Using SNOMED CT for Translational Genomics Data Integration

Joel Dudley, David Chen, Atul Butte

David Chen
KR-MED
June 2, 2008

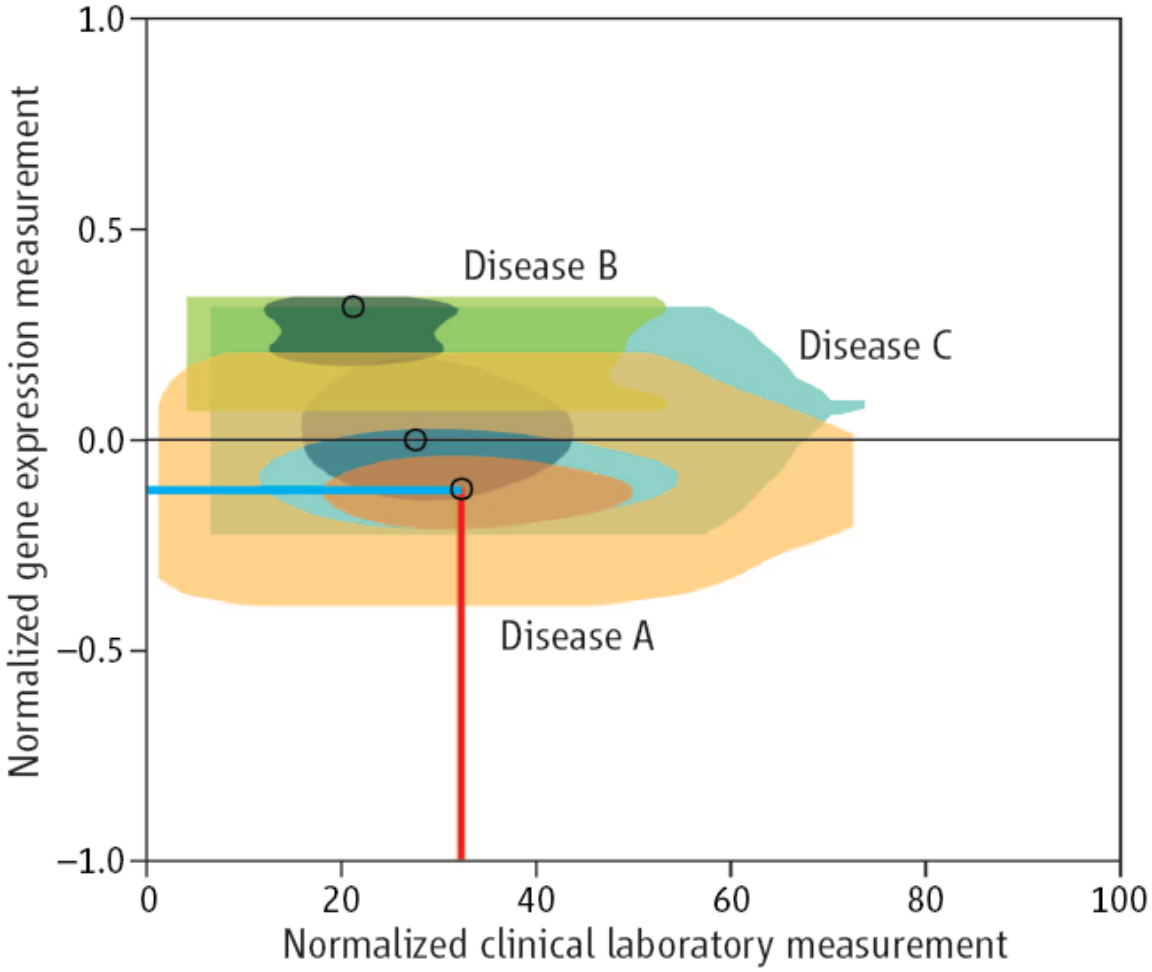
STANFORD
SCHOOL OF MEDICINE
Stanford University Medical Center

