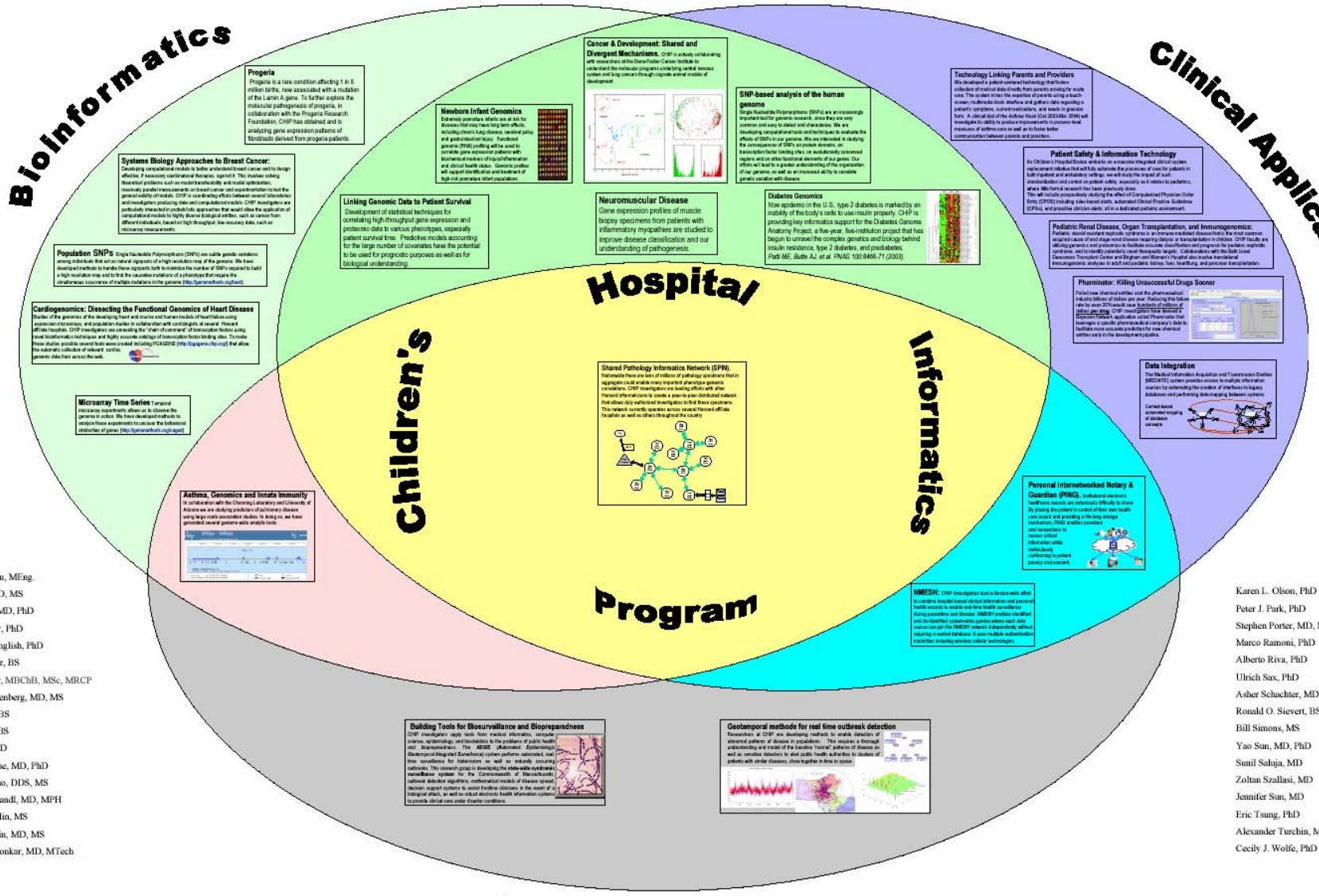
Finding the Glue to Fasten Clinical to Genomic Databases

Isaac "Zak" Kohane

Division of Health Sciences Technology of
Harvard and MIT



Shiu Chung Au, MEng.
Atul Butte, MD, MS
Zhaohui Cai, MD, PhD
Sara Dempster, PhD
Sangeeta B. English, PhD
Daniel Fleisher, BS
Hamish Fraser, MBChB, MSc, MRCP
Steven A. Greenberg, MD, MS
Rolf Hanson, BS
Albert Hong, BS
Alvin Kho, PhD
Isaac S. Kohane, MD, PhD
Winston P. Kuo, DDS, MS
Kenneth D. Mandl, MD, MPH
Maung Zaw Min, MS
Daniel J. Nigrin, MD, MS
Ashish Nimgaonkar, MD, MTech
Karen L. Olson, PhD
Peter J. Park, PhD
Stephen Porter, MD, MPH
Marco Ramoni, PhD
Alberto Riva, PhD
Ulrich Sax, PhD
Asher Schachter, MD, MMSc, MS
Ronald O. Sievert, BS, MSEE
Bill Simons, MS
Yao Sun, MD, PhD
Sunil Saluja, MD
Zoltan Szallasi, MD
Jennifer Sun, MD
Eric Tsung, PhD
Alexander Turchin, MD
Cecily J. Wolfe, PhD

Outline

- Professional context and problems
- Why this problem is at the center of genomic medicine
- How **distributed** and standardized solutions can be used to address the problem.

Clinical Informatics vs. Bioinformatics

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime
ompressed
ded to see

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Quick
Uncompre
needed t

Bioinformatics: A House Divided

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

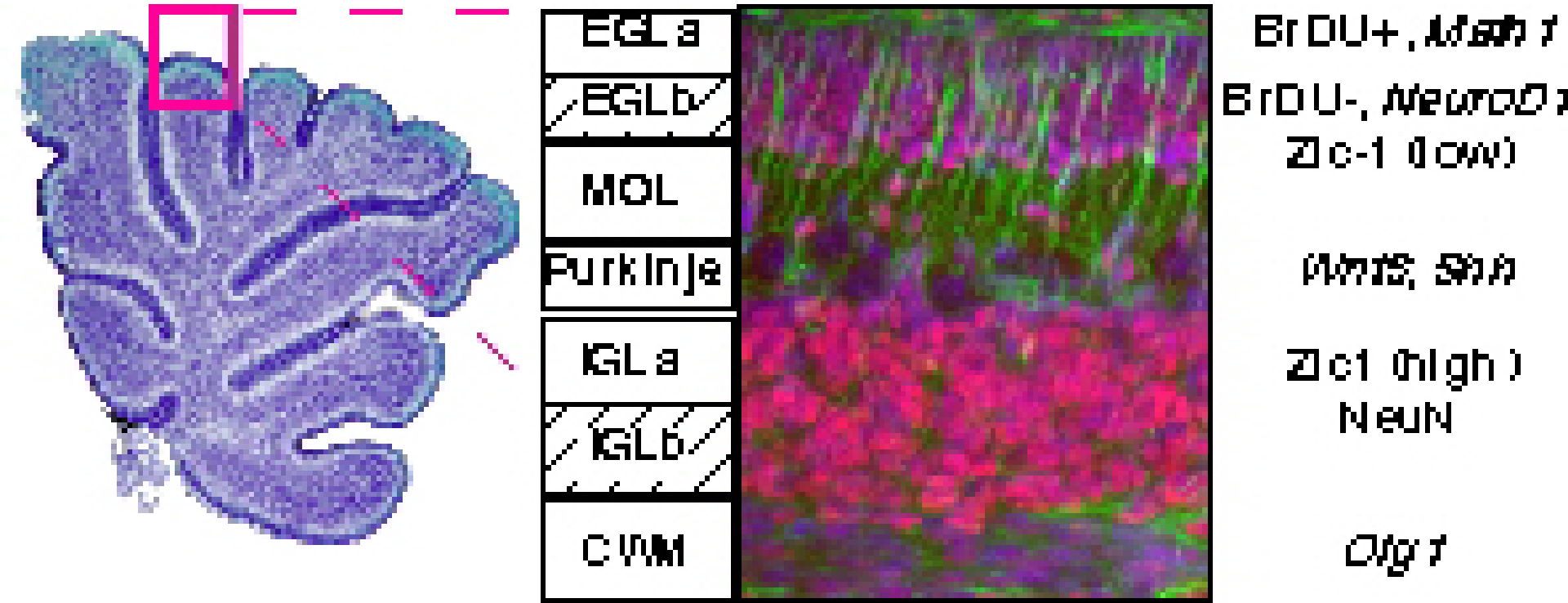
bioinformatics "is a tool, not a discipline, and tools have a way of getting absorbed into science." -L. Stein

Combining Genomics from Different Species: Sequence and Expression Based Comparative Genomics

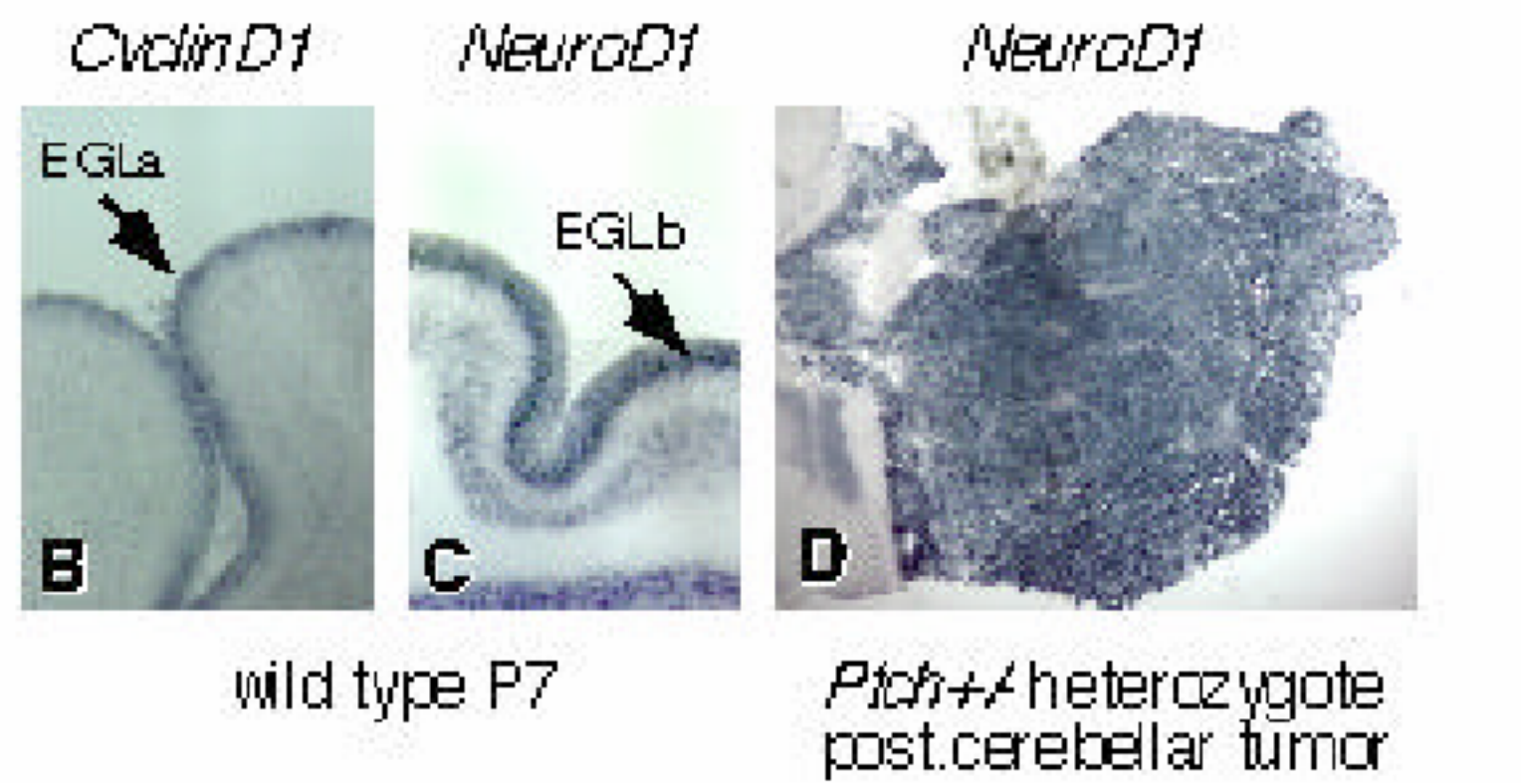
- The hypothesis that conservation of sequence across species has borne many fruit
- The harvest from the conservation of functional genomic dependencies (in expression or proteomics) has been less fruitful.
- We provide an example

Functional genomics of complex tissue

- Cerebellum has pivotal roles in the coordination of posture and locomotion
- Laminar organization of the cerebellar cortex has facilitated understanding its basic circuitry, functions and ontogeny



Sonic Hedgehog (Shh), development and tumorigenesis



Zhao et al. PNAS 2002

Integrative Approach to Comparative Genomics

- We studied time-series of day 1 to day 60 of life of the mouse
  - ✓ Each time point, cerebellum measured with a microarray
  - ✓ PNAS Zhao et al. 2002
- How can we leverage this developmental view of the mouse?
  - ✓ Does it matter for clinical medicine?
- Human medulloblastoma microarray data
  - ✓ Pomeroy et al, Nature 2002
- Principal Component Analysis to find the main sources of variance in the developmental time series

HoMoloGeNe

Nucleotide | Protein | Genome | Structure | Map Viewer

for [ ] Go Clear

Limits | Preview/Index | History | Clipboard | Details

Bu

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

HomoloGene Release Statistics

Initial numbers of genes from complete genomes, numbers of genes placed in a homology group, and the numbers of groups for each species

	HomoloGene Build 36		
Species	Number of genes		HomoloGene
	Input	Grouped	groups
H.sapiens	22,827	18,055	16,782
M.musculus	24,019	19,996	18,036
R.norvegicus	20,913	17,429	16,042
D.melanogaster	12,918	8,717	7,683
A.gambiae	12,012	8,543	7,577
C.elegans	19,109	6,502	5,260
S.pombe	4,947	3,625	3,359
S.cerevisiae	5,863	3,612	3,146
N.crassa	10,079	6,156	6,049
M.grisea	11,109	6,307	6,028
A.thaliana	26,281	8,022	4,791
P.falciparum	5,222	1,770	1,589

