Stanford University
	2006-2013	Programmer Anaylst II, Yale University
Industry	2002-2007	Sr. Bioinformatics Scientist, Illumina
	2004-2009	Data Integration Officer, St. Jude, Memphis
	2003-2010	Scientist, Celgene
	2004-2010	Quantitative Trader, Laurion Capital Mgt
	2005-2010	Director of Informatics, Bina Technologies Inc.
	2005-2010	Investigator, Novartis Institutes for BioMedical Research
	2004-2010	Sr. Developer, Schrodinger, Inc.
	2006-2011	Assoc Principal Scientist, Merck Company
	2005-2011	Product Manager & Bioinformatics Analyst, 5AM Solutions
	2005-2011	Financial firm in Beijing
	2006-2011	Quantitative Analyst, Google
	2005-2011	Data Analyst/NLP Specialist, Elsevier
	2007-2012	Lead Bioinformatics R&D Developer, Regeneron Pharmaceuticals Inc.
	2006-2012	Software Developer, Berkeley Nat Lab
	2009-2012	Information Technology and Services, Germany
	2008-2013	Economic Modeling Senior, Freddie Mac
	2007-2013	Analytics Consultant, SeqWise Next Generation Sequencing Consulting
	2008-2014	Research Scientist, GE Global Research
	2008-2014	Bioinformatics Scientist, Illumina
	2009-2014	Senior Consulting Engineer, Attivio, Inc.

Bigger Output Dataset (MG lab since '97)

Of 25 faculty positions split betw. **bio**, **cs** & **bioinfo** & later incr.

Faculty

<= postdocs

1999 – 2002	Johns Hopkins
1999 – 2004	McGill U
1999 – 2002	Yale
2000 – 2004	Univ. College London
2002 – 2004	U of Toronto
2003 – 2005	Miami U.
2003 – 2006	McGill U
2003 – 2006	Cincinnati Children's Hospital
2003 – 2005	Royal Inst. of Technology, Sweden
2003 – 2007	Albert Einstein College of Medicine
2003 – 2005	U of London
2004 – 2008	U of Toronto
2005 – 2010	Albert Einstein College of Medicine
2005 – 2007	EMBL
2006 – 2011	Cornell Medical School
2008 – 2011	Tsinghua University
2008 – 2012	Dartmouth University
2008 – 2014	Mayo Clinic/U of Minnesota
2008 – 2014	Weill Cornell Medical College
2007 – 2014	NYU (Shanghai)

PhD students=>

1998 – 2005	EBI (Cambridge)
2000 – 2005	Cornell U
2004 – 2007	Uppsala U
2004 – 2009	CUHK

Industry

<= postdocs

1998 – 2004	Goldman Sachs
2000 – 2002	Incyte
2000 – 2003	Sigma-Aldrich
2002 – 2004	ExxonMobil
2002 – 2004	Genelogic
2002 – 2004	McKinsey Consulting
2002 – 2005	UCB Pharma
2003 – 2006	McKinsey Consulting
2005 – 2006	Glaxosmithkline
2005 – 2007	British Telecom
2005 – 2009	Quantitative consulting & writing
2007 – 2011	BASF
2011 – 2012	NEC
2013 – 2014	BioMarin Pharmaceutical