Butte Lab
Stanford Center for Biomedical Informatics Research

Motivation

- For over 10 years analytical methods have been applied to relate gene sequence and molecules to particular diseases.
- So many diseases have been studied that we can find common gene expression changes among them using publically available data repositories.
- Parallel measurements of physiological variables have also been successfully linked to genetic markers. (Stoll et al. Science 2001)
- Vast amounts of physiological measurements lie within electronic medical records

Finding trends between genes and clinical measurements



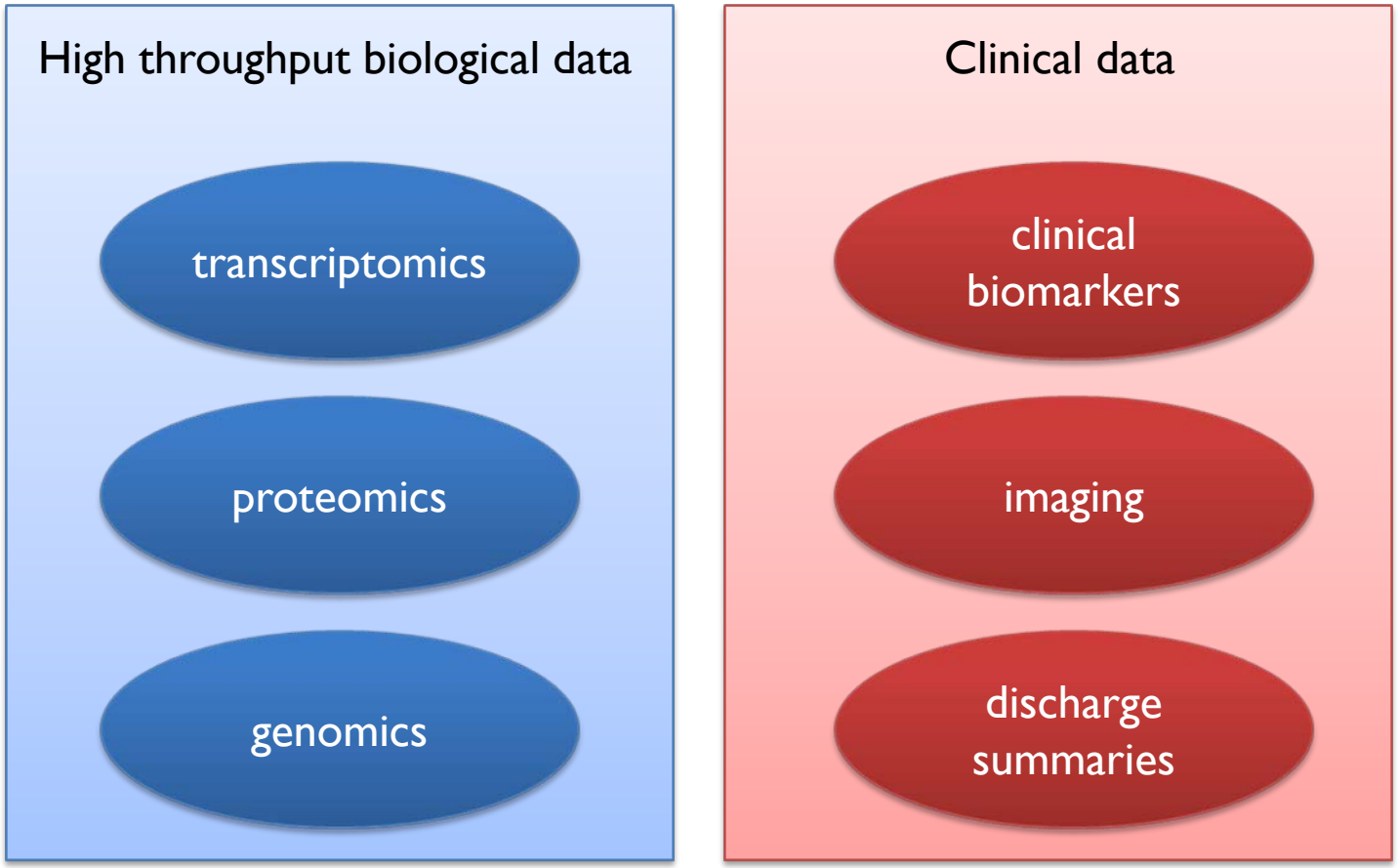
A.J Butte. The Ultimate Model Organism. Science. 2008