Last updated on: 05/25/2004

Related Resources

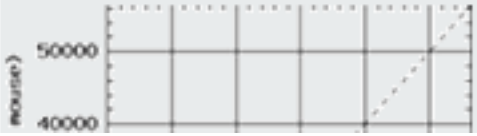
Entrez Genome

A collection of complete genome sequences that includes more than 1000 viruses and over 100 microbes

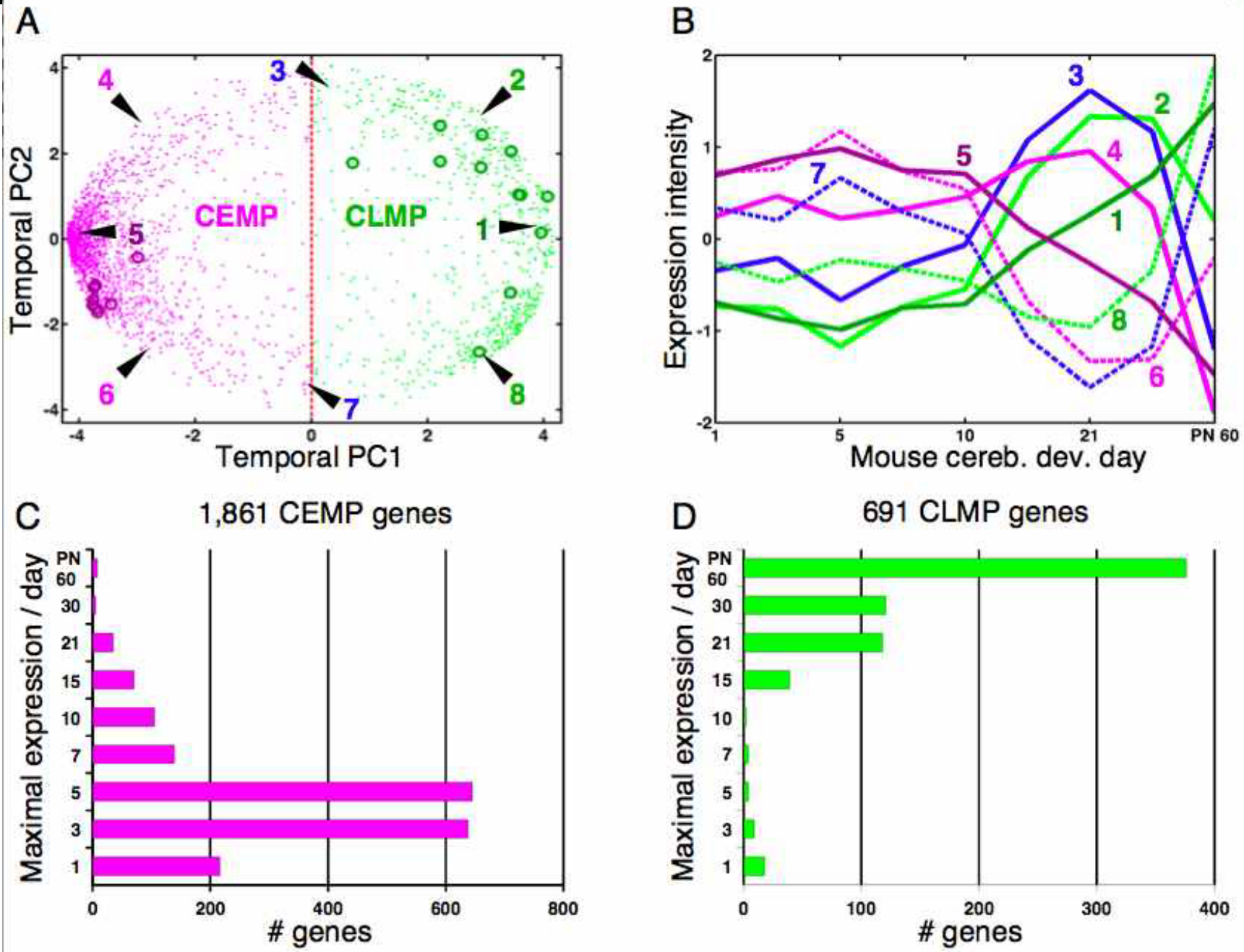
- Archaea
- Bacteria
- Eukaryota
- Viruses

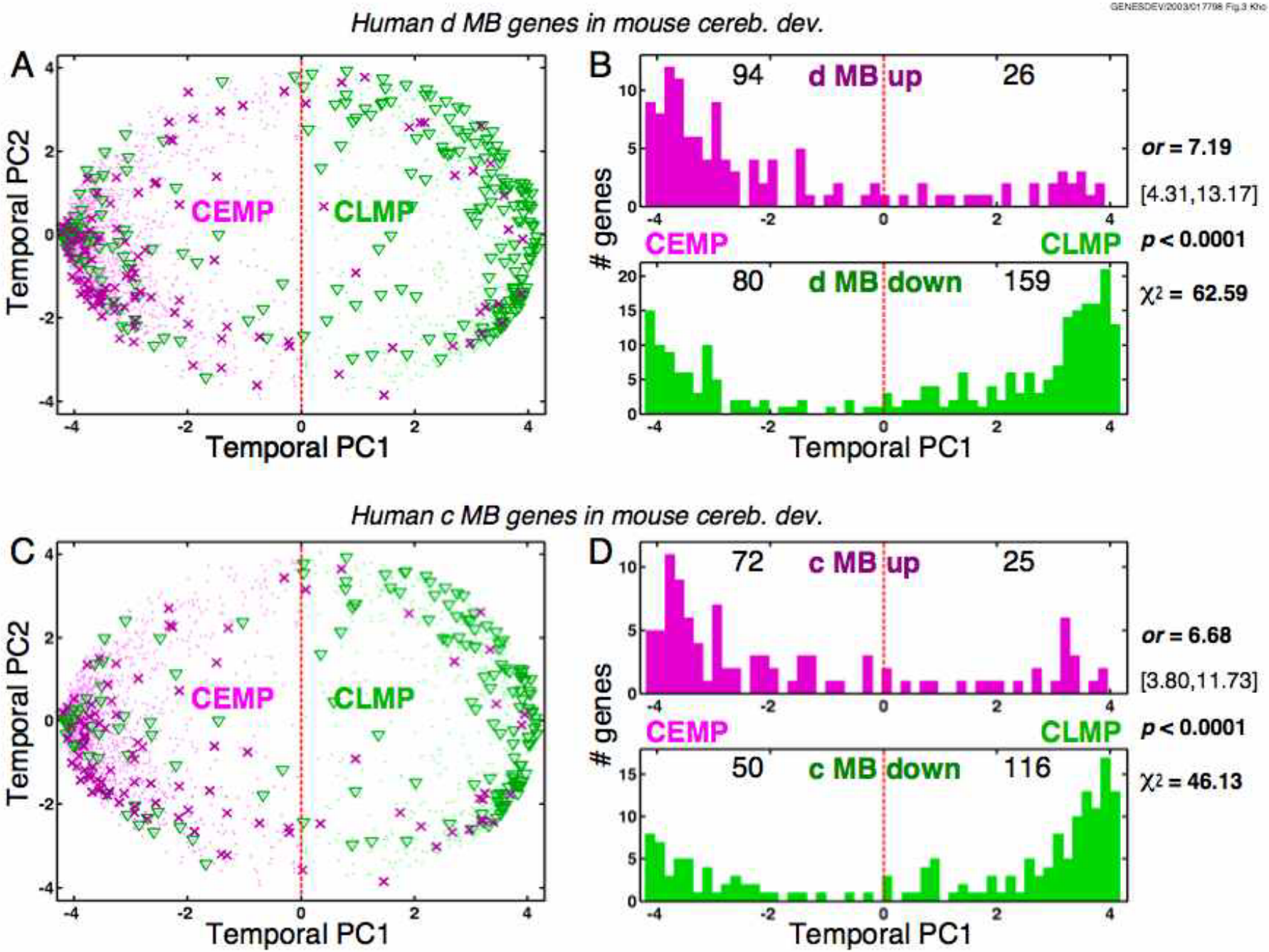
Tax Plot

Three-way view of genome similarities

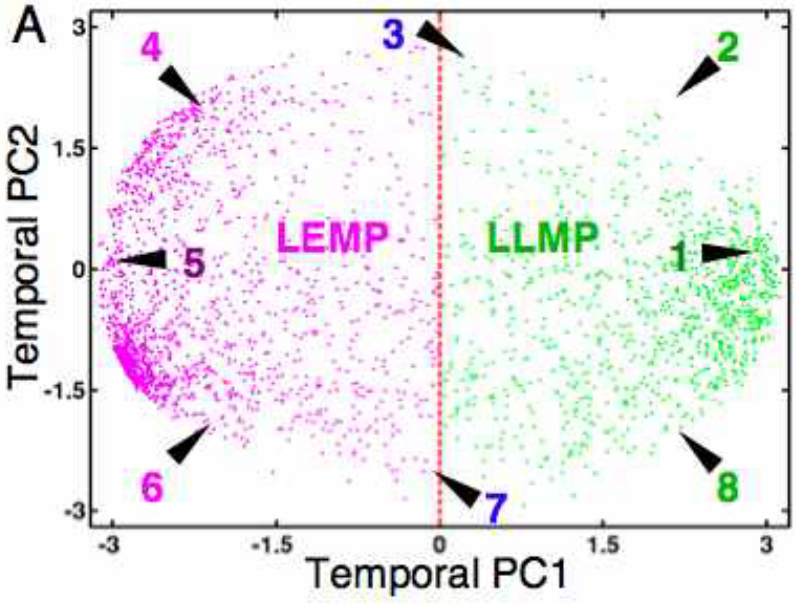


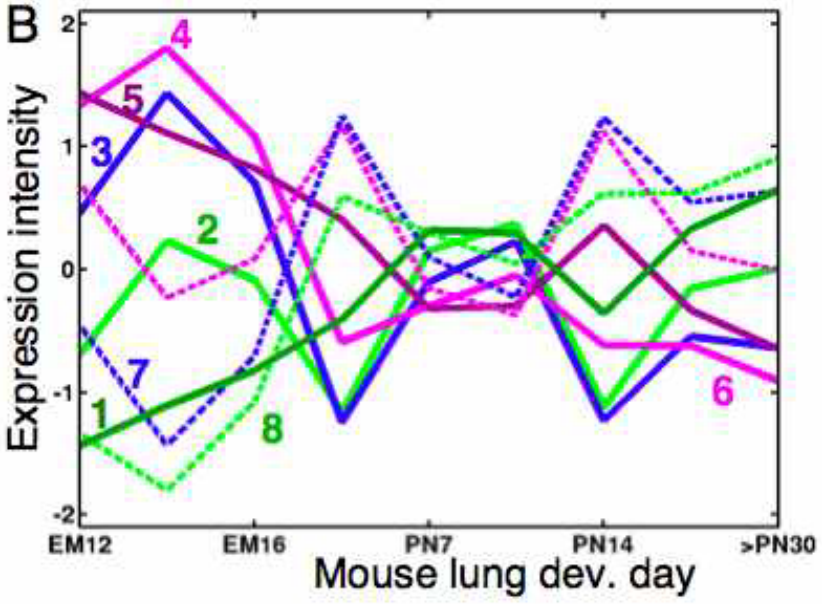
Principal Components


A
Temporal PC2
Temporal PC1
CEMP
CLMP
1
2
3
4
5
6
7
8
B
Expression intensity
Mouse cereb. dev. day
PN 60
C
1,861 CEMP genes
Maximal expression / day
genes
D
691 CLMP genes
Maximal expression / day
genes

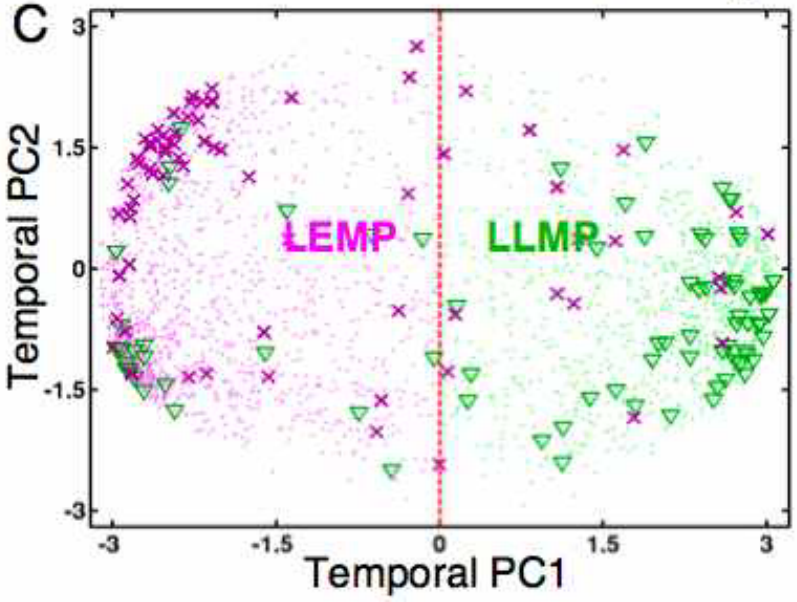

Human d MB genes in mouse cereb. dev.
A
Temporal PC2
Temporal PC1
CEMP
CLMP
B
94
d MB up
26
80
d MB down
159
genes
CEMP
CLMP
or = 7.19
[4.31,13.17]
p < 0.0001
χ² = 62.59
Human c MB genes in mouse cereb. dev.
C
Temporal PC2
Temporal PC1
CEMP
CLMP
D
72
c MB up
25
50
c MB down
116
genes
CEMP
CLMP
or = 6.68
[3.80,11.73]
p < 0.0001
χ² = 46.13

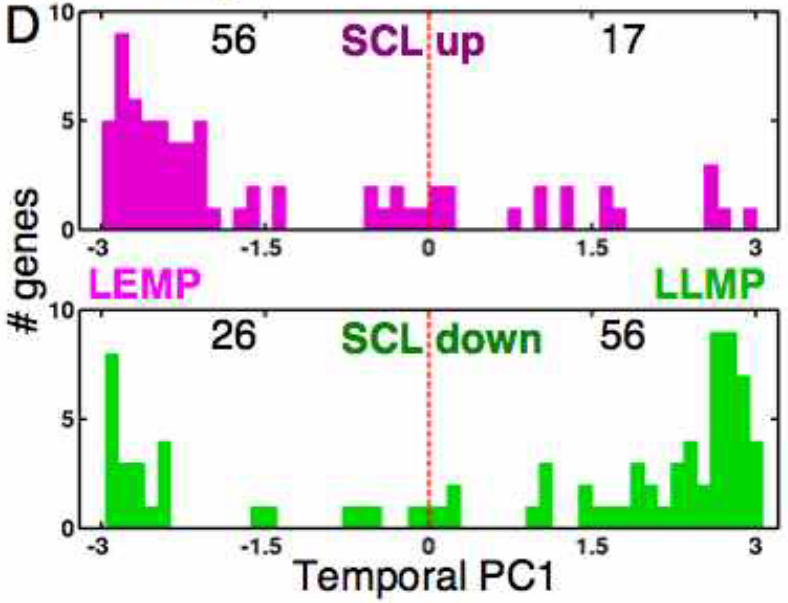
Does this pattern generalize to other tissues?