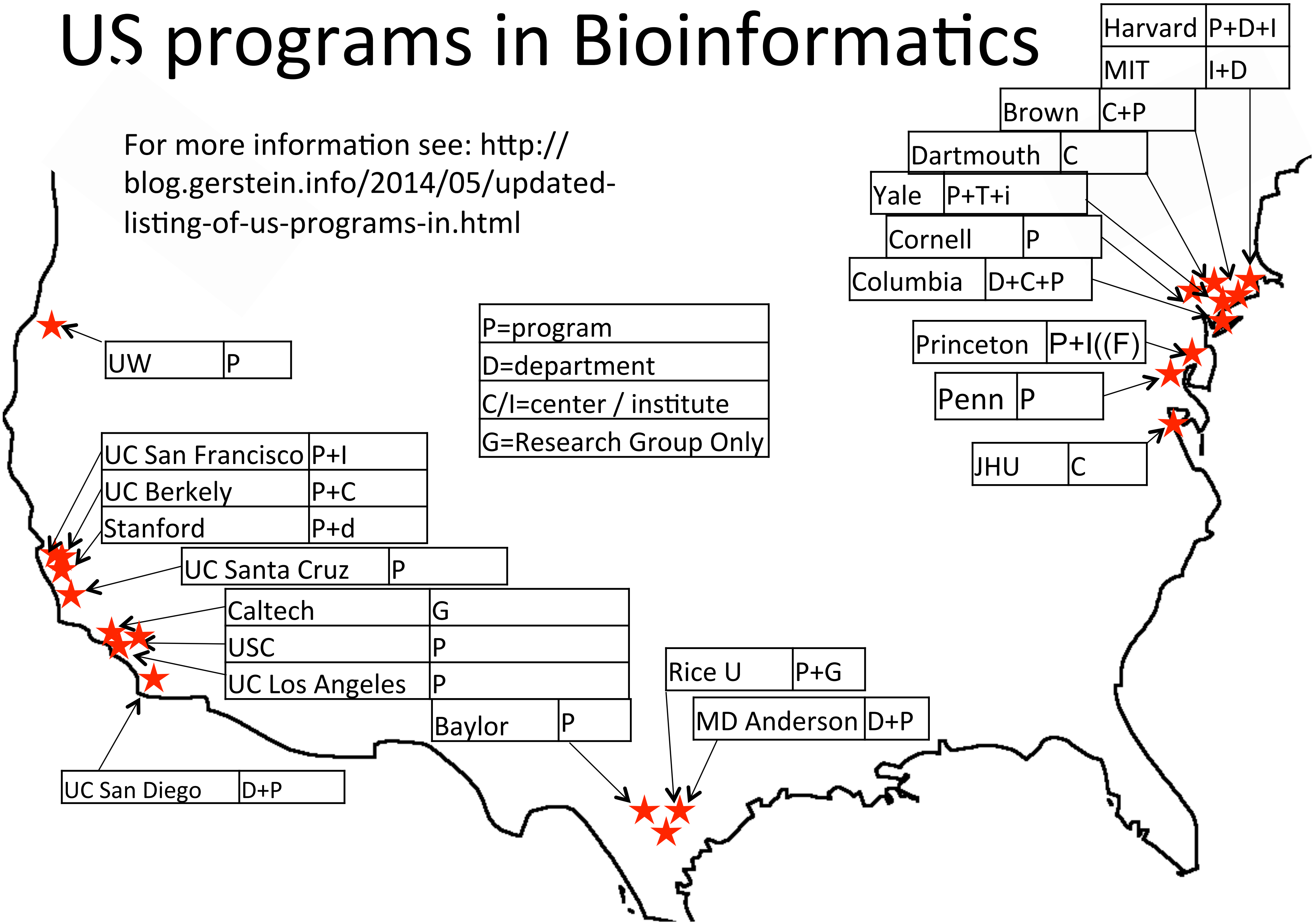
PhD students=>

1996 – 2001	Bank of America
1997 – 2002	Goldman Sachs
1998 – 2003	Psychogenics
1999 – 2004	Pearl Cohen Zedek Latzer
2002 – 2007	Illumina
2002 – 2007	Bristol-Myers Squibb
2004 – 2010	JP Morgan
2005 – 2011	MF Global
2005 – 2010	23andme
2006 – 2006	Merrill Lynch
2001 – 2007	Latham & Watkins
2007 – 2012	LEK Consulting
2009 – 2014	Illumina

Majority of industry positions in **generalized data-science** rather than traditional **bioinfo. in biotech/pharma**