SNOMED CT enables clinical and genomic data integration

- Clinical physiological measurements are rarely collected in parallel with genetic or genomic measurements
- We wish to take advantage of petabytes of clinical measurements on patients who have not had genetic or genomic measurements taking
- SNOMED CT enables this type of analysis

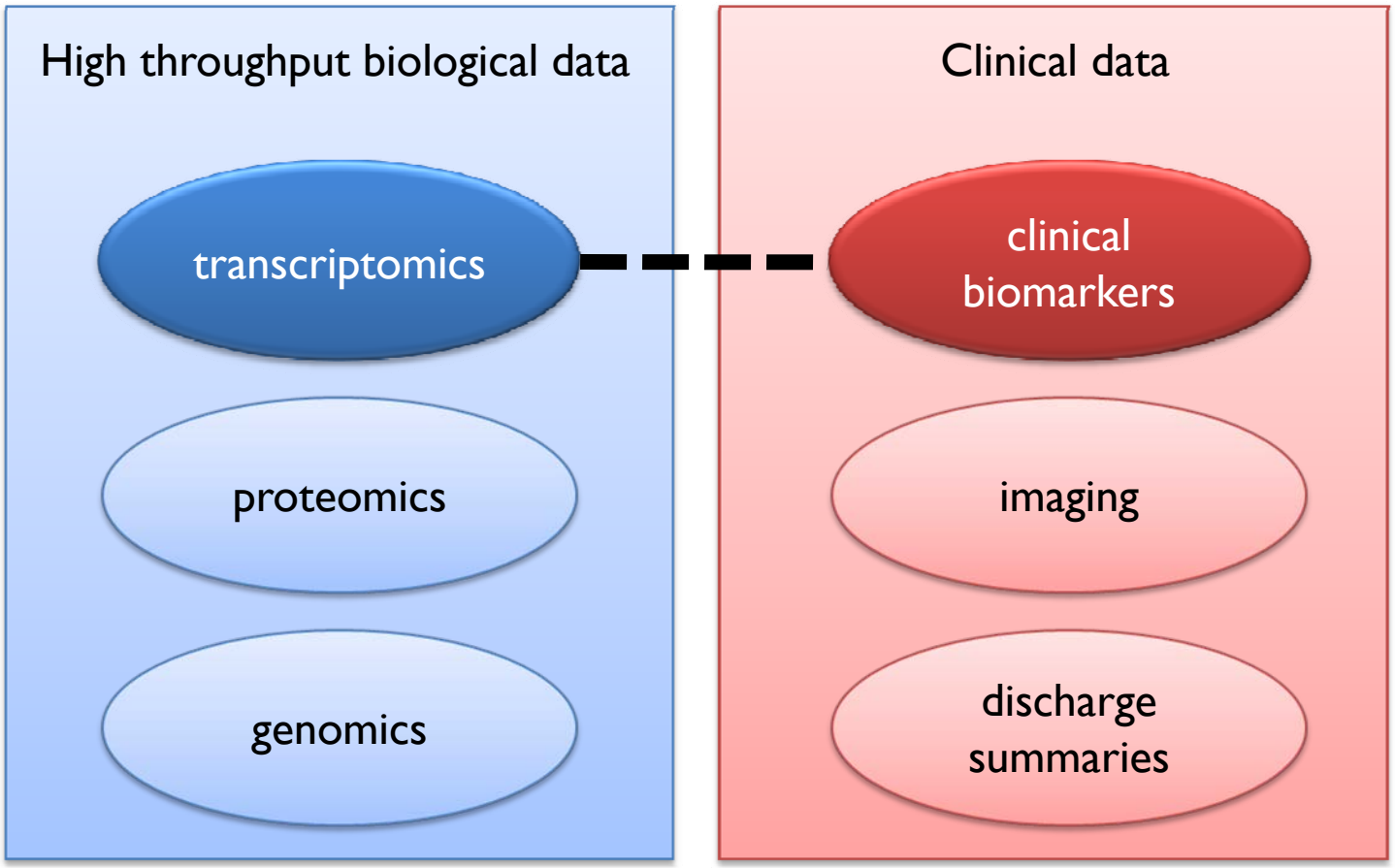
Bridge the divide between high throughput biological data and clinically relevant data

Samples measured in high throughput biological data are not from the same patients that clinical data is gathered from



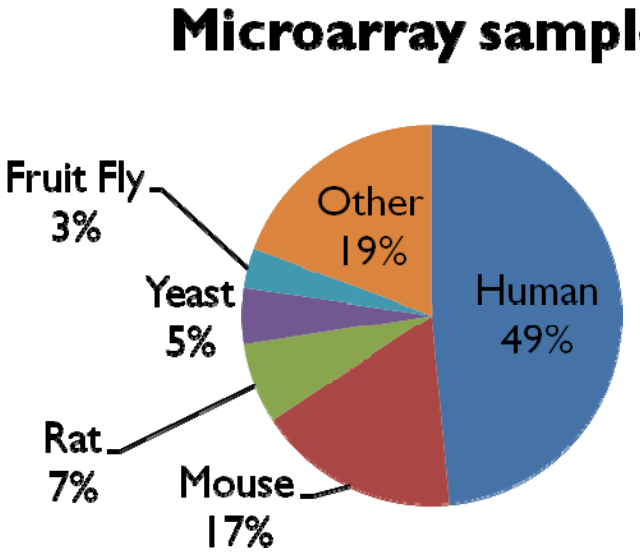
Bridge the divide between high throughput biological data and clinically relevant data

Samples measured in high throughput biological data are not from the same patients that clinical data is gathered from



NCBI Gene Expression Omnibus (GEO) contains a wealth of clinically relevant microarray data

- 7264 Experiments (GEO Series)
- 188,539 Microarray Samples
- 3,146 Array platforms
- More than 10,000 samples annotated with "disease state" or "infection"



GEO statistics as of May 29, 2008

8,702 experiments
227,156 microarray samples
4,709 platform

Mapping GEO Data Sets (GDS) to SNOMED CT

GDS Summary

Accession:	GDS10 ☞ View Expression (GEO profiles)
Title:	Type 1 diabetes gene expression profiling
Citations:	Eaves IA, Wicker LS, Ghandour G, Lyons PA et al. Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res* 2002 Feb;12(2):232-43. PMID: 11827943

↓ NCBI eUtils

MeSH

MH - Animals
MH - Diabetes Mellitus, Type 1*/genetics
MH - *Disease Models, Animal
…

↓ UMLS MRCONSO table

UMLS CUI for each mesh heading

Dudley, J. Butte, AJ. Enabling Integrative Genomic Analysis of High-Impact Human Diseases Through Text Mining. Pacific Symposium on Biocomputing 2008.

Mapping GEO Data Sets (GDS) to SNOMED CT

UMLS CUI for each mesh heading

UMLS MRSTY table

Examine semantic type: "Injury or poison", "pathological function", "disease or syndrome", etc.

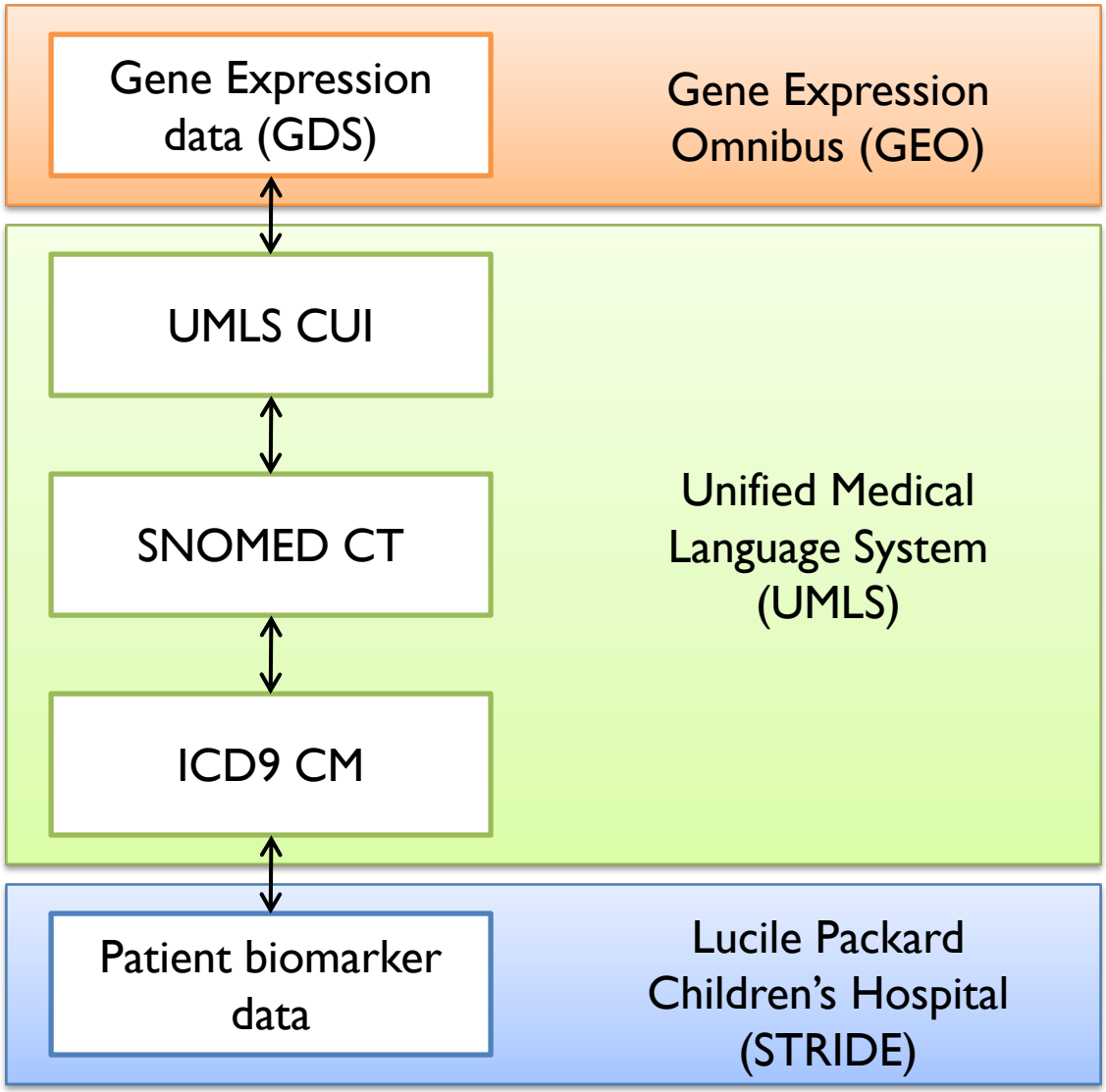
Disease related GDS

4 assigned subsets			
Samples	Type		Description
☑ (6)	☑	disease state	type 2 diabetes
☑ (6)		disease state	non-diabetic
☑ (6)	☑	age	8 week
☑ (6)		age	16 week

Use NLP methods to map free-text annotation of GDS subsets to SNOMED CT disease terms using UMLS

Dudley, J. Butte, AJ. Enabling Integrative Genomic Analysis of High-Impact Human Diseases Through Text Mining. Pacific Symposium on Biocomputing 2008.

Data integration overview



De-identified clinical data can be used for translational data integration

- Stanford Translational Research Integrated Database Environment (STRIDE) contains de-identified patient data from Lucile Packard Children's hospital.
- 49,414 patients
- 9,997 ICD9 CM codes
- Over 7.5 million clinical biomarker measurements

Integrating GEO disease subsets with Clinical ICD9 CM

- SNOMED CT identifiers are translated to their relevant ICD9 CM codes via MRMAP
- The Lucile Packard Children's Hospital database is queried for patients diagnosed with any of the ICD9 CM codes
- GDS subsets with mappings to SNOMED CT disease CUIS were joined to ICD9 CMs that mapped to the same CUI

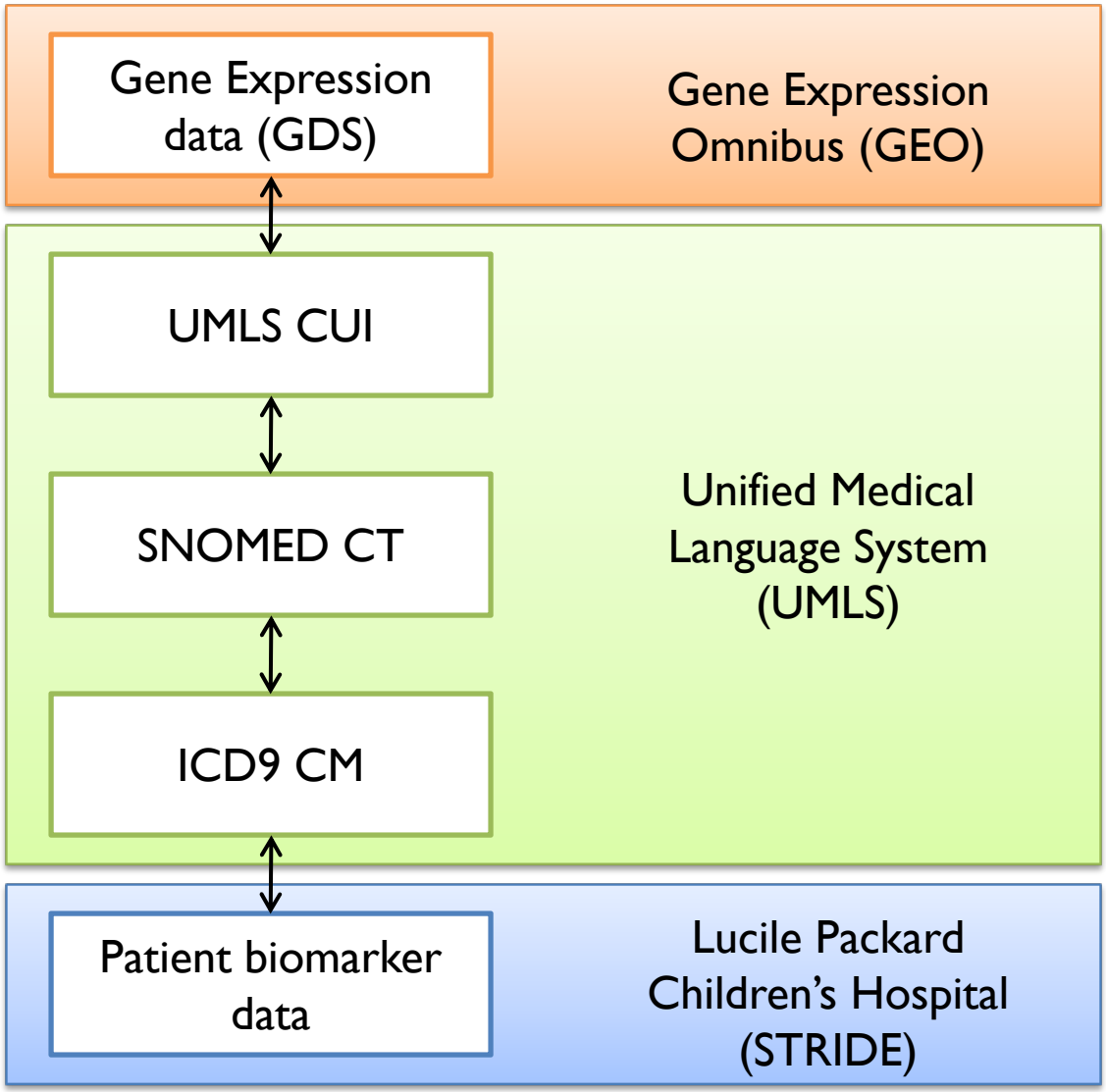
Why do we use SNOMED?