Human SCL genes in mouse lung dev.





***or* = 7.10**

[3.47,14.51]

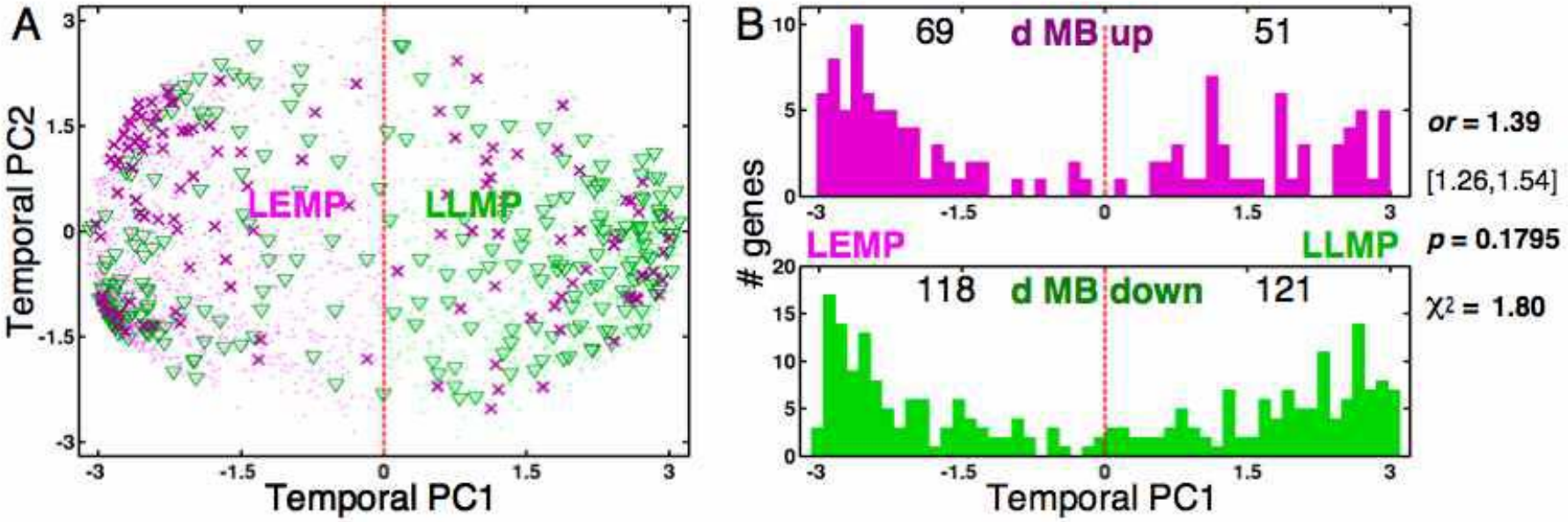
***p* < 0.0001**

χ^2 = 29.61

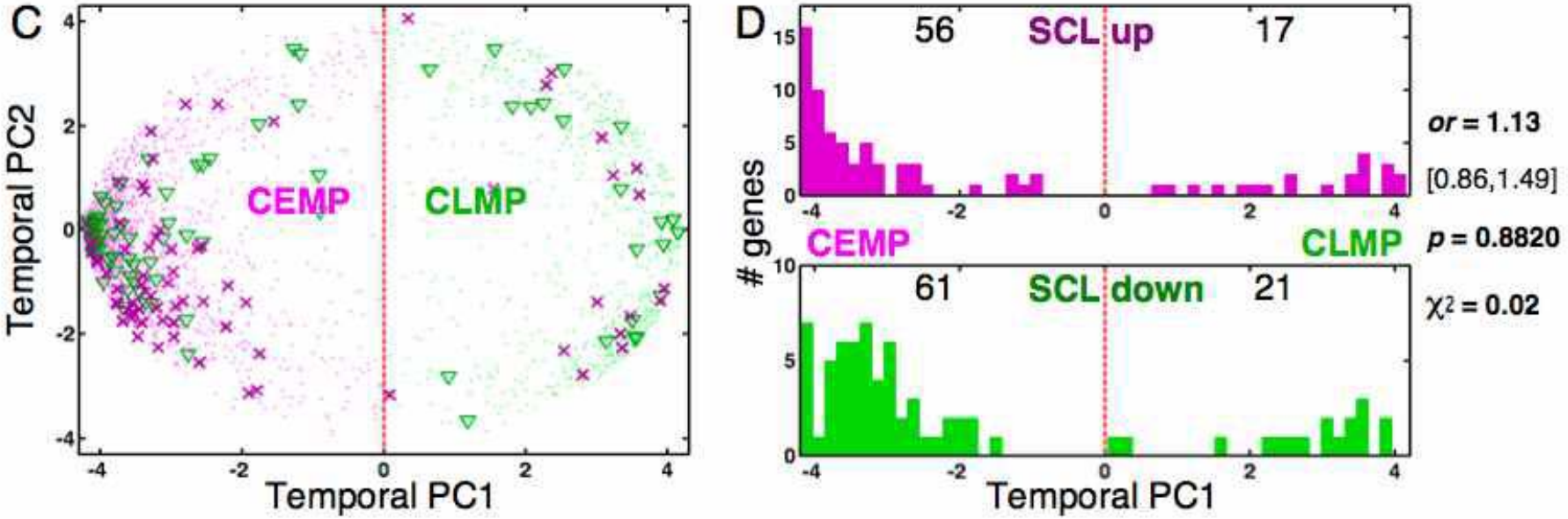
Is this phenomenon, merely reporting proliferation?

Or, does it reflect the tissue-specific developmental program?

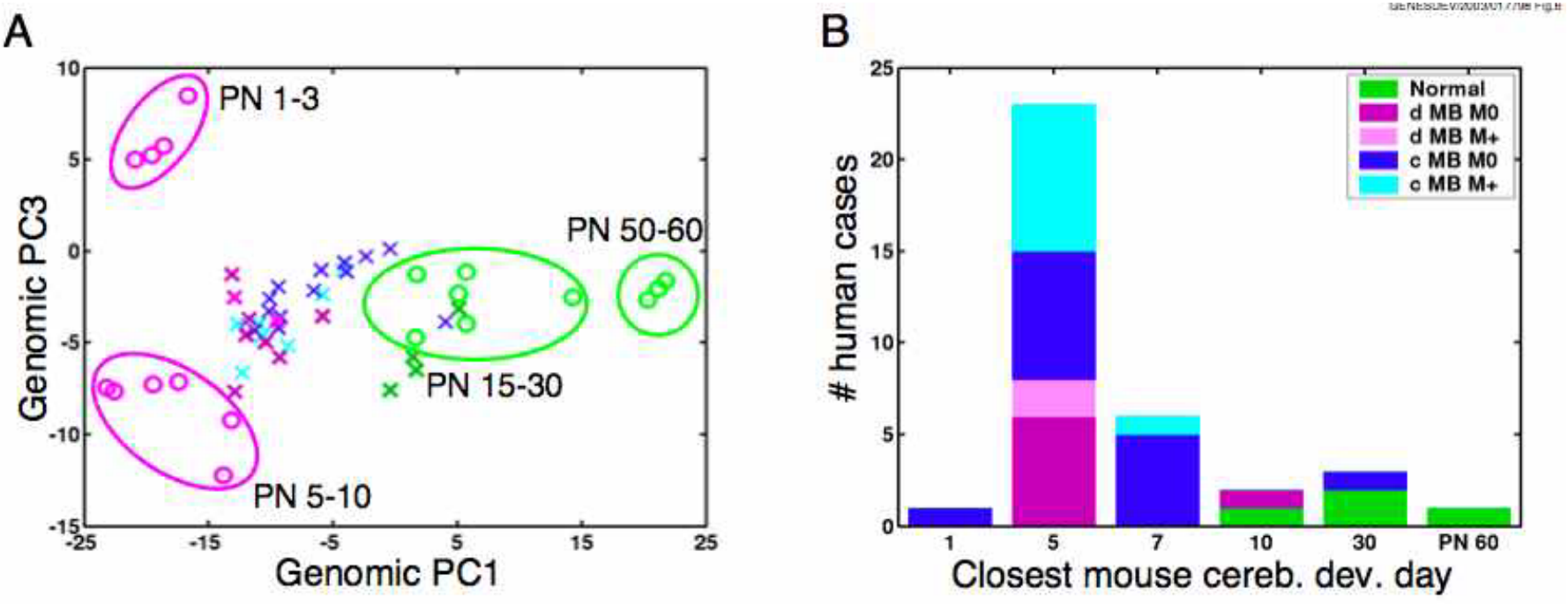
Human d MB genes in mouse lung dev.



Human SCL genes in mouse cereb. dev.



What about a macro view?



Interim Conclusion (I)

- Revisiting an old idea with the quantitation and precision of the genomic era.
- Lobstein (1829) and Cohnheim (1887) were amongst the first to theorize similarities between human embryogenesis and the biology of cancer cells .
- The brain tumor classification system of Bailey and Cushing (1926),
  - ✓ from which modern taxonomies derive,
  - ✓ emphasizes the histologic resemblance to cells of the developing central nervous system

Kho et al. *Genes and Development*, March 2004

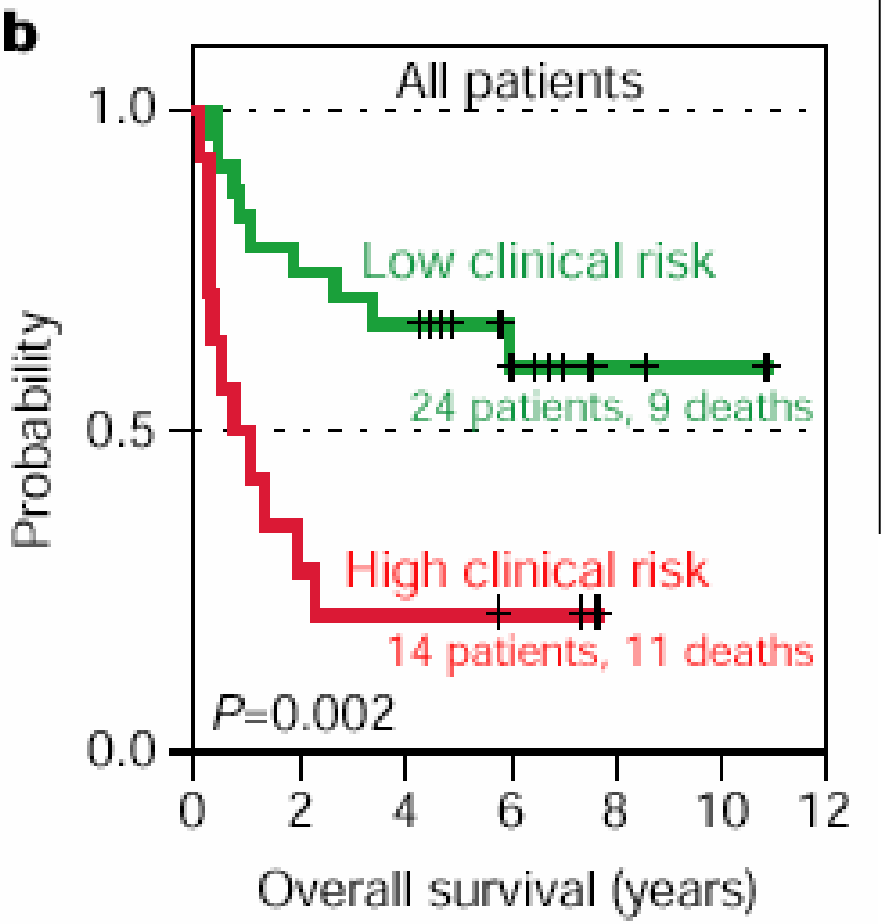
Conclusion (II)

- Projecting human solid tumors against a background of mouse models provides insight
  - ✓ Into diagnostic staging
  - ✓ Into biological process that characterize the tumors
  - ✓ May have biologically ground prognostic value
- We demonstrate that this may be generalizable to many other systems of human disease
- Demonstrate that comparative **functional** genomic data sets DEPENDS crucially: on a **COMMON VOCABULARY** and cross-species mapping.

Integrative Biology

- With the multiplicity of data sets, increasing efforts at integration to learn more about the underlying biology.
- Integration however present fundamental methodological problems.