US programs in Bioinformatics

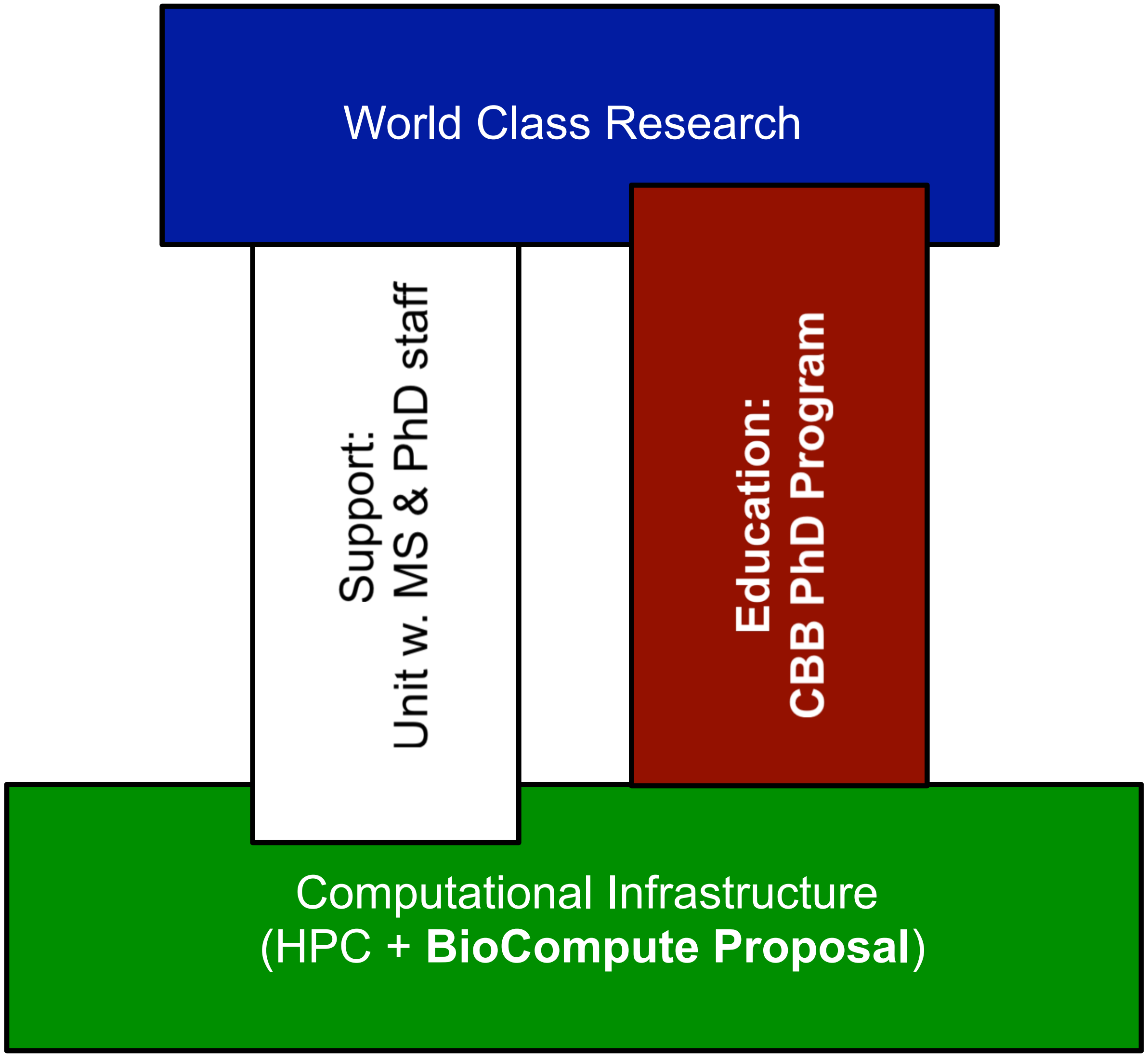
For more information see: http://blog.gerstein.info/2014/05/updated-listing-of-us-programs-in.html



P=program
D=department
C/I=center / institute
G=Research Group Only

Harvard	P+D+I
MIT	I+D
Brown	C+P
Dartmouth	C
Yale	P+T+i
Cornell	P
Columbia	D+C+P
Princeton	P+I((F)
Penn	P
JHU	C
UW	P
UC San Francisco	P+I
UC Berkely	P+C
Stanford	P+d
UC Santa Cruz	P
Caltech	G
USC	P
UC Los Angeles	P
UC San Diego	D+P
Baylor	P
Rice U	P+G
MD Anderson	D+P

Computational Biology at Yale



World Class Research

Support:
Unit w. MS & PhD staff

Education:
CBB PhD Program

Computational Infrastructure
(HPC + **BioCompute Proposal**)

Yale Life Sciences HPC

- Current workhorses
  - BulldogN [W Campus Seq. Ctr.]: 2Pb, 2.6K cores
    - used by ~20 groups (at 1% level) w/ 5 big users on each (~5% level)
  - Louise [300 George]: 1Pb, 3.5K cores
    - Similar usage profile to BulldogN ("20 & 5")
  - Omega: 1.4Pb, 8.5K cores
    - Phys. Sci. cluster, small use by ~10 bio. groups
- Future
  - Grace: 1 Pb, 1.6K cores
  - Louise & BulldogN to fold into Grace,
    most compute hardware moving to WC
  - Expanding Grace storage
    & mounting it on all clusters as a shared resource

XXXXXXX – See Shadow

- XXXXXXXX – See Shadow

XXXXXXX – See Shadow

.