- SNOMED allows for greater coverage of ICD9 CM codes
- We can traverse the SNOMED hierarchy depending on our desired level of resolution

Results

- 737 GEO Data Sets that were related to human disease
- 238 disease concepts were associated with GDS subsets
- 29,541 microarray samples were coded with SNOMED CT identifiers
- Note, we only included GDS that compared disease state to normal state
- 13,452 patients (of 49,414) mapped to 211 (of 238) of the disease concepts

Data integration overview



Results of mapping GEO data sets to patients

Disease	SNOMED Terms	ICD9CM Terms	Patients
Allergic asthma	1	1	2240
Asthma	1	1	2240
Allergic asthma NEC	1	1	2240
Esophageal Reflux	1	1	1895
H. Pylori infection	1	2	1322
Colitis	1	1	1299
Primary Hypertension	1	1	1017
Hypertension	1	1	1017
Obesity	2	1	1010
Type 1 Diabetes	1	1	843

Caveats

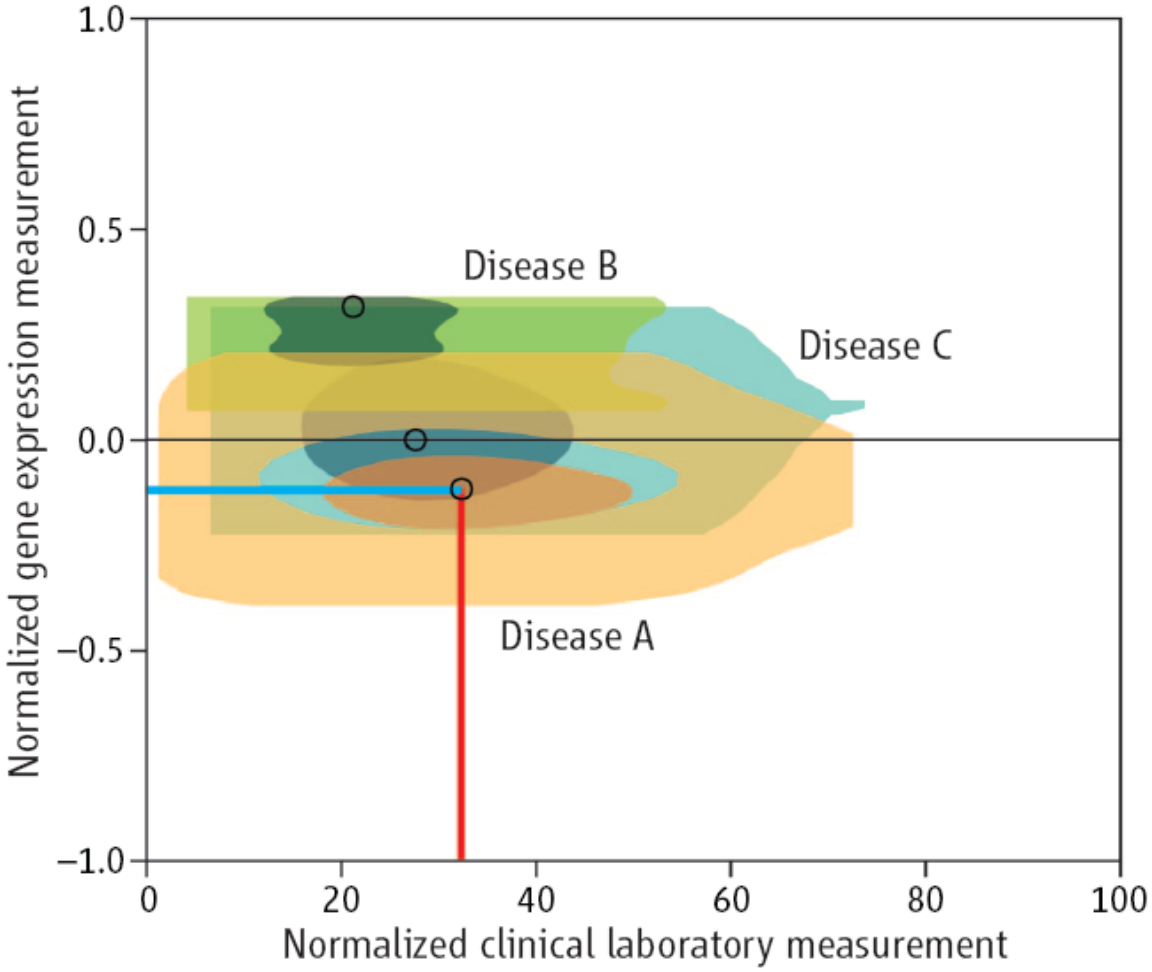
- Some microarray data could not be mapped to clinical laboratory data. This reflects the fact that only pediatric data was used
  - Parkinson's Disease
  - Alzheimer's Disease
- The study did not apply the same technique to alternative disease terminologies to offer quantitative comparison
- The investigation offers no statistical characterization to assess overall quality and reliability

Some ambiguities between ICD9CM and SNOMED remain

- Distinct SNOMED CTs may map to the same ICD9 CM
- Single UMLS CUIs can map to multiple SNOMED terms
- There is some ambiguity between SNOMED CT and ICD9 CM mappings

Disease	SNOMED Terms	ICD9CM Terms	Patients
Follicular lymphoma	4	3	136
Hamman-Rich Syndrome	4	2	18
Mycobacterial Infection	3	2	26
Mixed hyperlipidemia	3	2	90
Hepatoma	3	2	67
Fetal alcohol syndrome	3	1	10

Finding trends between genes and clinical measurements



A.J Butte. The Ultimate Model Organism. Science. 2008

Conclusions

- Using this method we can now integrate patient clinical data with microarray data at a population level
- This integration does not require parallel measurement acquisition
- Nothing was explicitly encoded with SNOMED
- SNOMED acts mediator between different systems that we can't join otherwise
- This method can be extended to other types of clinical data and molecular data

Thanks to

- Buttle Lab:
  - Atul Bute, MD, PhD
  - Joel Dudley
  - Annie Chiang, PhD
  - Rong Chen, PhD
  - Sangeeta English, PhD
  - Alex Morgan
  - Shai Shen-Orr, PhD
  - Marina Sirota
  - Shiv Venkatasubrahmanyam, PhD
- STRIDE:
  - Philip Constantinou
  - Todd Ferris, MD
  - Henry Lowe, MD
  - Susan Weber, PhD
- Computing Support:
  - Alex Skrenchuck
- Funding:
  - Lucile Packard Foundation for Children's Health
  - National Library of Medicine (K22 LM008621)
  - National Library of Medicine (T15 LM007033)
  - National Institute of General Medical Sciences (R01 GM079719)
  - National Human Genome Research Institute (P50 HG003389)
  - Howard Hughes Medical Institute
  - Pharmaceutical Research and Manufacturers of America

Thank you

Questions?

davidpchen@stanford.edu