Example: Alizadeh 2000



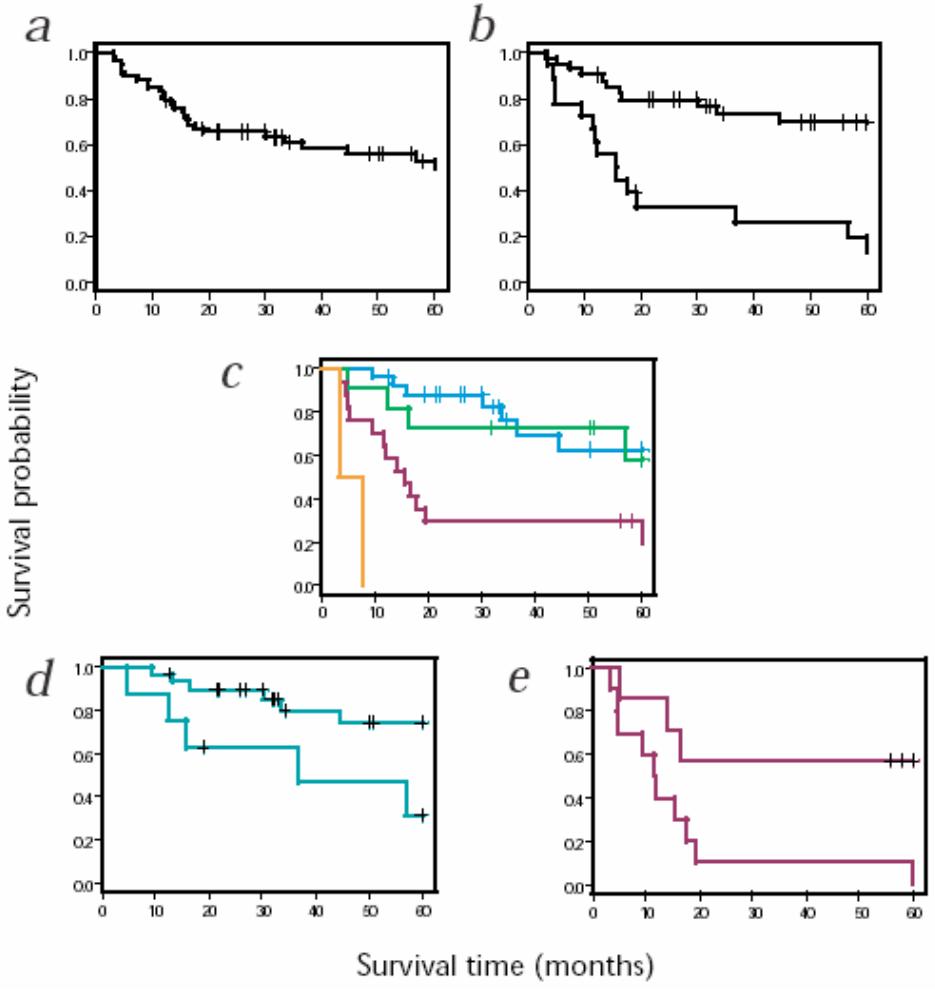
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[5], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], Hajeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], David B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], James O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], Patrick O. Brown[1,18] & Louis M. Staudt[5]

Departments of [1]Biochemistry, [3]Genetics, [14]Pathology, [6]Medicine, [10]Pediatrics and [11]Health Research & Policy and Statistics, and [18]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA
[5]Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[7]Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA
[8]CBER, FDA, Bethesda, Maryland 20892, USA
[9]Research Genetics, Huntsville, Alabama 35801, USA
Departments of [12]Pathology and Microbiology, and [13]Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA
[15]Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[16]Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA
[17]Walter Reed Army Medical Center, Washington, DC 20307, USA
[2]These authors contributed equally to this work

Alizadeh et al.

Example: Shipp 2002



Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning

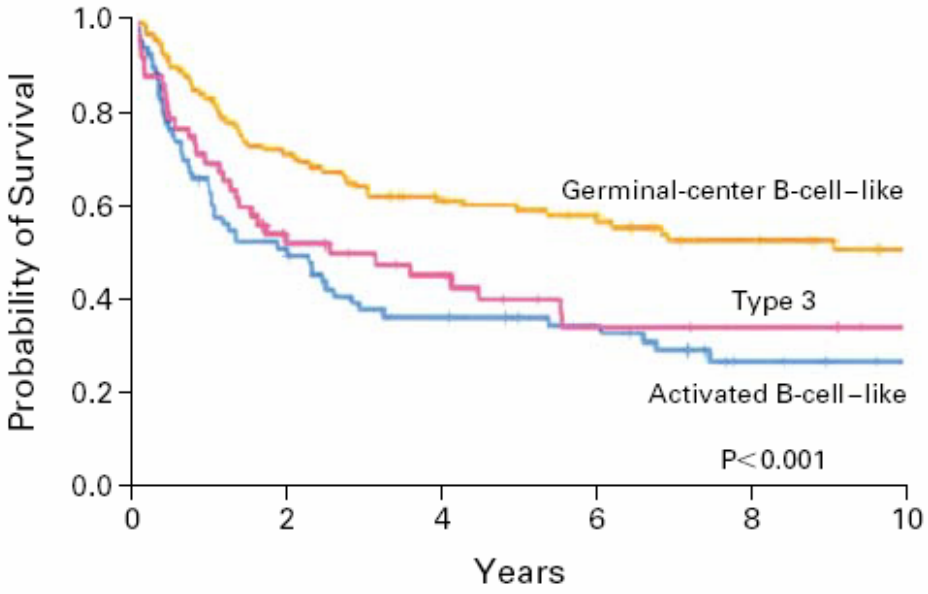
Margaret A. Shipp[1], Ken N. Ross[2], Pablo Tamayo[3], Andrew P. Weng[3], Jeffery L. Kutok[3], Ricardo C.T. Aguiar[1], Michelle Gaasenbeek[2], Michael Angelo[2], Michael Reich[2], Geraldine S. Pinkus[3], Tane S. Ray[6], Margaret A. Koval[1], Kim W. Last[4], Andrew Norton[5], T. Andrew Lister[4], Jill Mesirov[2], Donna S. Neuberg[1], Eric S. Lander[2,7], Jon C. Aster[3] & Todd R. Golub[1,2]

[1]Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA
[2]Whitehead Institute for Biomedical Research/Massachusetts Institute of Technology Center for Genome Research, Cambridge, Massachusetts, USA
[3]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[4]ICRF Medical Oncology Unit and [5]Pathology Unit, St. Bartholomew's Hospital, London, UK
[6]Department of Computer Science, Maths and Physics, University of West Indies, Bridgetown, Barbados
[7]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
K.N.R. and P.T. contributed equally to this study.
Correspondence should be addressed to M.A.S.: email: margaret_shipp@dfci.harvard.edu, or T.R.G.: email: golub@genome.wi.mit.edu

Fig. 4 Overall survival predictions for DLBCL study patients. ***a***, 5-year OS for the entire study group. 33 of 58 DLBCL study patients remained alive at a median of a 58-month follow-up. The predicted 5-year OS for the group as a whole was 54%. ***b***, 5-year OS for favorable and unfavorable risk groups defined by the 13-gene model (70% versus 12%, $P = 0.00004$). Top line, cured; bottom, fatal/refractory. ***c***, 5-year OS for patients in L-risk (green line), LI-risk (blue line), HI-risk (red line) and H-risk (orange line) categories as defined by the IPI: L, 26 pts; LI, 11 pts; HI, 17 pts; H, 2 pts. ***d***, 5-year OS for combined L/LI-risk patients with favorable or unfavorable disease as defined by the molecular model (75% versus 32%, nominal $P = 0.02$) Top line, cured; bottom, fatal/refractory. ***e***, 5-year OS for HI-risk patients with favorable or unfavorable disease as defined by the molecular model (57% versus 0%; nominal $P = 0.02$). Top line, cured; bottom, fatal/refractory.

Example: Rosenwald

C



No. at Risk						
Germinal-center B-cell–like	115	81	60	46	32	19
Type 3	52	24	18	10	8	5
Activated B-cell–like	73	35	23			

The New England Journal of Medicine

VOLUME 346 JUNE 20, 2002 NUMBER 25

THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA

ANDREAS ROSENWALD, M.D., GEORGE WRIGHT, PH.D., WING C. CHAN, M.D., JOSEPH M. CONNORS, M.D., ELIAS CAMPO, M.D., RICHARD I. FISHER, M.D., RANDY D. GASCOYNE, M.D., H. KONRAD MULLER-HERMELINK, M.D., ERLEND B. SMELAND, M.D., PH.D., AND LOUIS M. STAUDT, M.D., PH.D., FOR THE LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT

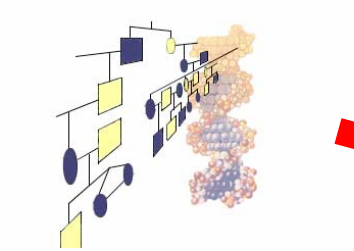
Interim Conclusion

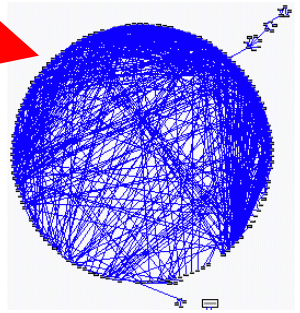
- Less than 100 patients is small in any clinical study
- With thousands of genomic variables, 100's is even smaller with respect to the dangers of:
  - ✓ Over-fitting
  - ✓ Multiple-hypothesis testing

- Many methodological ills can be forgiven by large numbers of cases

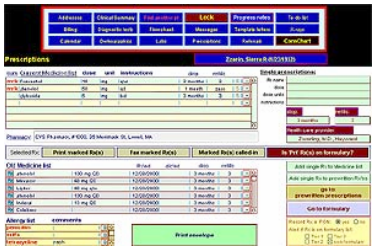
So, how do we get enough samples?