Technical Architecture

- XXXXXXX – See Shadow

Cancer Genomics & PDX Use Case

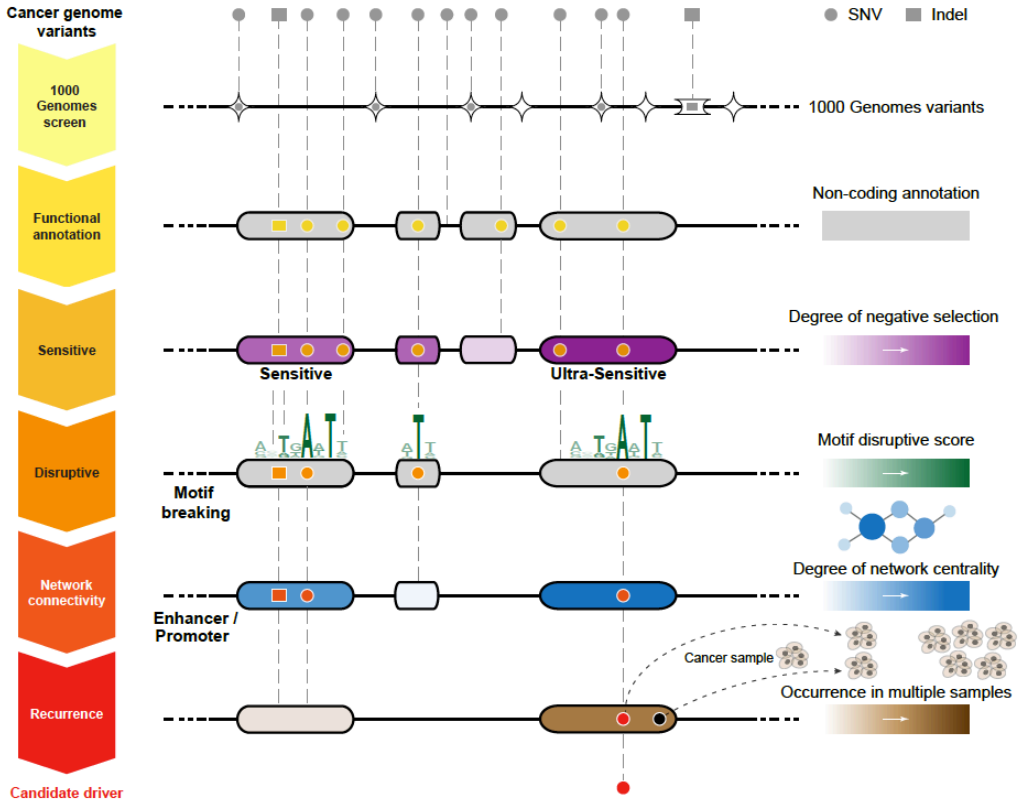
MTB

Mouse Tumor Biology Database

- Importance of topic obvious
- JAX is rapidly accruing genomics data for many PDX (Patient-derived xenograft models) samples
  - Expect the scale of data in next year to be 100-200 TB.
- Desire to analyze data, collaborate, merge data & compare with public cancer genomics information

At Yale: Researchers developing systems for analyzing cancer genomes

- Variant Calling
- Recurrence Analysis
- Mutation Prioritization
- All req. access to many sequenced genomes for context



[Khurana et al., *Science* ('13)]

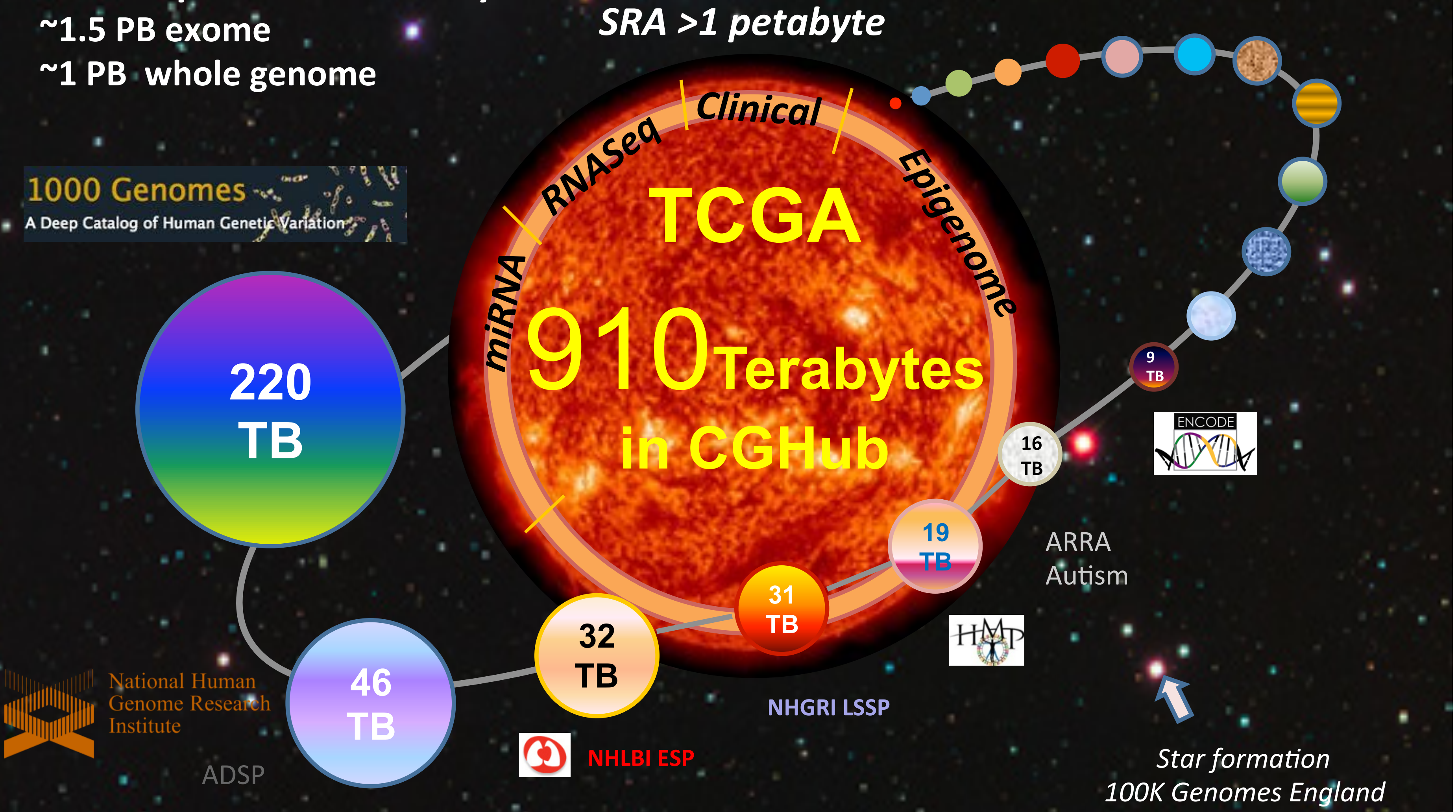
Seq Universe

[from Heidi Sofia, NHGRI]