Family Hx





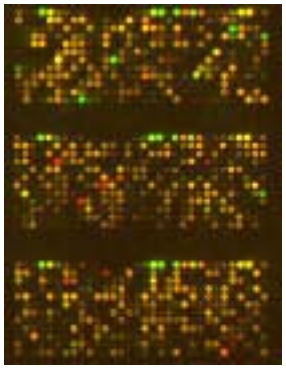
Consent



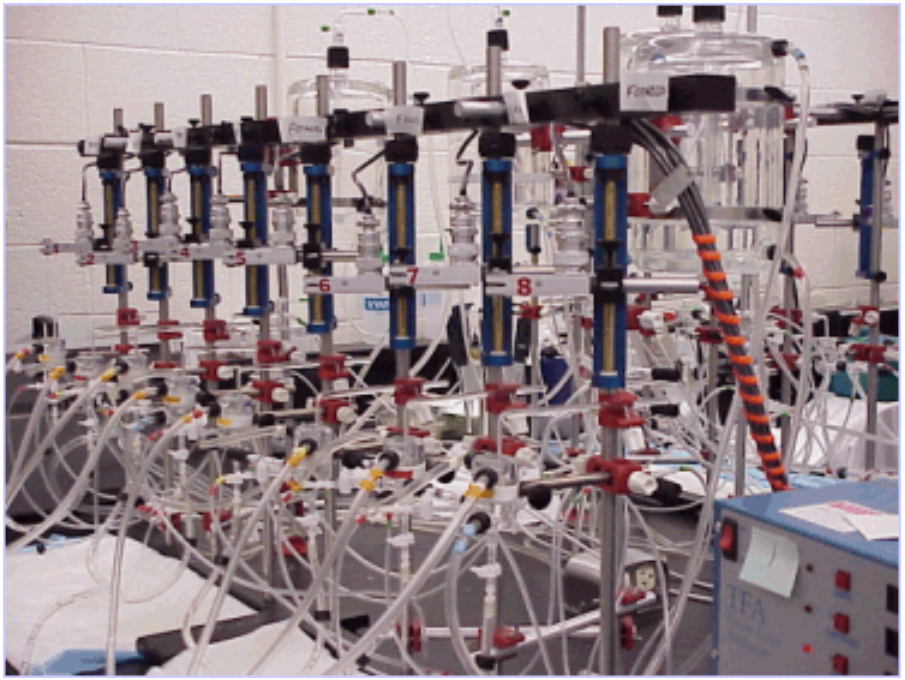
Annotation

Banking

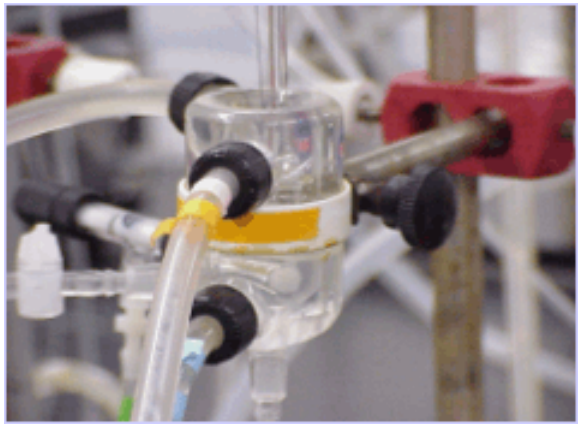
Genomics

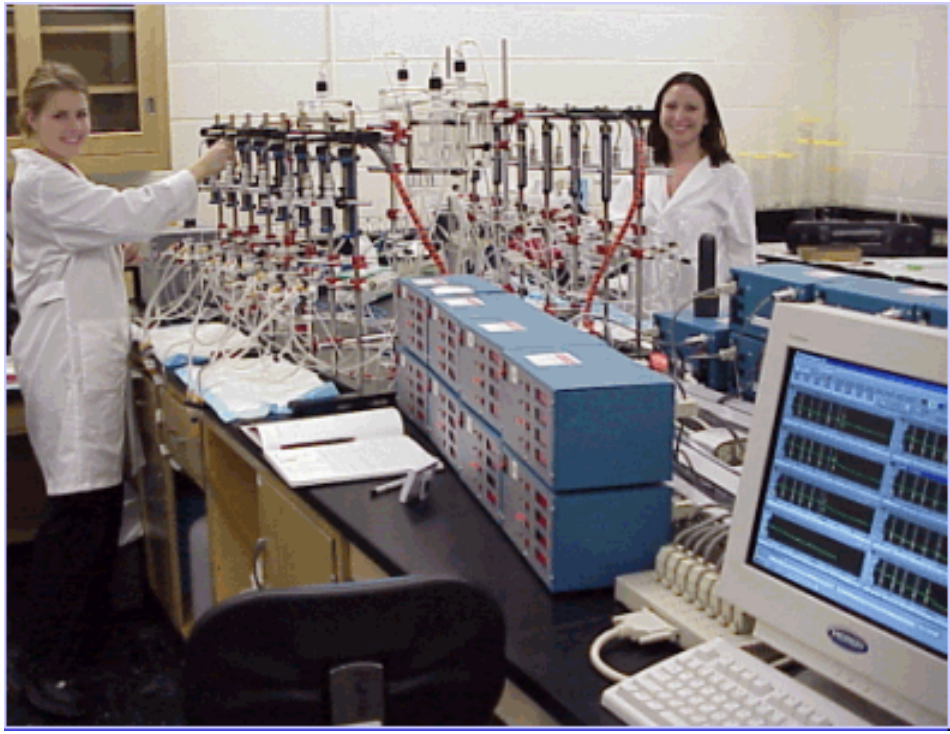


Isaac S. Kohane BECON/BISTI-2004



High-throughput phenotyping at MCW





Not Registered? ENTER HERE

REQUEST ACCOUNT

username: demo2

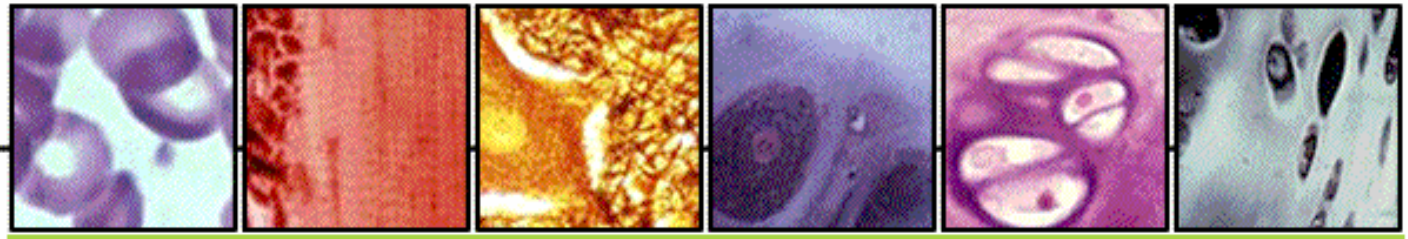
password:

REGISTERED USER LOGIN

National Institutes of Health
National Cancer Institute

Brigham & Women's Hospital
Beth Israel Deaconess Medical Center
Cedars-Sinai Medical Center
Children's Hospital
Dana-Farber Cancer Institute
Massachusetts General Hospital
Olive View Medical Center
UCLA Medical Center
VA Greater LA Healthcare System
University of Pittsburgh Medical Center

SPiN

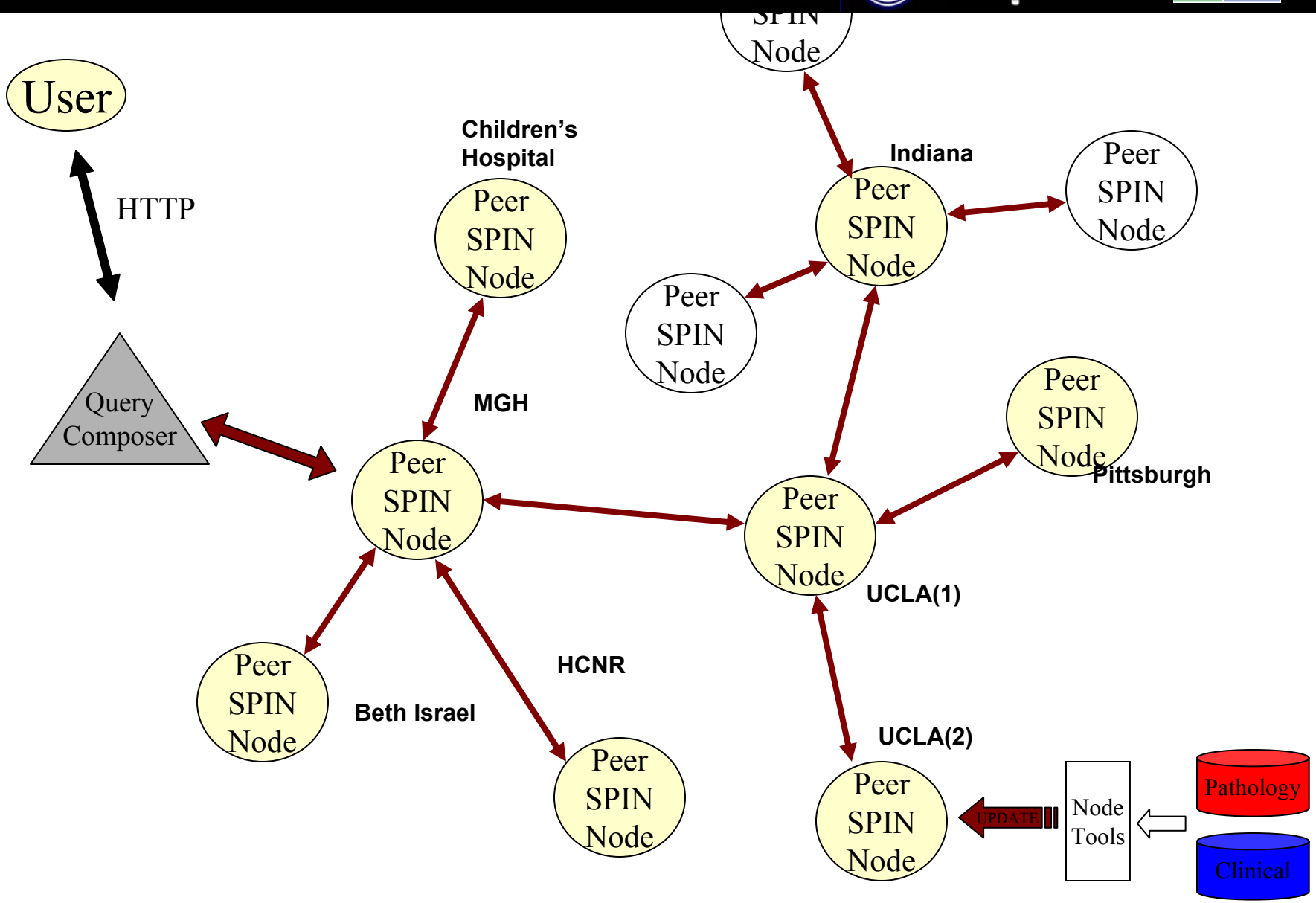


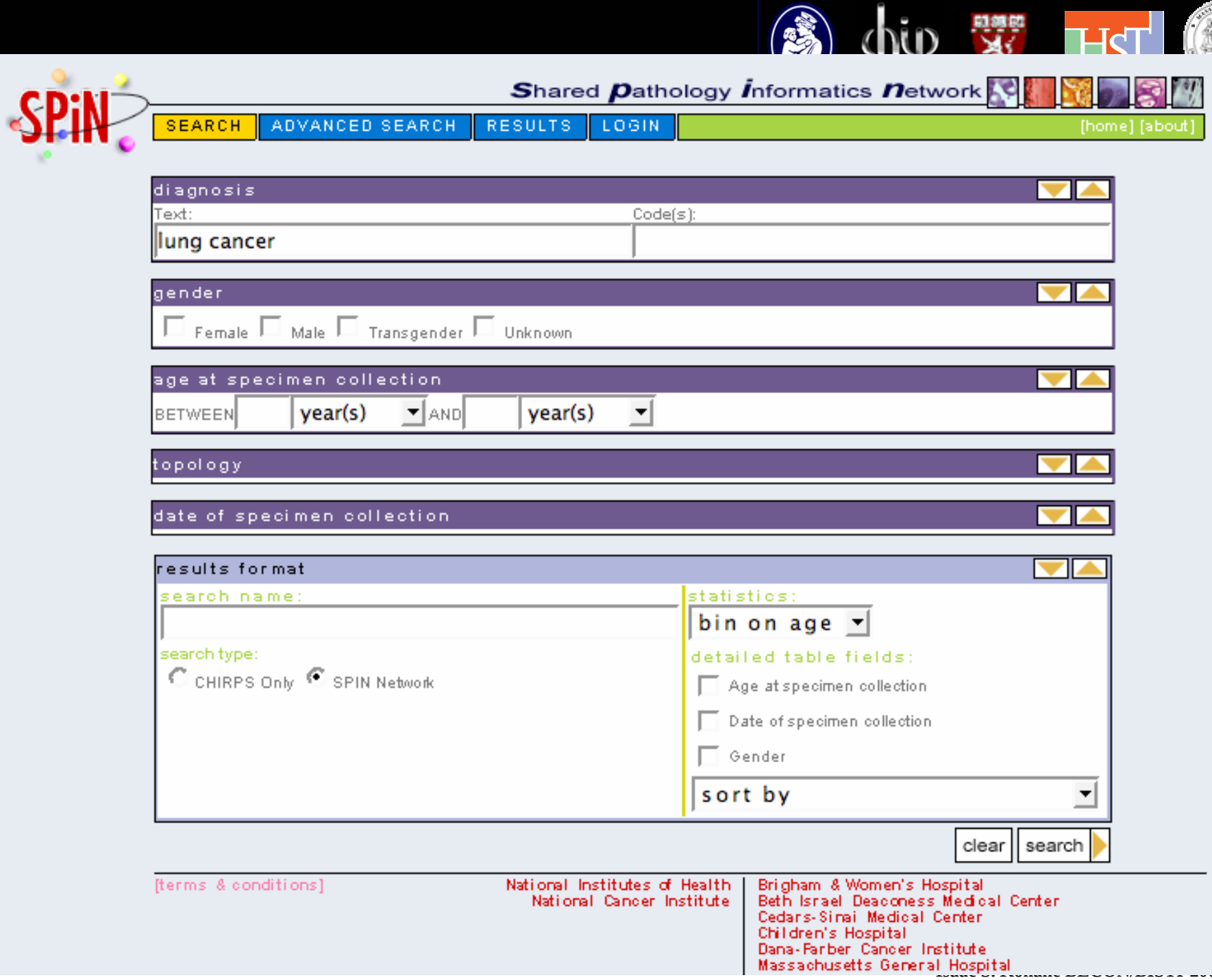
Shared **p**athology **i**nformatics **n**etwork

The objective of this initiative is to use state-of-the-art informatics techniques to establish an Internet-based virtual database that will allow investigators to locate appropriate human tissue specimens for their research.

The SPIN software will allow approved researchers access to data that describe archived tissue specimens across multiple institutions while still allowing those institutions to maintain local control of the data. The need for this capability has been fueled by the growing use of tissues, diagnostic specimens, and their related clinical data in modern biomedical research.

[terms & conditions]


User
HTTP
Query Composer
Children's Hospital
Peer SPIN Node
MGH
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Beth Israel
HCNR
Peer SPIN Node
Peer SPIN Node
Indiana
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Pittsburgh
Peer SPIN Node
UCLA(1)
Peer SPIN Node
UCLA(2)
UPDATE
Node Tools
Pathology
Clinical


Shared Pathology Informatics Network
SEARCH
ADVANCED SEARCH
RESULTS
LOGIN
[home] [about]
diagnosis
Text:
lung cancer
Code(s):
gender
Female
Male
Transgender
Unknown
age at specimen collection
BETWEEN
year(s)
AND
year(s)
topology
date of specimen collection
results format
search name:
search type:
CHIRPS Only
SPIN Network
statistics:
bin on age
detailed table fields:
Age at specimen collection
Date of specimen collection
Gender
sort by
clear
search
[terms & conditions]
National Institutes of Health
National Cancer Institute
Brigham & Women's Hospital
Beth Israel Deaconess Medical Center
Cedars-Sinai Medical Center
Children's Hospital
Dana-Farber Cancer Institute
Massachusetts General Hospital

Shared Pathology Informatics Network

SEARCH | ADVANCED SEARCH | RESULTS | LOGIN | [home] [about]

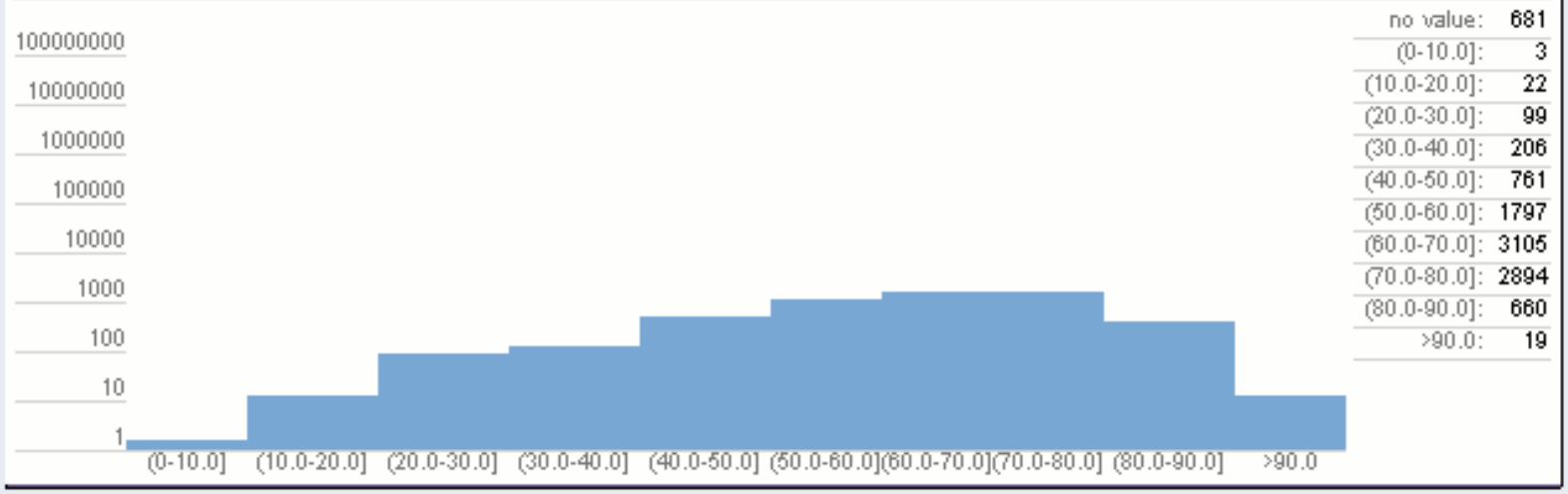
Select ALL

delete selected results

RESULTS | SUBQUERY | NODES Fetching.

name: records: 10247 nodes: 7 sent: 2004-05-27 23:49:36

age at sample collection



no value:	681
(0-10.0]:	3
(10.0-20.0]:	22
(20.0-30.0]:	99
(30.0-40.0]:	206
(40.0-50.0]:	761
(50.0-60.0]:	1797
(60.0-70.0]:	3105
(70.0-80.0]:	2894
(80.0-90.0]:	660
>90.0:	19

RESULTS | SUBQUERY | NODES

name: records: 0 nodes: 6 sent: 2004-05-27 23:48:55

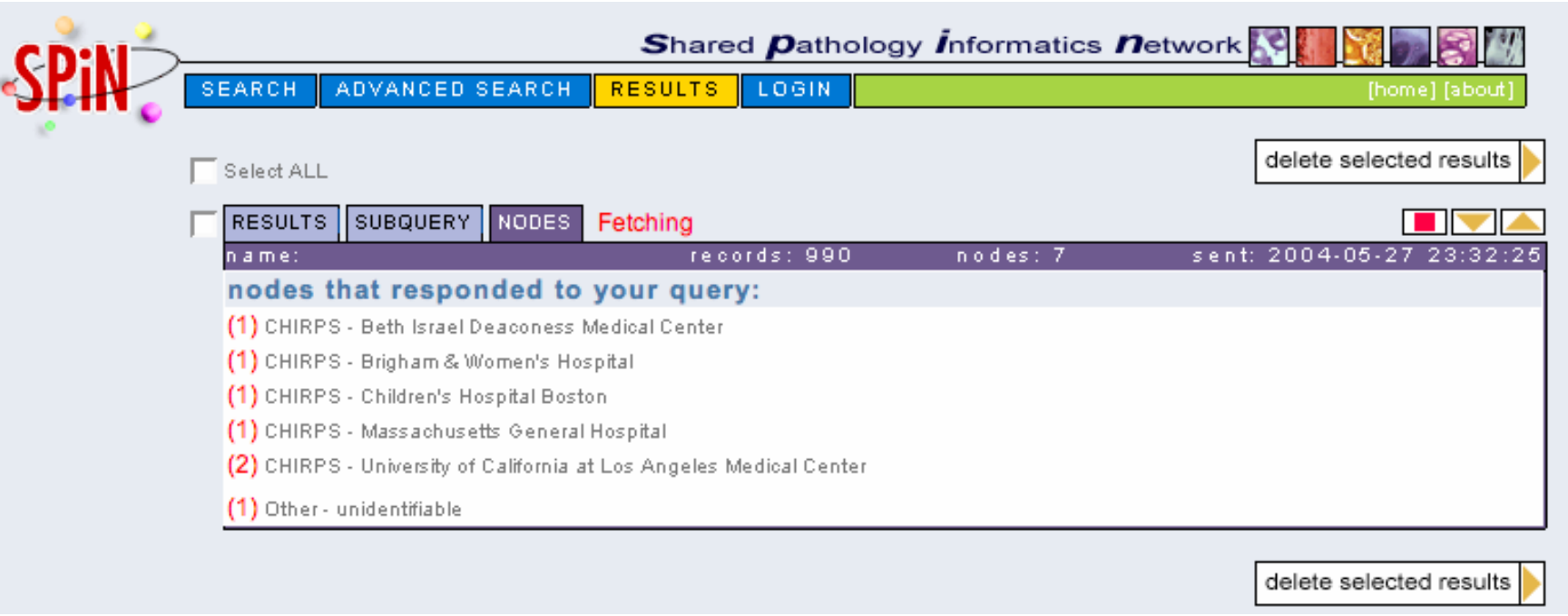
RESULTS | SUBQUERY | NODES

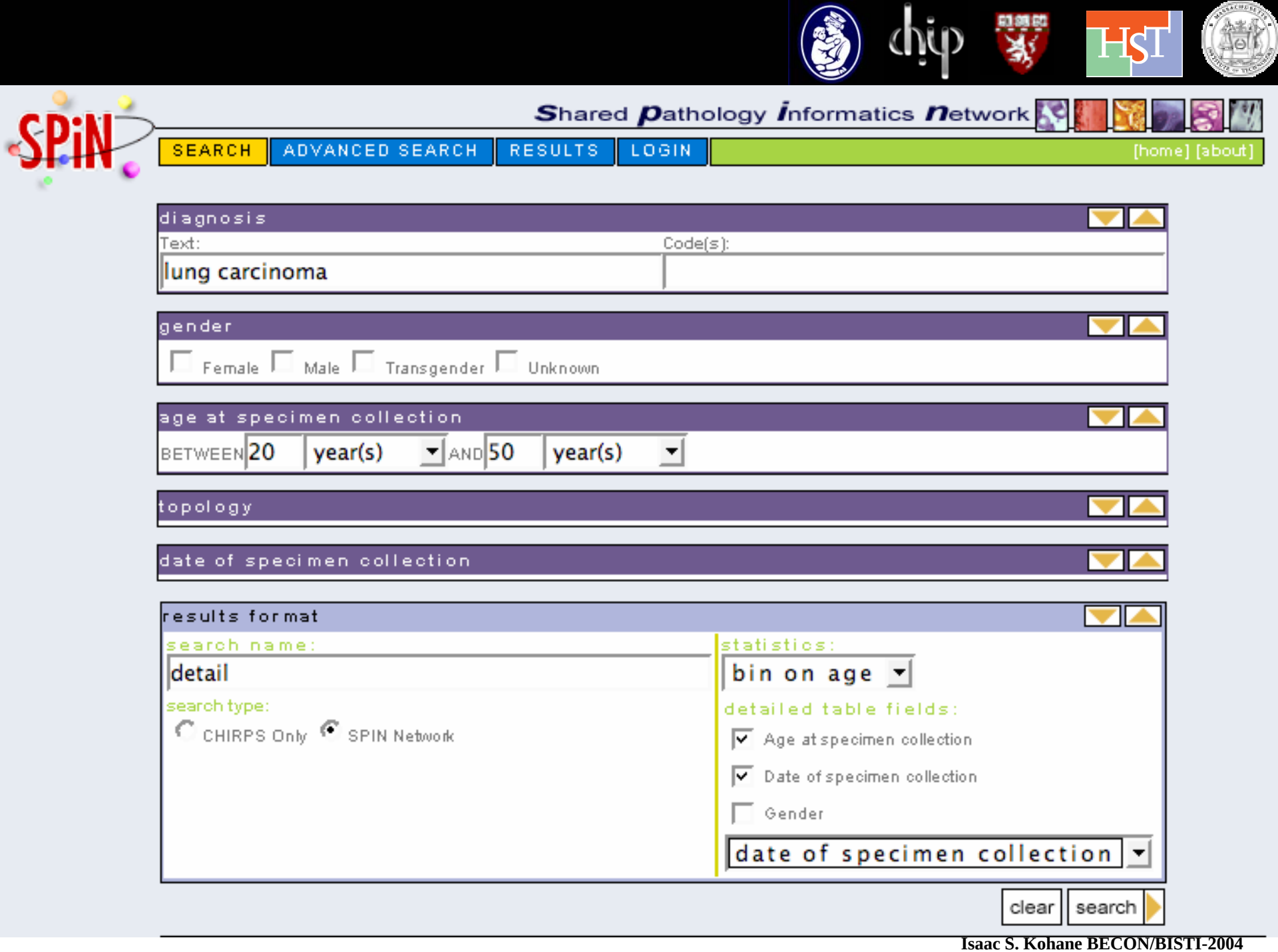
name: detail records: 990 nodes: 7 sent: 2004-05-27 23:45:23

RESULTS | SUBQUERY | NODES

name: records: 0 nodes: 7 sent: 2004-05-27 23:43:57

delete selected results


SPiN
Shared Pathology Informatics Network
SEARCH
ADVANCED SEARCH
RESULTS
LOGIN
[home] [about]
Select ALL
delete selected results
RESULTS
SUBQUERY
NODES
Fetching
name:
records: 990
nodes: 7
sent: 2004-05-27 23:32:25
nodes that responded to your query:
(1) CHIRPS - Beth Israel Deaconess Medical Center
(1) CHIRPS - Brigham & Women's Hospital
(1) CHIRPS - Children's Hospital Boston
(1) CHIRPS - Massachusetts General Hospital
(2) CHIRPS - University of California at Los Angeles Medical Center
(1) Other - unidentifiable
delete selected results


Shared Pathology Informatics Network
SEARCH
ADVANCED SEARCH
RESULTS
LOGIN
[home] [about]
diagnosis
Text:
lung carcinoma
Code(s):
gender
Female
Male
Transgender
Unknown
age at specimen collection
BETWEEN 20 year(s) AND 50 year(s)
topology
date of specimen collection
results format
search name:
detail
search type:
CHIRPS Only
SPIN Network
statistics:
bin on age
detailed table fields:
Age at specimen collection
Date of specimen collection
Gender
date of specimen collection
clear
search

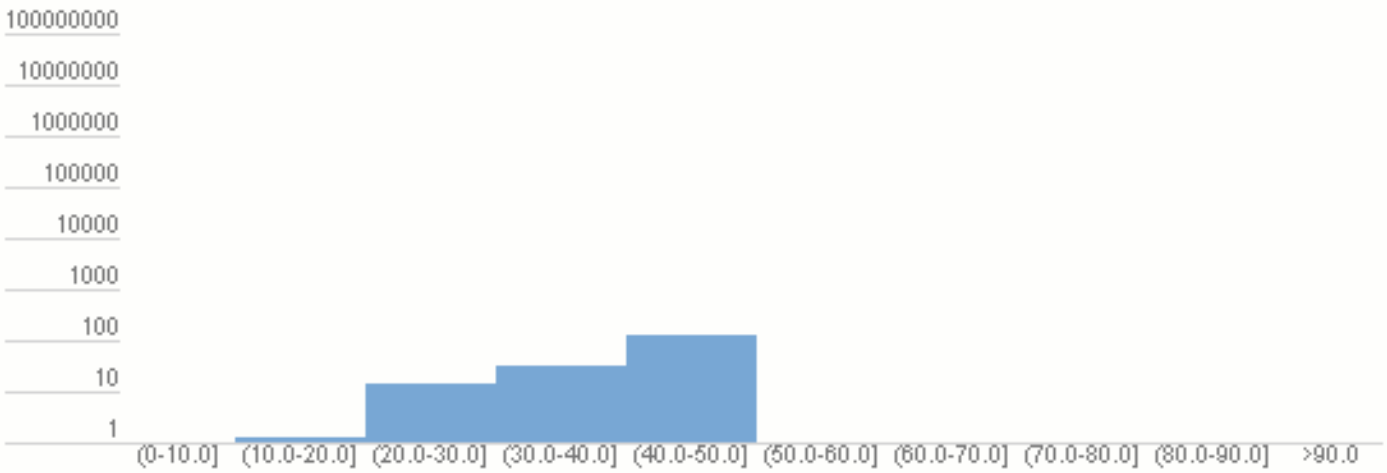
Select ALL

delete selected results

RESULTS | SUBQUERY | NODES | Fetching

name: detail records: 301 nodes: 6 sent: 2004-05-27 23:54:58

age at sample collection



no value:	
(0-10.0]:	
(10.0-20.0]:	2
(20.0-30.0]:	24
(30.0-40.0]:	57
(40.0-50.0]:	218
(50.0-60.0]:	
(60.0-70.0]:	
(70.0-80.0]:	
(80.0-90.0]:	
>90.0:	

detail

Age at Specimen Collection	Tissue Acquisition Date
50	1988-03-04
50	1989-08-10
46	1989-09-28
45	1989-10-26
46	1989-11-30
48	1990-01-11
34	1990-03-09
36	1990-04-11

Interim Conclusion

- Tasteful delegation of control and access enables large data sharing at the scale required of the genomic age.
- A **COMMON VOCABULARY** across our clinical systems is essential.
- A common set of **STANDARDIZED PROTOCOLS** across distributed systems is the only way to scale to the national level.
- Taking seriously the building of **LEVELS of ABSTRACTION** and **ABSTRACTION BARRIERS** is the key to large system construction

But if medical records are incomplete, how do we get the entire patient history/phenotype?

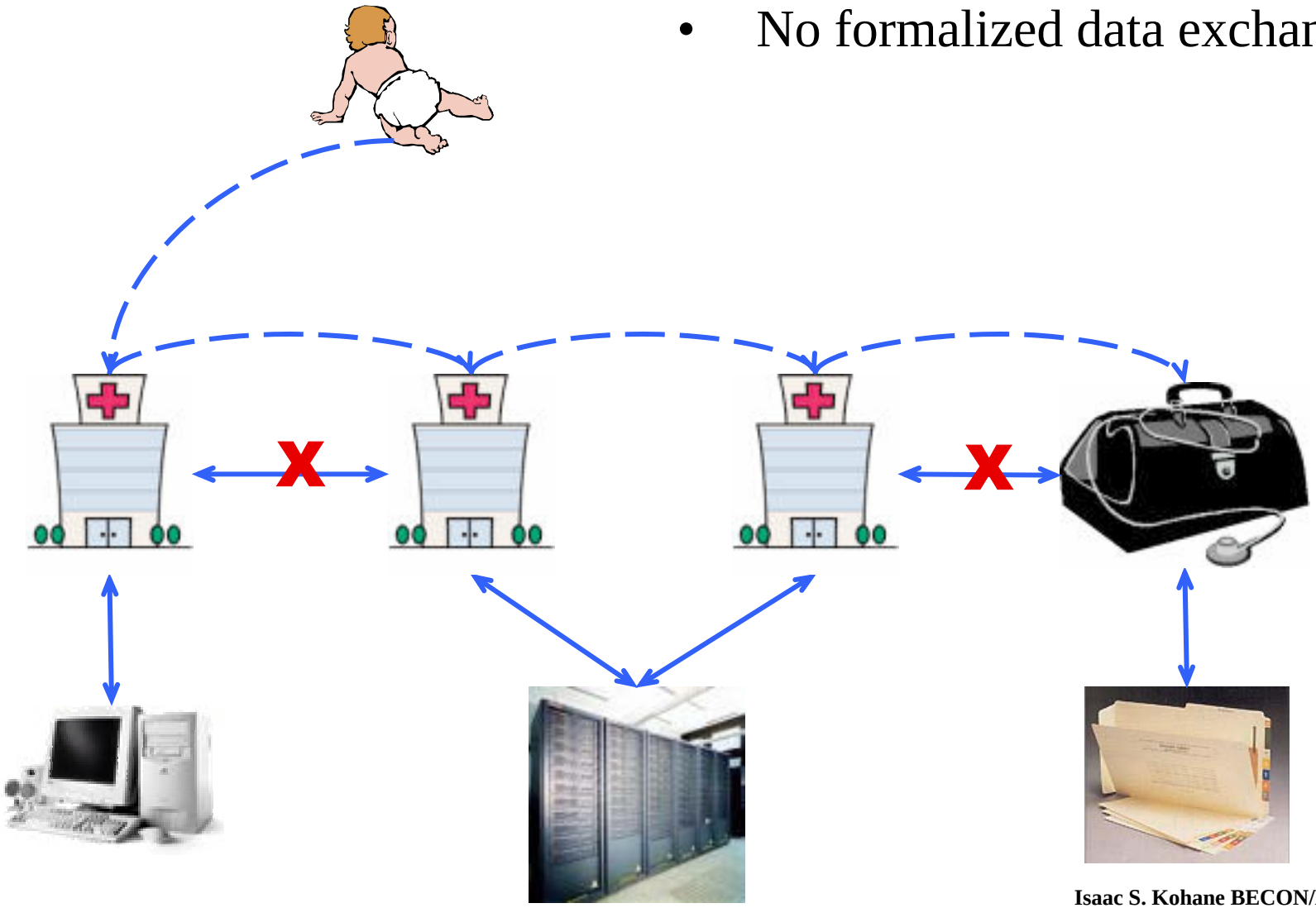
- ✓ Daily medications
- ✓ Exercise level
- ✓ Adverse events with over the counter medications
- ✓ Absence from work
- ✓ …

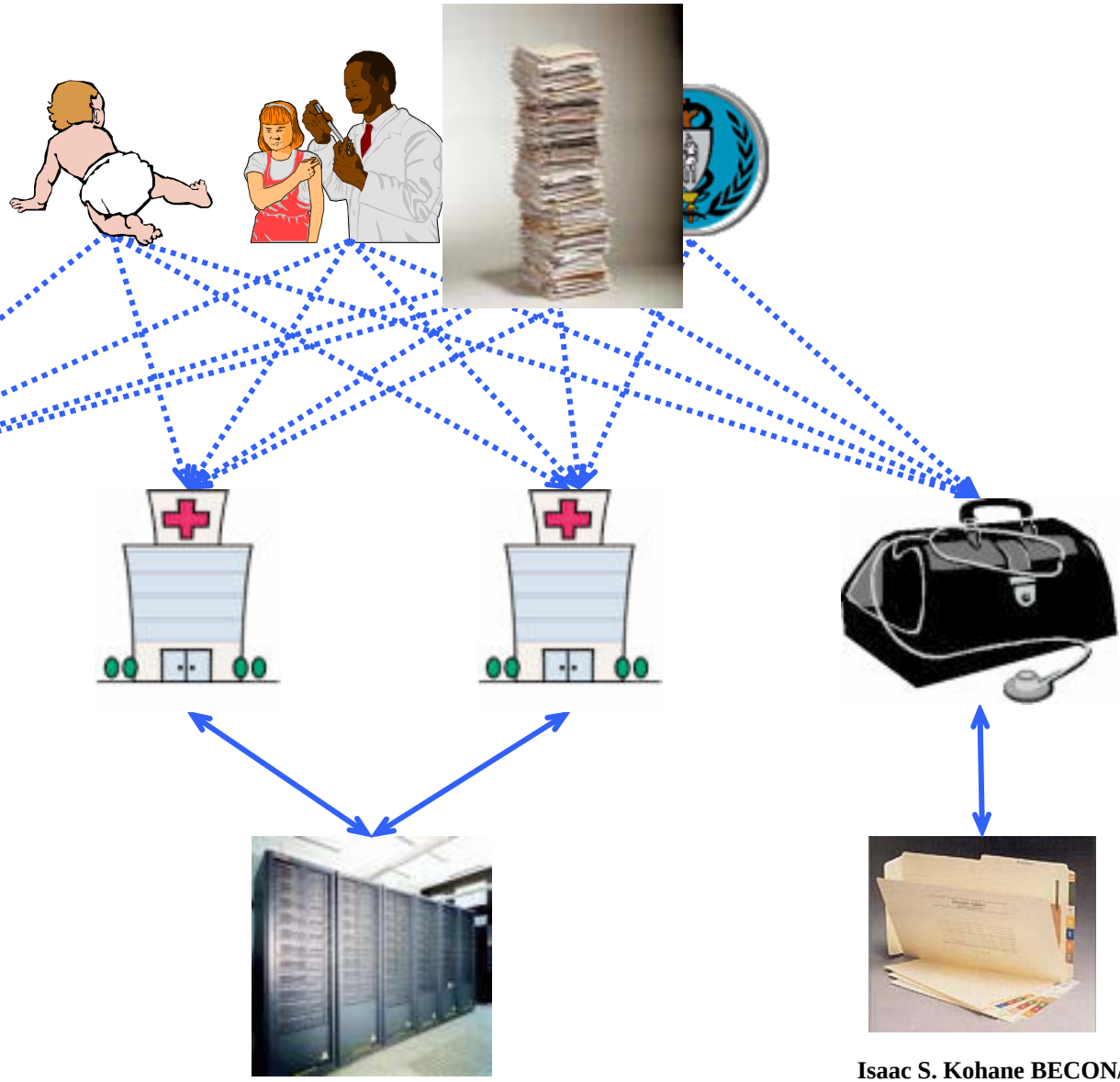
- Given persistent failure to get institutional buy-in
- Can we harness the patient?

The Problem

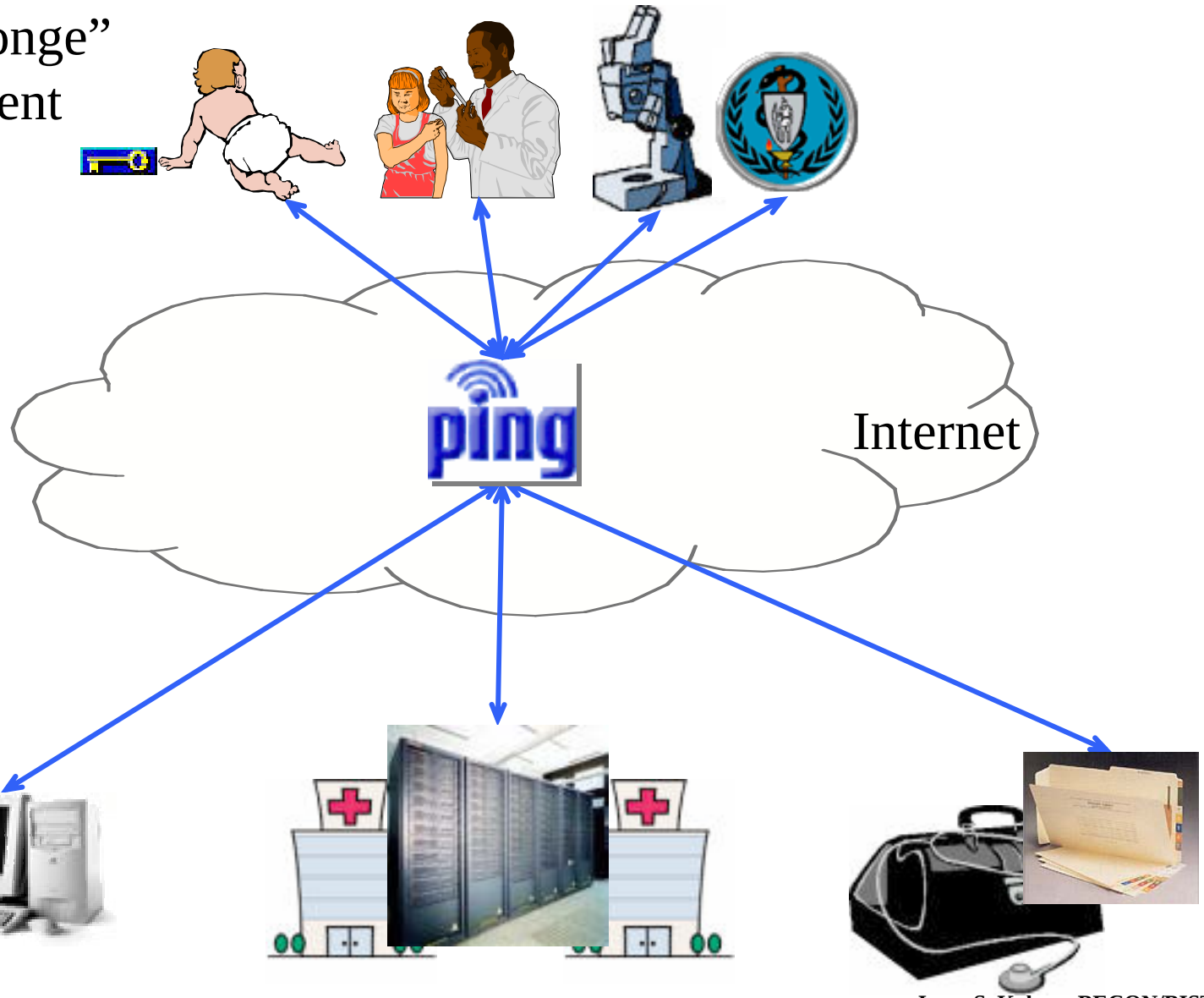
- Medical records are fragmented across multiple institutions
- There's no unified view of a patient's record
- Patients have difficulty accessing their medical information

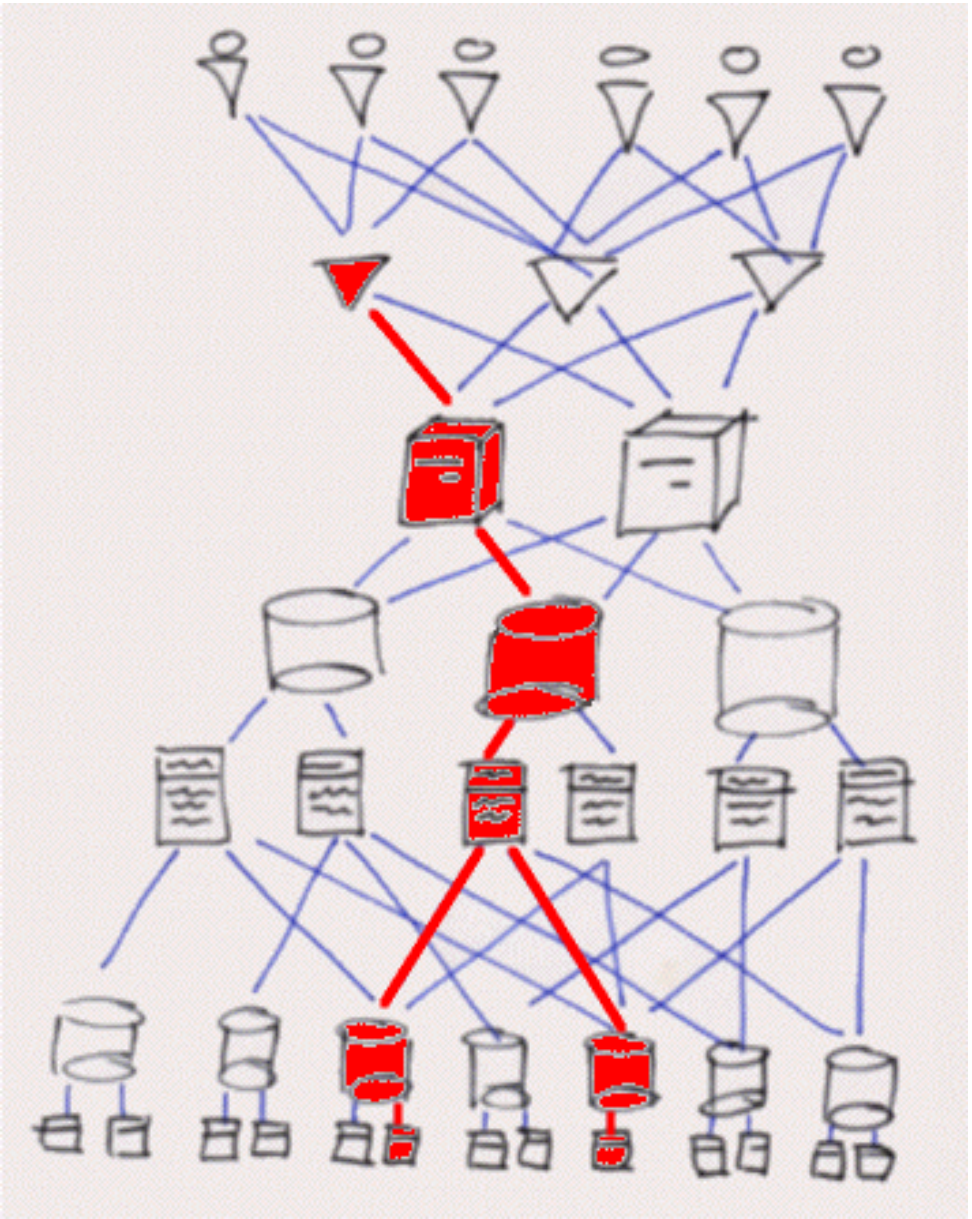
- Highly mobile patients
- No formalized data exchange



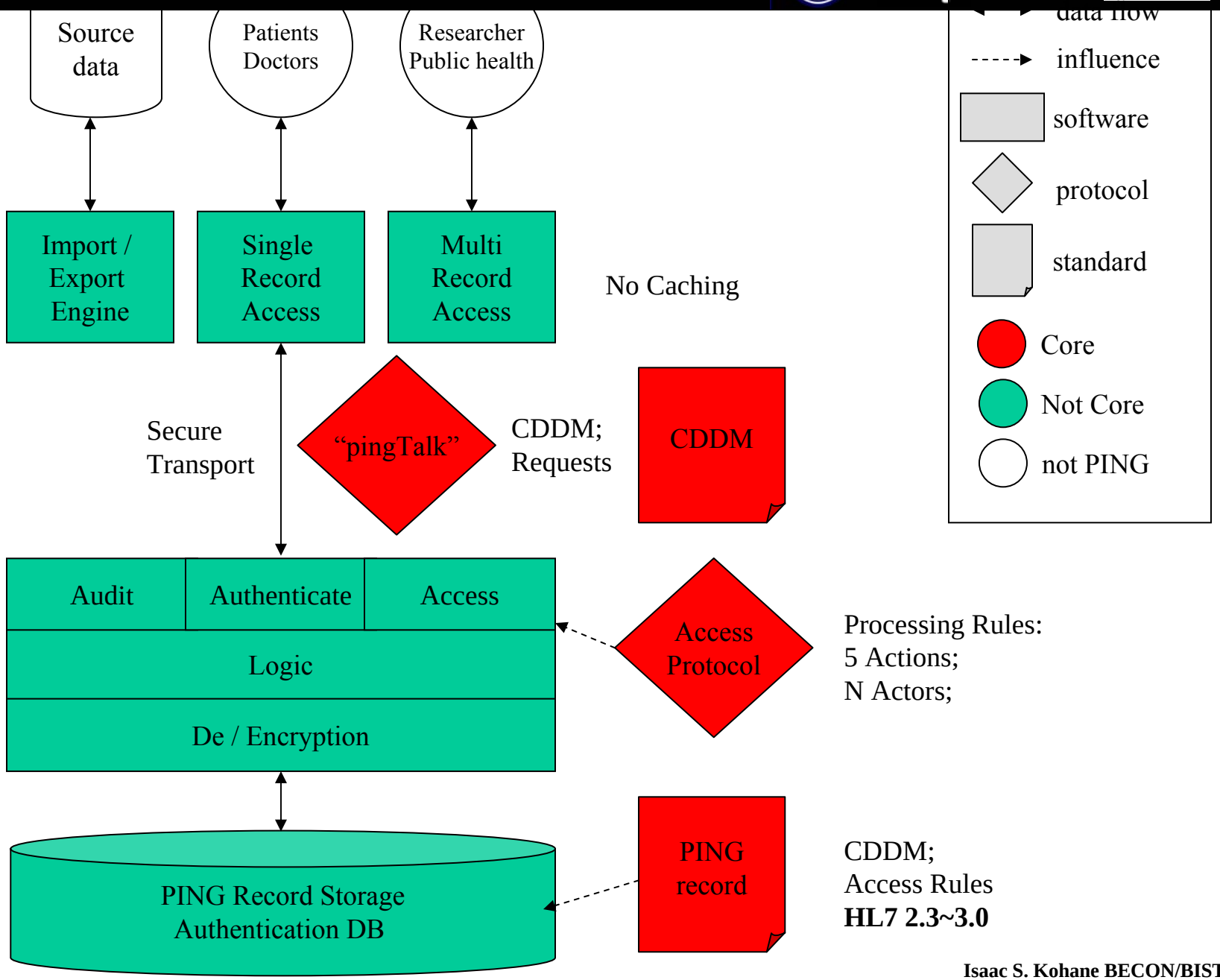


- "Data sponge"
- Independent





- Clients access the PING server, which in turn, accesses the patient's record
- The PING server authenticates the agent and can perform the following four atomic operations:
  - ✓ Create
  - ✓ Read
  - ✓ Modify
  - ✓ Annotate
- The server manages the PING records which may be stored by *service bureaus* or on any server of the patient's choice
- Medical record data and images may be stored with the PING record, or the record may contain pointers to information and images stored elsewhere.

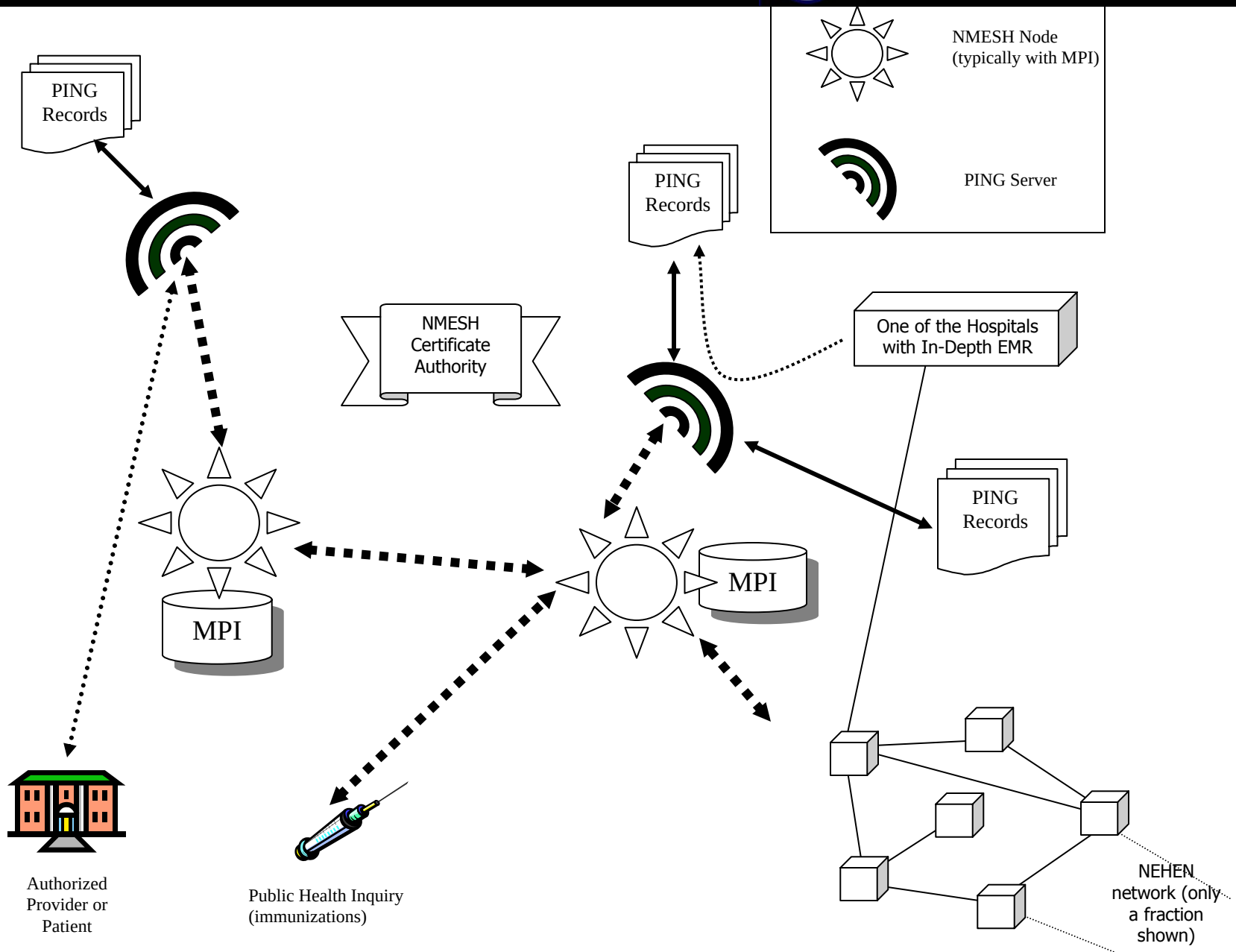

Source data
Patients Doctors
Researcher Public health
data flow
influence
software
protocol
standard
Core
Not Core
not PING
Clients
Gateways
Import / Export Engine
Single Record Access
Multi Record Access
No Caching
Secure Transport
"pingTalk"
CDDM; Requests
CDDM
Server
Audit
Authenticate
Access
Logic
De / Encryption
Access Protocol
Processing Rules:
5 Actions;
N Actors;
Storage
PING Record Storage
Authentication DB
PING record
CDDM;
Access Rules
HL7 2.3~3.0

Personal Internetworked Notary and Guardian

- For large-scale disasters (PING-Response)
- For patients poorly tied into the healthcare system.
  - ✓ PING-Citizen (Canada)
- For immunization record
  - ✓ PING 'baby book'
- ***PING Genome***

And Let's Bring it All Together

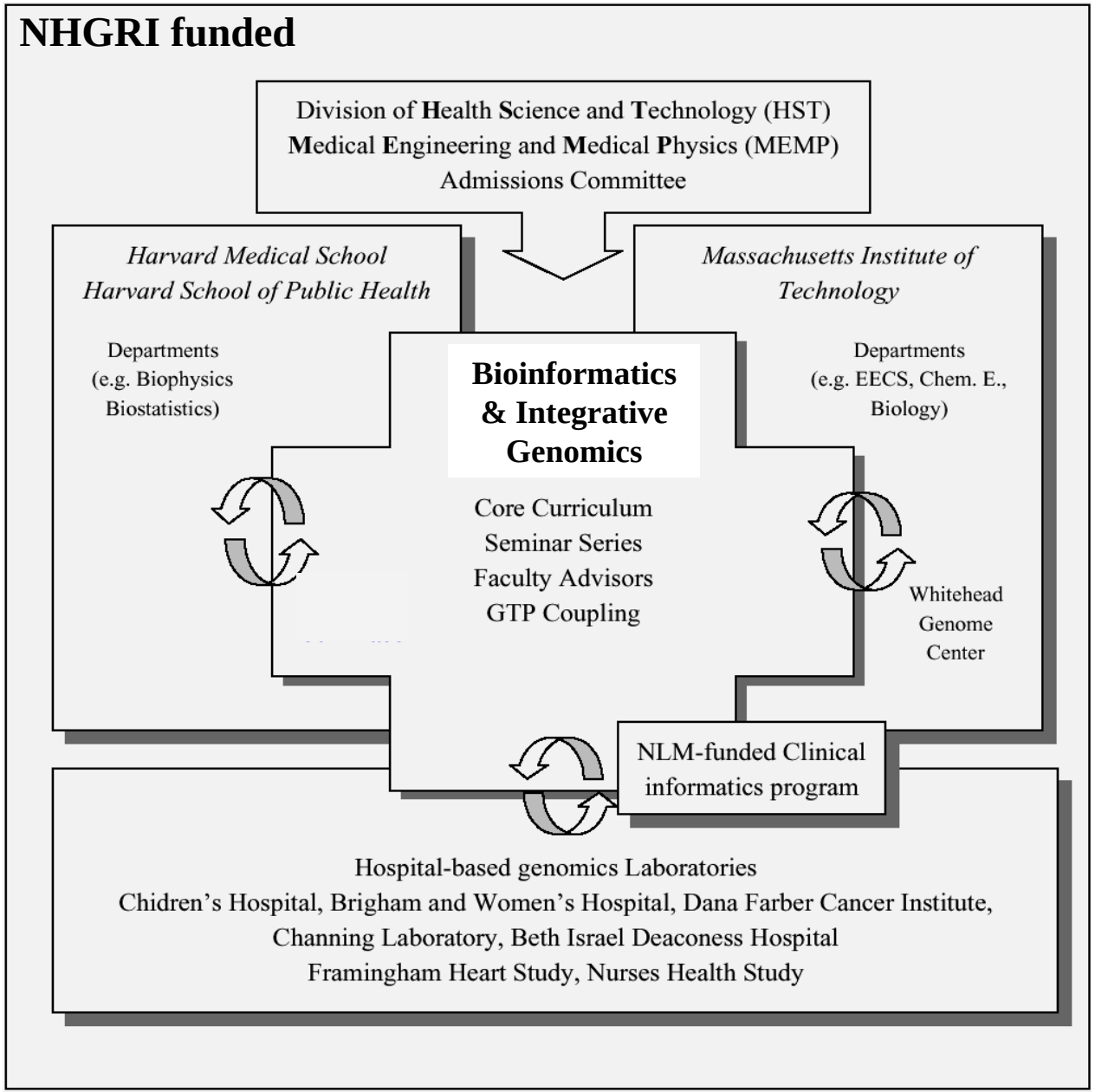
- The National Multi-Protocol Ensemble for Self-scaling Systems for Health (NMESH)
- Cover an entire region (Northeast)
- Use HL7 data flow from multiple hospitals into PING records
- Use SPIN to provide a peer-peer mechanism to query all the PING records
- Apply for research, clinical care and disaster management.


NMESH Node
(typically with MPI)
PING Server
PING
Records
PING
Records
NMESH
Certificate
Authority
One of the Hospitals
with In-Depth EMR
PING
Records
MPI
MPI
Authorized
Provider or
Patient
Public Health Inquiry
(immunizations)
NEHEN
network (only
a fraction
shown)

Summary

- **STANDARDIZED VOCABULARIES** and **PROTOCOLS** are the essential **glue** to allow
  - ✓ Linking of disparate data types and sources
  - ✓ Leveraging existing methodologies and tools
    - ☞ E.g. homology maps and phenotypic classification
- **Lightweight social engineering is necessary**
  - ✓ Preserve local autonomy
  - ✓ Enable local curation
  - ✓ Represent the first and most important steps to unifying clinical informatics to bioinformatics and the tool builders and discoverers.
- **Do not try to solve all data representations problems**
  - ✓ **Tasteful, partial solutions will incrementally bring you to your goal**
  - ✓ **All out effort will bring you nowhere.**

Bioinformatics and Integrative Genomics (BIG)



Thank you

Isaac_Kohane@harvard.edu