TCGA endpoint: ~2.5 Petabytes

~1.5 PB exome

~1 PB whole genome



SRA >1 petabyte

1000 Genomes

A Deep Catalog of Human Genetic Variation

TCGA

910 Terabytes in CGHub

miRNA

RNASeq

Clinical

Epigenome

220 TB

46 TB

ADSP

National Human Genome Research Institute

32 TB

NHLBI ESP

31 TB

NHGRI LSSP

19 TB

HMP

ARRA Autism

16 TB

ENCODE

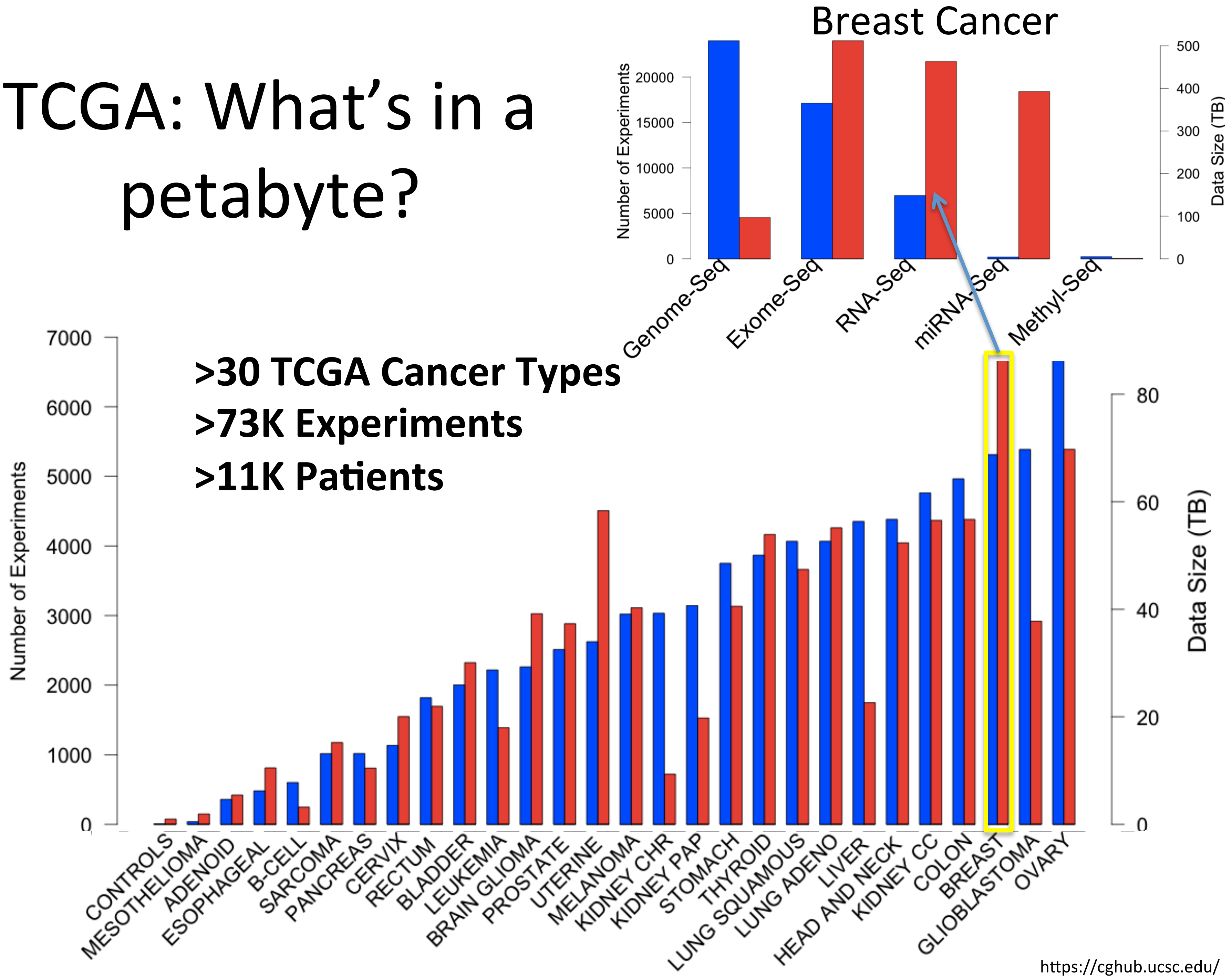
9 TB

Star formation
100K Genomes England

Sofia, 2-28-14

JESS3

TCGA: What's in a petabyte?



>30 TCGA Cancer Types

>73K Experiments

>11K Patients

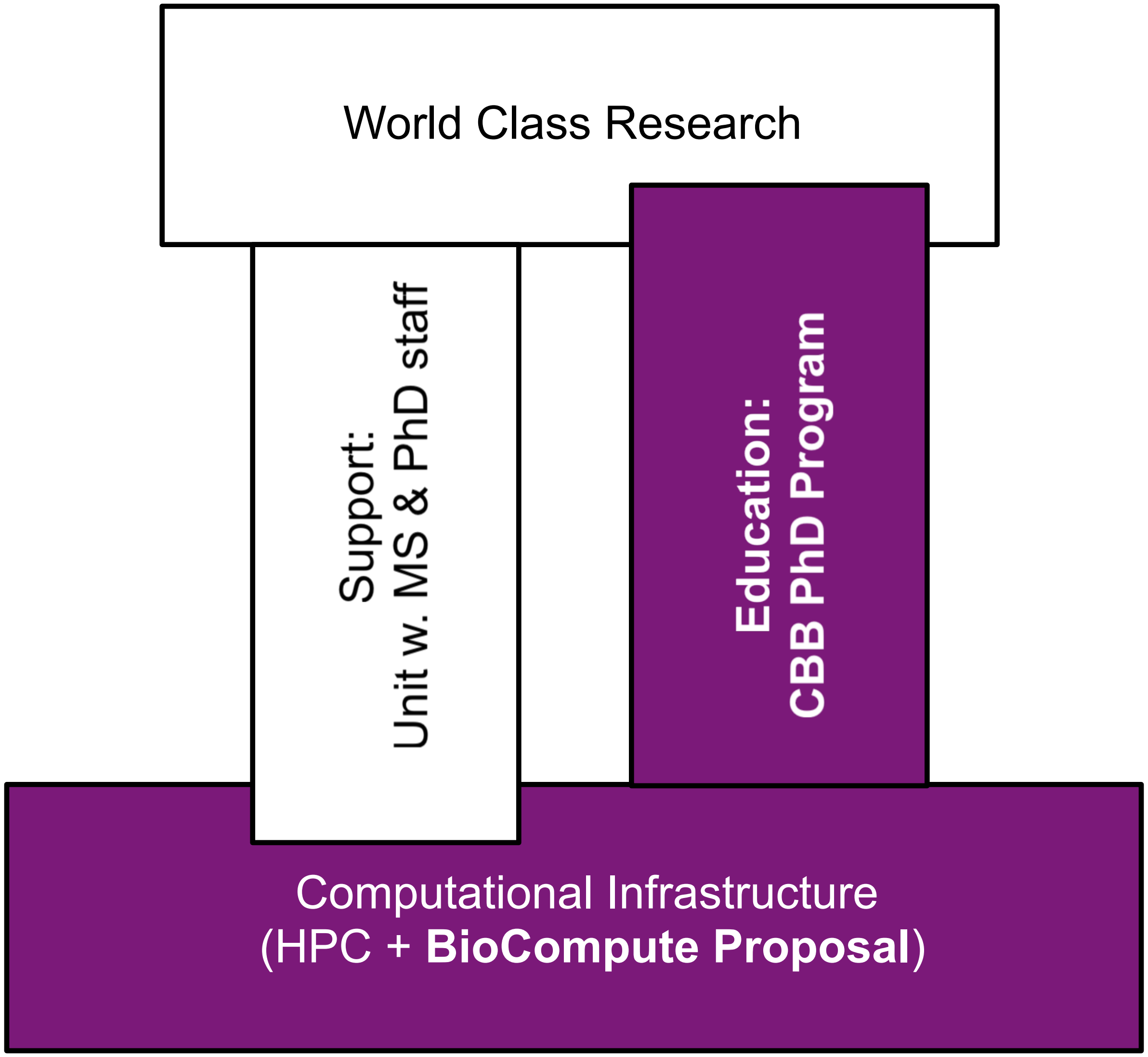
https://cghub.ucsc.edu/

Biocompute Comparables

(Extracted from public websites)

- **Princeton** (only FAS)
  - Della Cluster - 2816 cores, 2PB storage
- **Columbia** (FAS+med+seq. ctr.)
  - C2B2 - 6336 CPU cores, 73,728 GPU cores, 1.4PB storage
  - NY Genome Center - 2,000 CPU cores, 2PB storage
- **Harvard**
  - Odyssey Cluster - 60,000 cores, 79,872 CPU cores, 14PB storage
  - Massachusetts Green High Performance Computing Center
    - Incl. part of Odyssey
    - MIT, Harvard, NEU, BU, UMASS
    - $95M
- **Texas**
  - Texas Advanced Computing Center (TACC): 203K CPU cores, 319K GPU cores, 14PB storage, 200Tb of RAM!

Computational Biology at Yale



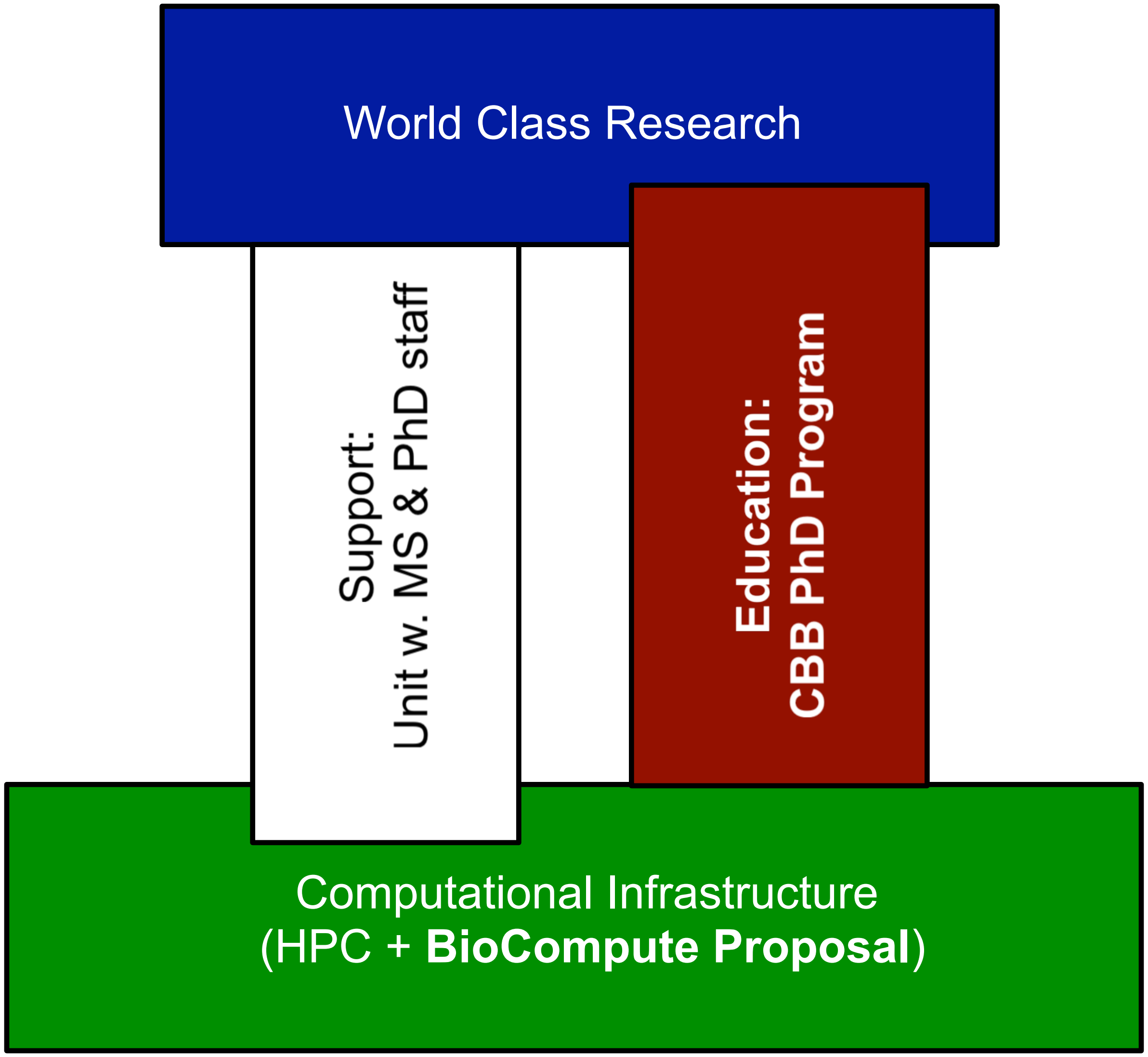
World Class Research

Support:
Unit w. MS & PhD staff

Education:
CBB PhD Program

Computational Infrastructure
(HPC + **BioCompute Proposal**)

Computational Biology at Yale



World Class Research

Support:
Unit w. MS & PhD staff

Education:
CBB PhD Program

Computational Infrastructure
(HPC + **BioCompute Proposal**)

Key points & challenges

- Current PhD program with many students & grads (>75,>35)
  - Balanced combination of Bio., Informatics & focused Bioinformatics
  - "Happy" students & diverse outcomes
  - Rise of Data Science as a driver for education
  - Students studying over whole campus
- Importance of robust computational infrastructure
  - Expertise for cloud computing
  - Necessary to tackle future problems in cancer genomics
  - More so than physical buildings!
- Challenge: Quality People!
  - Importance of getting highest quality faculty, students & computational staff
  - Often it's hard for people outside the field to judge & recruit
- Challenge: Unifying 3 locations for CBB at Yale
  - "Embedding" computational faculty, students & fellows but still giving them a coherent identify
    - Addressed by program, but what for faculty & postdocs ?
    - XXXXXXX – See Shadow

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2014.
  - Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- For SeqUniverse slide, please contact Heidi Sofia, NHGRI

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt