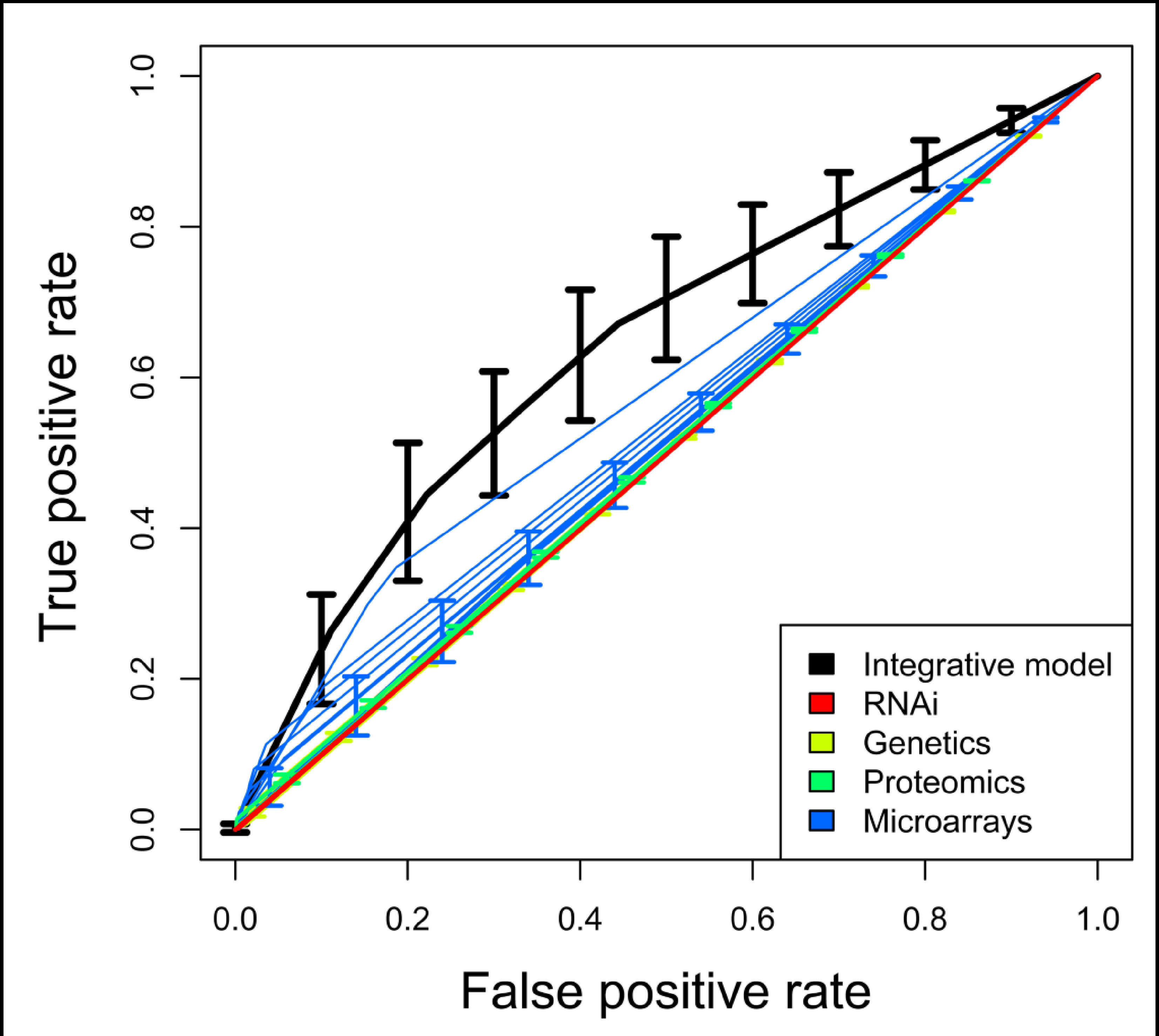
Translating publicly-available molecular data into new biomarkers and therapeutics

Joel Dudley
Division of Systems Medicine
Stanford University School of Medicine

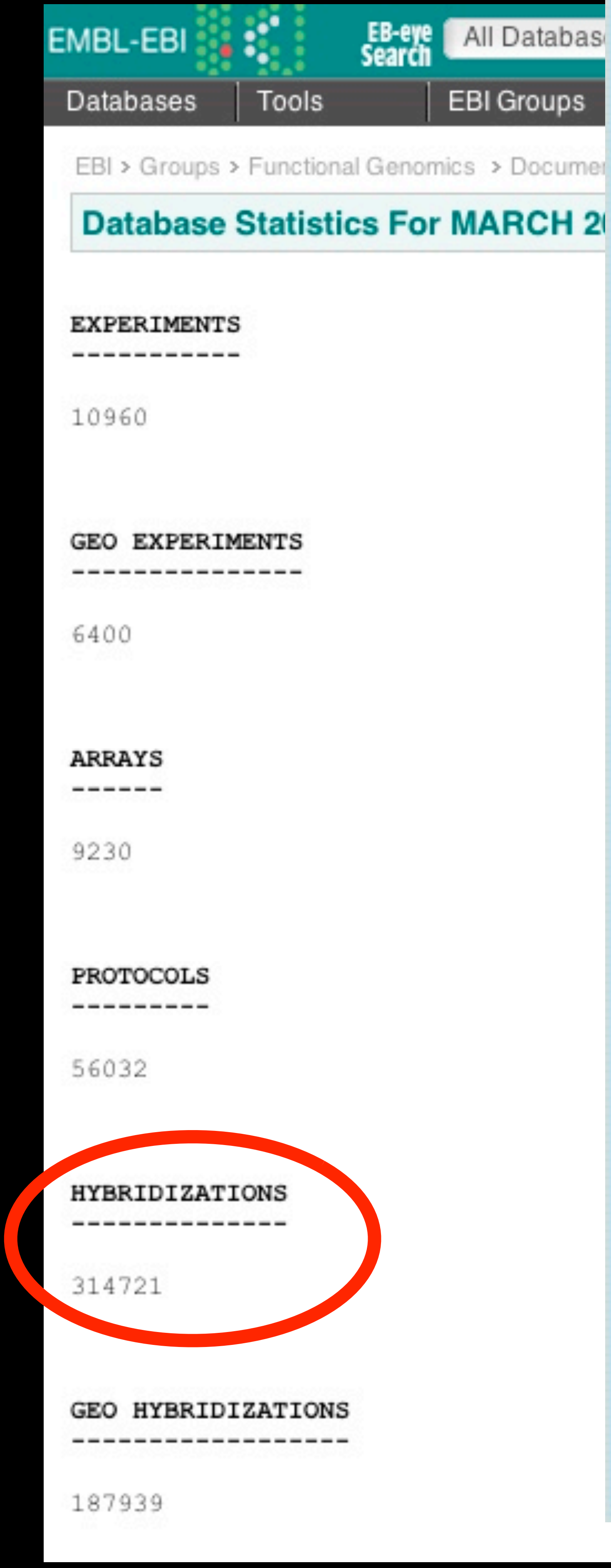
Lucile Packard
Children's Hospital
AT STANFORD

Butte Lab

More Data Wins

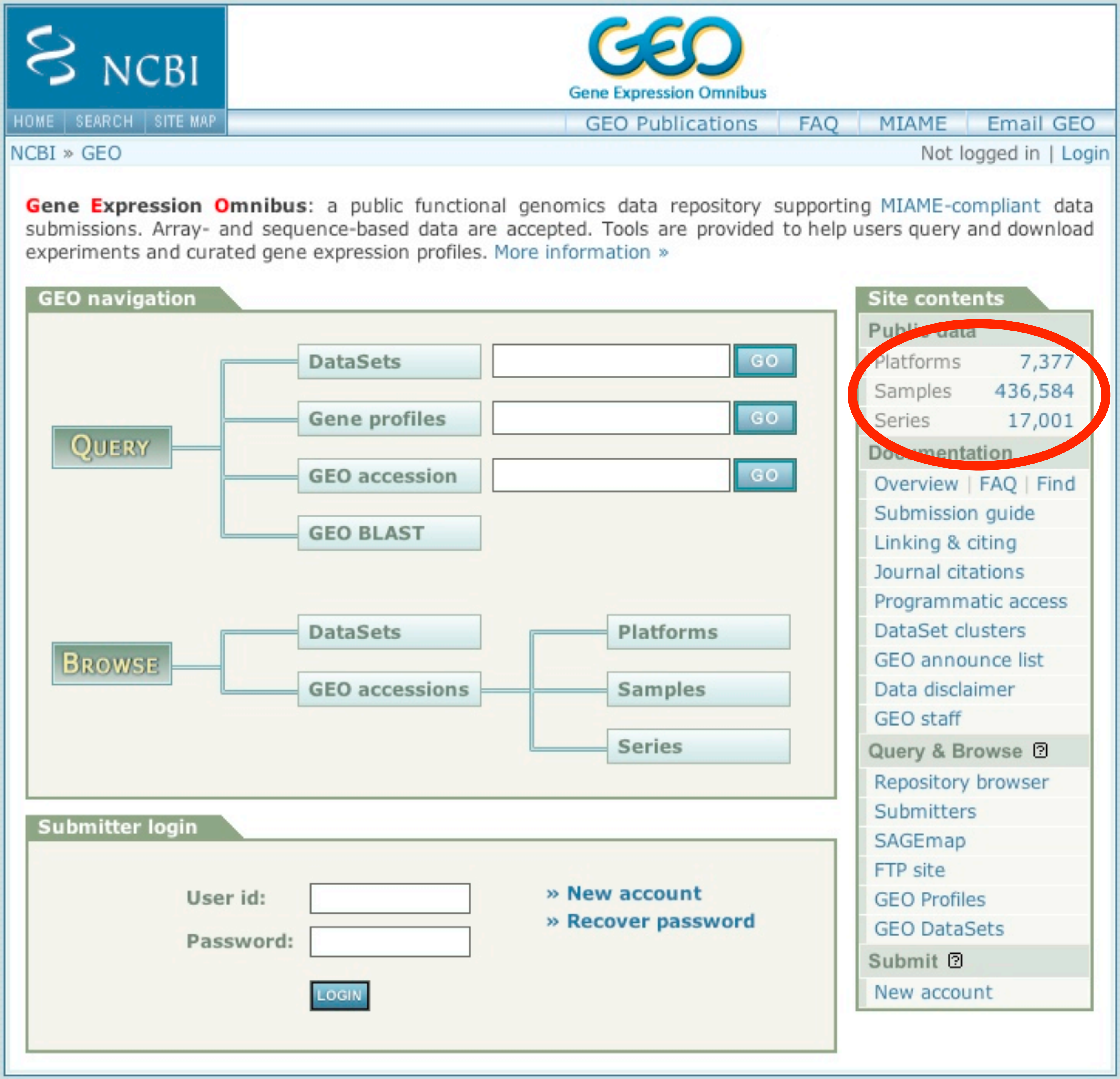


English S and Butte AJ. "Evaluation and Integration of 49 Genome-wide Experiments and the Prediction of Previously Unknown Obesity-related Genes." Bioinformatics (2007) vol. 23 (21) pp. 291



Database Statistics For MARCH 2

```
EXPERIMENTS
-----------

10960


GEO EXPERIMENTS
---------------

6400


ARRAYS
------

9230


PROTOCOLS
---------

56032


HYBRIDIZATIONS
--------------

314721


GEO HYBRIDIZATIONS
------------------

187939
```



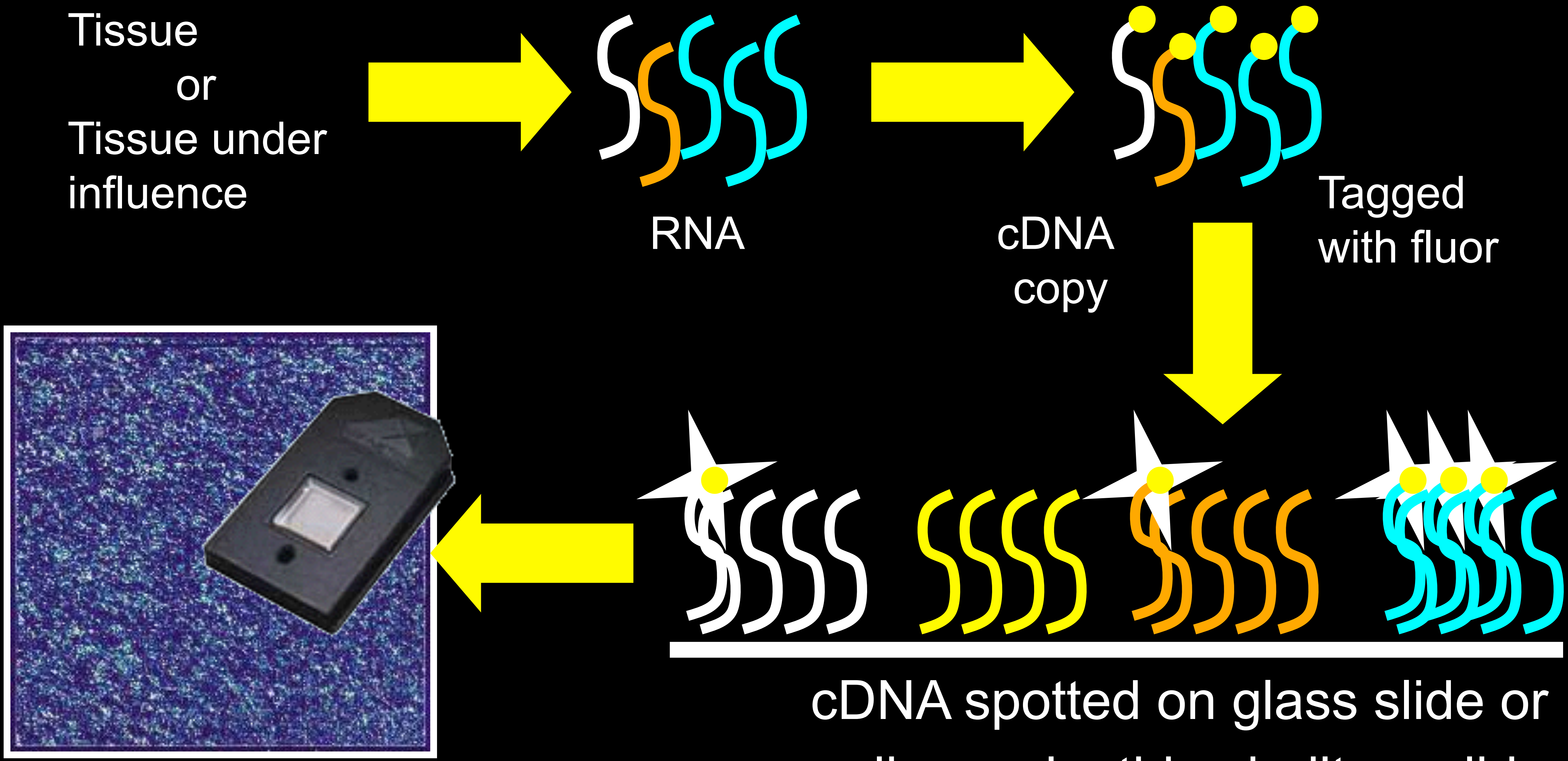
Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

Public data

Platforms	7,377
Samples	436,584
Series	17,001

AFFYMETRIX
GeneChip®
Human Genome
U133A 2.0
SUNIT- J-4
P/N 520065
Lot # 4022127
Exp. Date 06/12/07
For Research Use Only

RNA expression detection chips



Tissue
or
Tissue under
influence

RNA

cDNA
copy

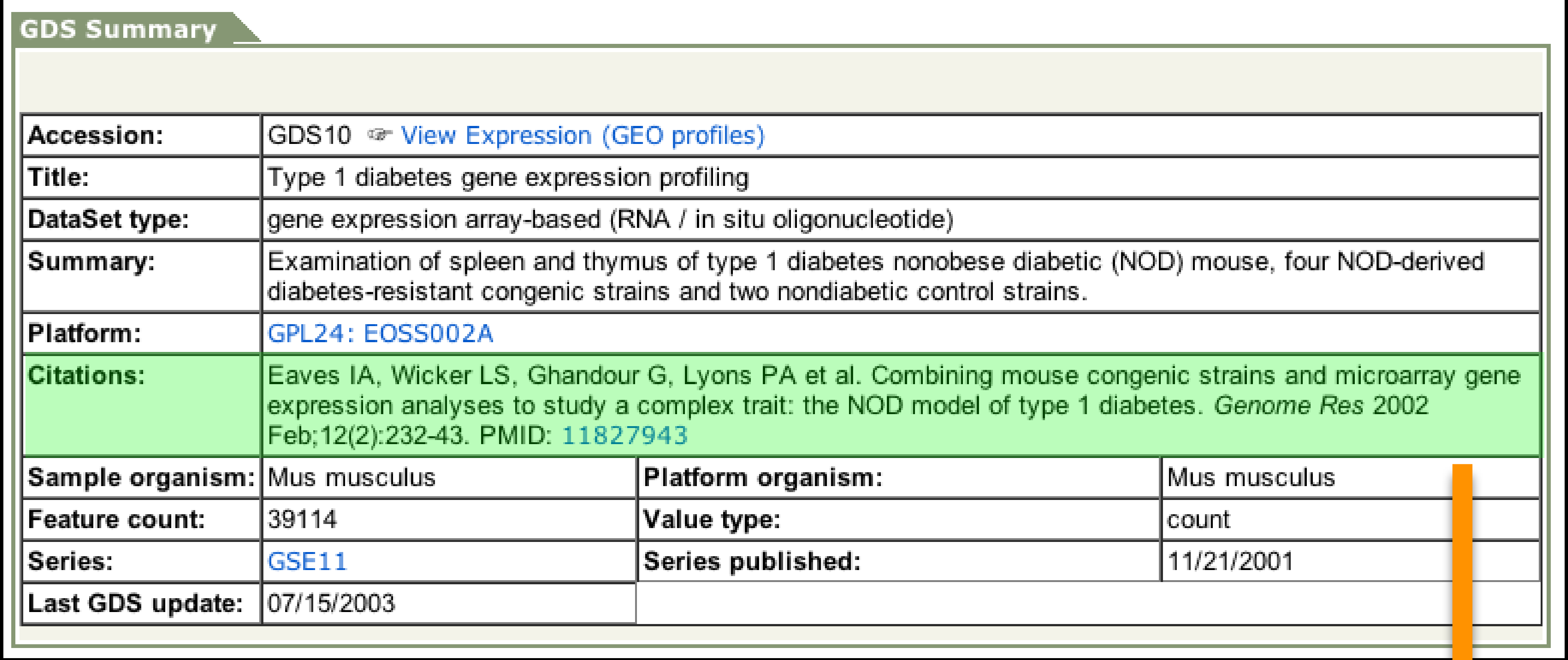
Tagged
with fluor

cDNA spotted on glass slide or
oligonucleotides built on slide

- Genome-wide, quantitative
- Commodity items
- International repositories of data

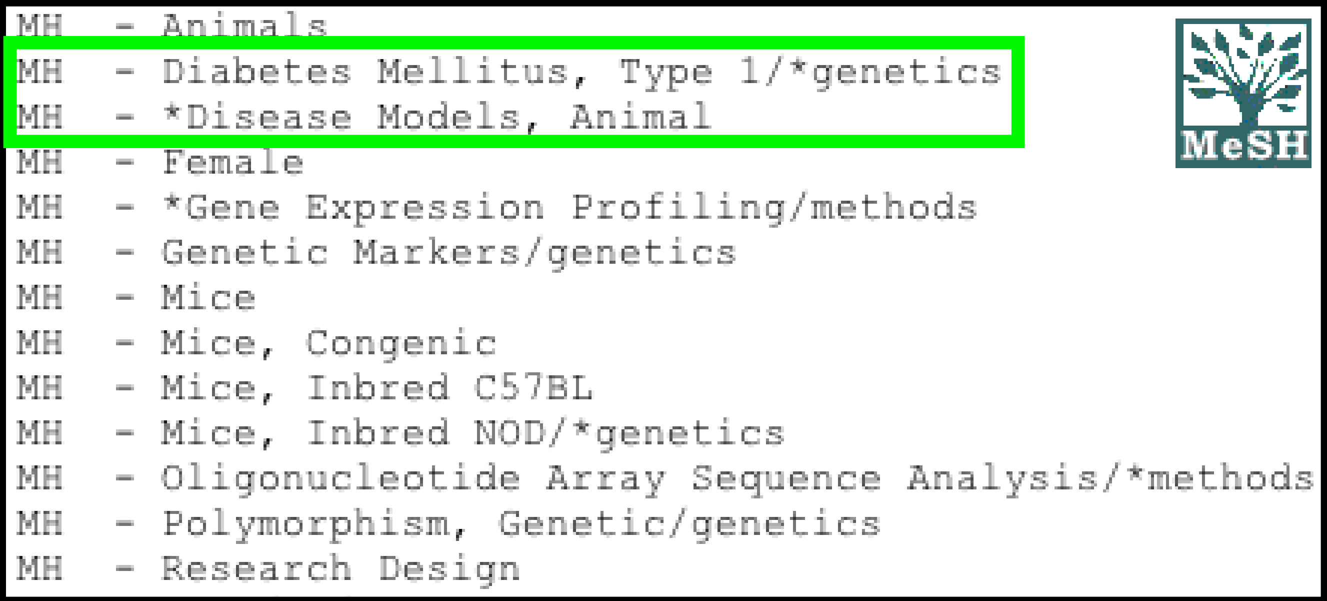
Schena M, et al. PNAS 93:10614 (1996).

Nature Genetics, 21: supplement (Jan 1999).



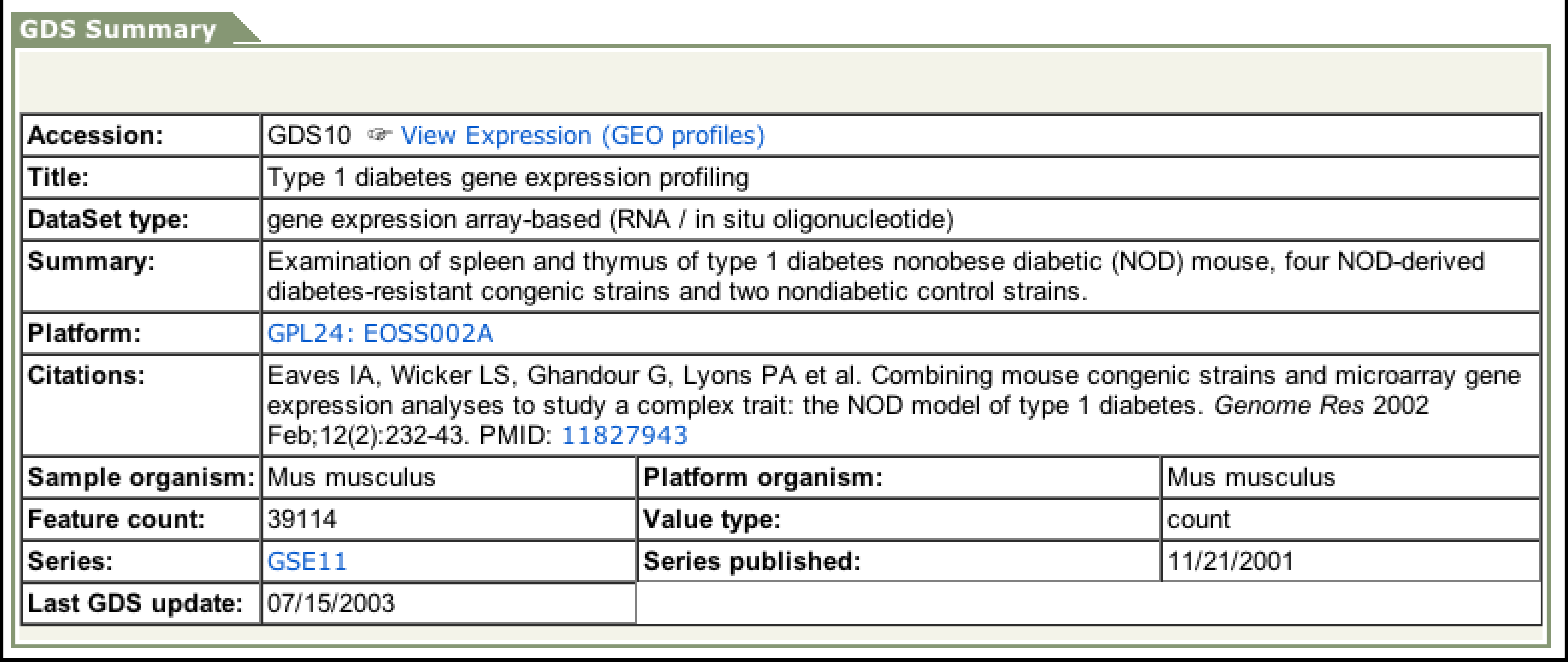
GDS Summary

Accession:	GDS10 ☞ View Expression (GEO profiles)		
Title:	Type 1 diabetes gene expression profiling		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Examination of spleen and thymus of type 1 diabetes nonobese diabetic (NOD) mouse, four NOD-derived diabetes-resistant congenic strains and two nondiabetic control strains.		
Platform:	GPL24: EOSS002A		
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		
Sample organism:	Mus musculus	**Platform organism:**	Mus musculus
Feature count:	39114	**Value type:**	count
Series:	GSE11	**Series published:**	11/21/2001
Last GDS update:	07/15/2003		



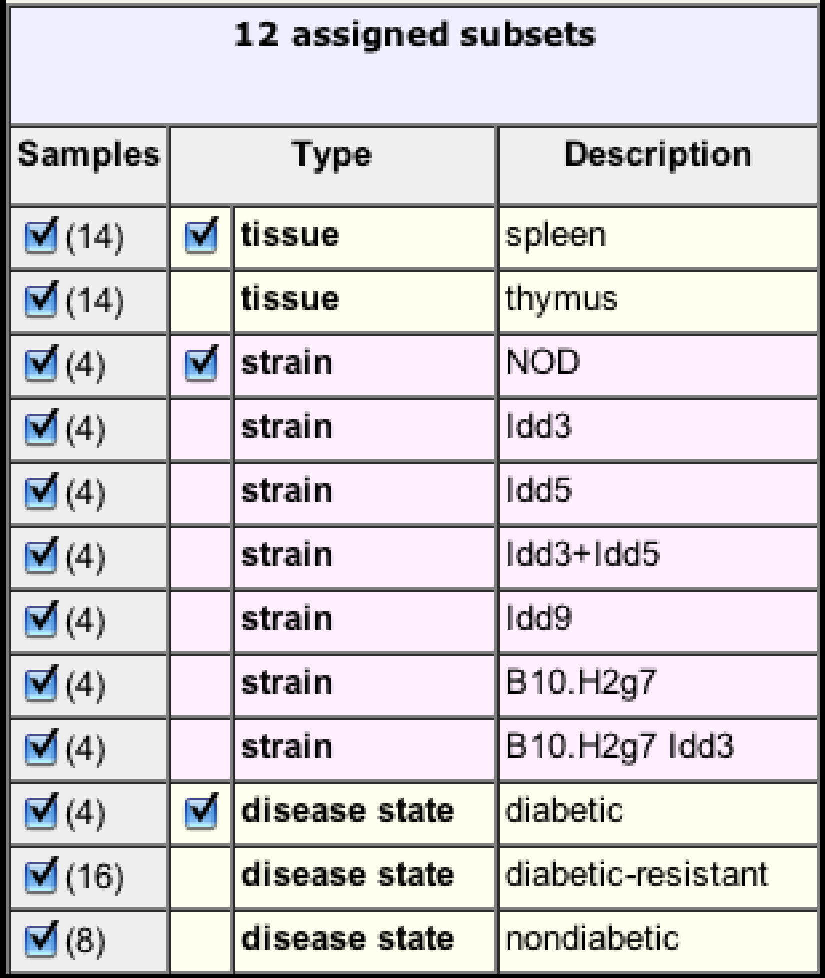
```
MH  - Animals
MH  - Diabetes Mellitus, Type 1/*genetics
MH  - *Disease Models, Animal
MH  - Female
MH  - *Gene Expression Profiling/methods
MH  - Genetic Markers/genetics
MH  - Mice
MH  - Mice, Congenic
MH  - Mice, Inbred C57BL
MH  - Mice, Inbred NOD/*genetics
MH  - Oligonucleotide Array Sequence Analysis/*methods
MH  - Polymorphism, Genetic/genetics
MH  - Research Design
```

Butte AJ, Chen R "Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics." AMIA Annu Symp Proc 2006; 106-10



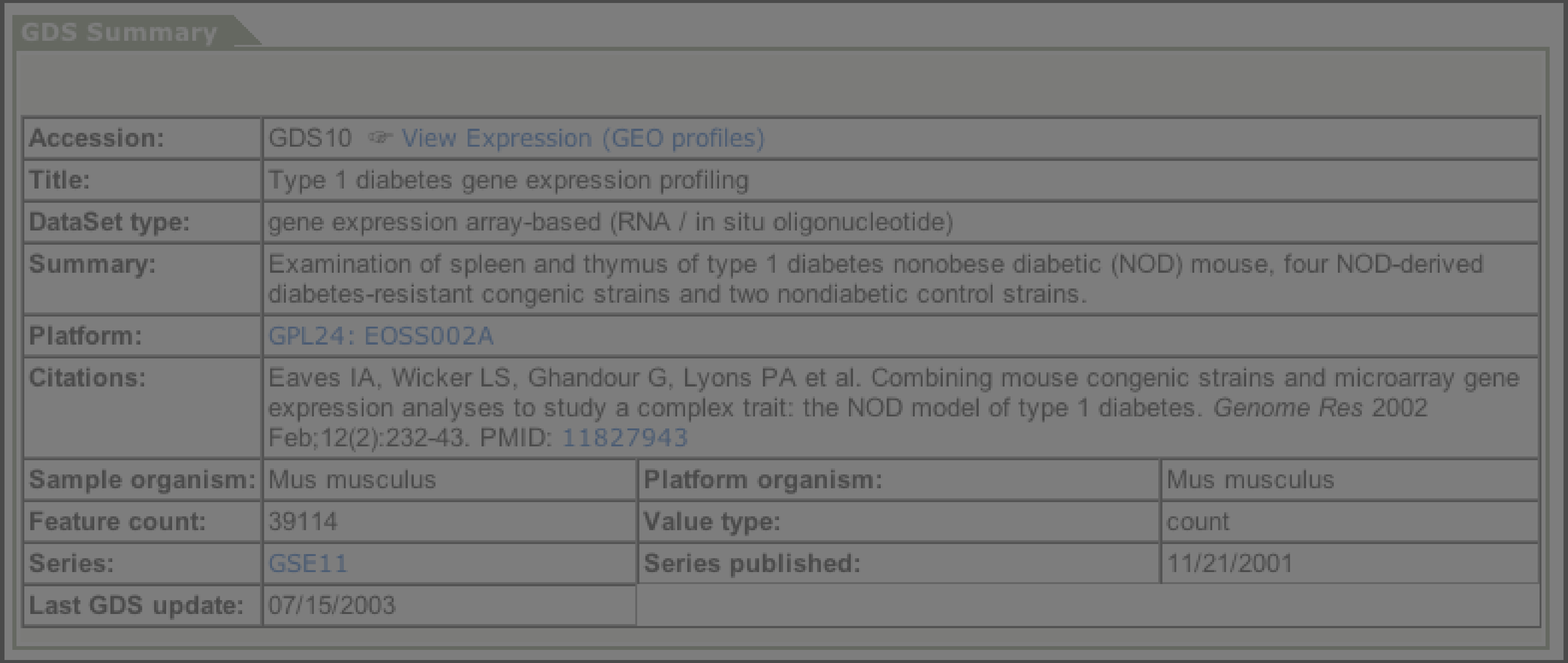
GDS Summary

Accession:	GDS10 ☞ View Expression (GEO profiles)		
Title:	Type 1 diabetes gene expression profiling		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Examination of spleen and thymus of type 1 diabetes nonobese diabetic (NOD) mouse, four NOD-derived diabetes-resistant congenic strains and two nondiabetic control strains.		
Platform:	GPL24: EOSS002A		
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		
Sample organism:	Mus musculus	**Platform organism:**	Mus musculus
Feature count:	39114	**Value type:**	count
Series:	GSE11	**Series published:**	11/21/2001
Last GDS update:	07/15/2003		



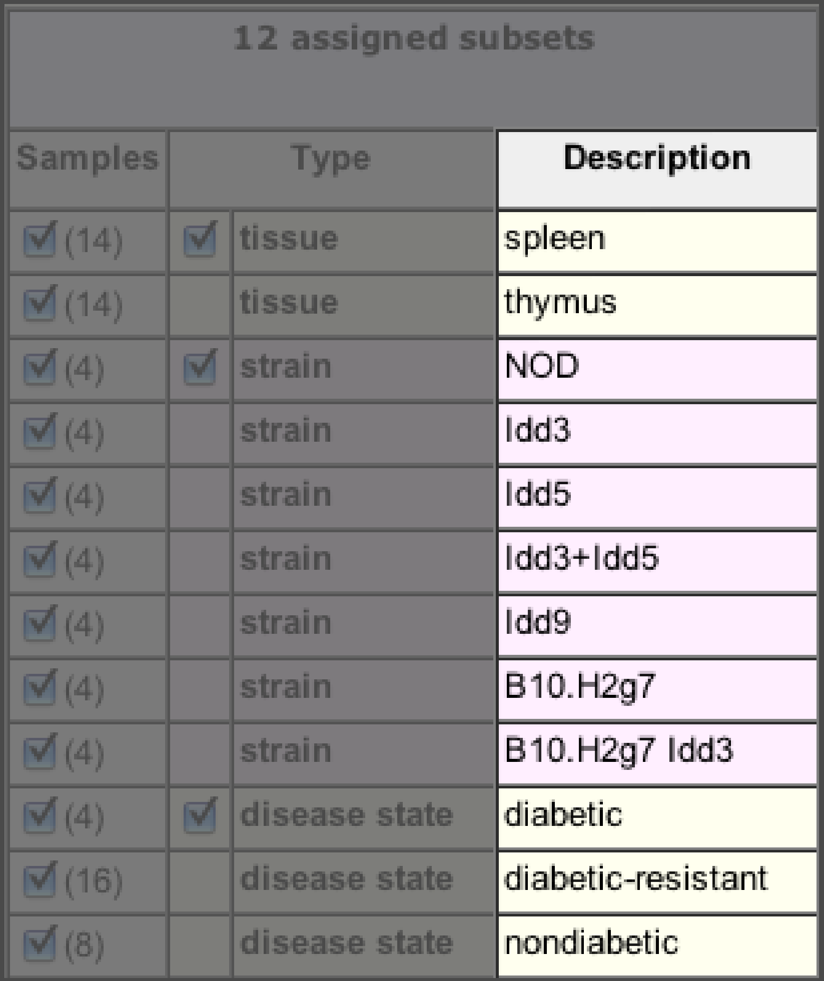
12 assigned subsets

Samples	Type		Description
☑ (14)	☑	**tissue**	spleen
☑ (14)		**tissue**	thymus
☑ (4)	☑	**strain**	NOD
☑ (4)		**strain**	Idd3
☑ (4)		**strain**	Idd5
☑ (4)		**strain**	Idd3+Idd5
☑ (4)		**strain**	Idd9
☑ (4)		**strain**	B10.H2g7
☑ (4)		**strain**	B10.H2g7 Idd3
☑ (4)	☑	**disease state**	diabetic
☑ (16)		**disease state**	diabetic-resistant
☑ (8)		**disease state**	nondiabetic



GDS Summary

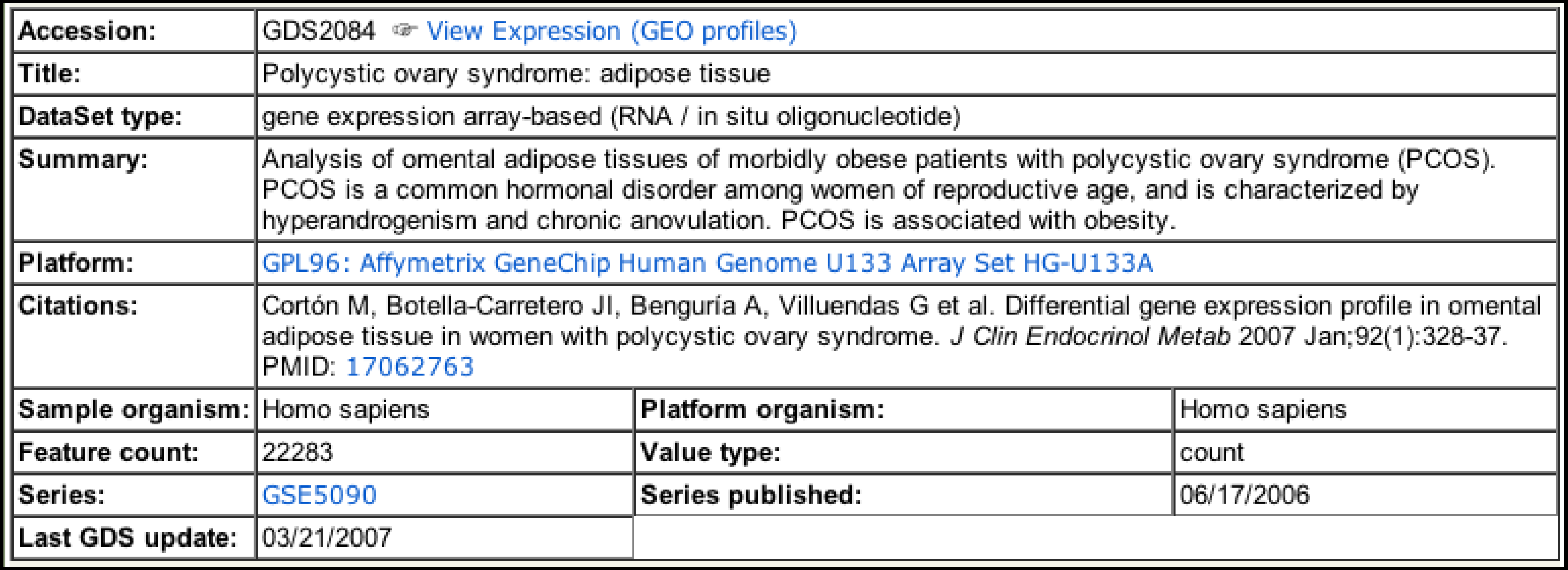
Accession:	GDS10 ☞ View Expression (GEO profiles)		
Title:	Type 1 diabetes gene expression profiling		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Examination of spleen and thymus of type 1 diabetes nonobese diabetic (NOD) mouse, four NOD-derived diabetes-resistant congenic strains and two nondiabetic control strains.		
Platform:	GPL24: EOSS002A		
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		
Sample organism:	Mus musculus	Platform organism:	Mus musculus
Feature count:	39114	Value type:	count
Series:	GSE11	Series published:	11/21/2001
Last GDS update:	07/15/2003		



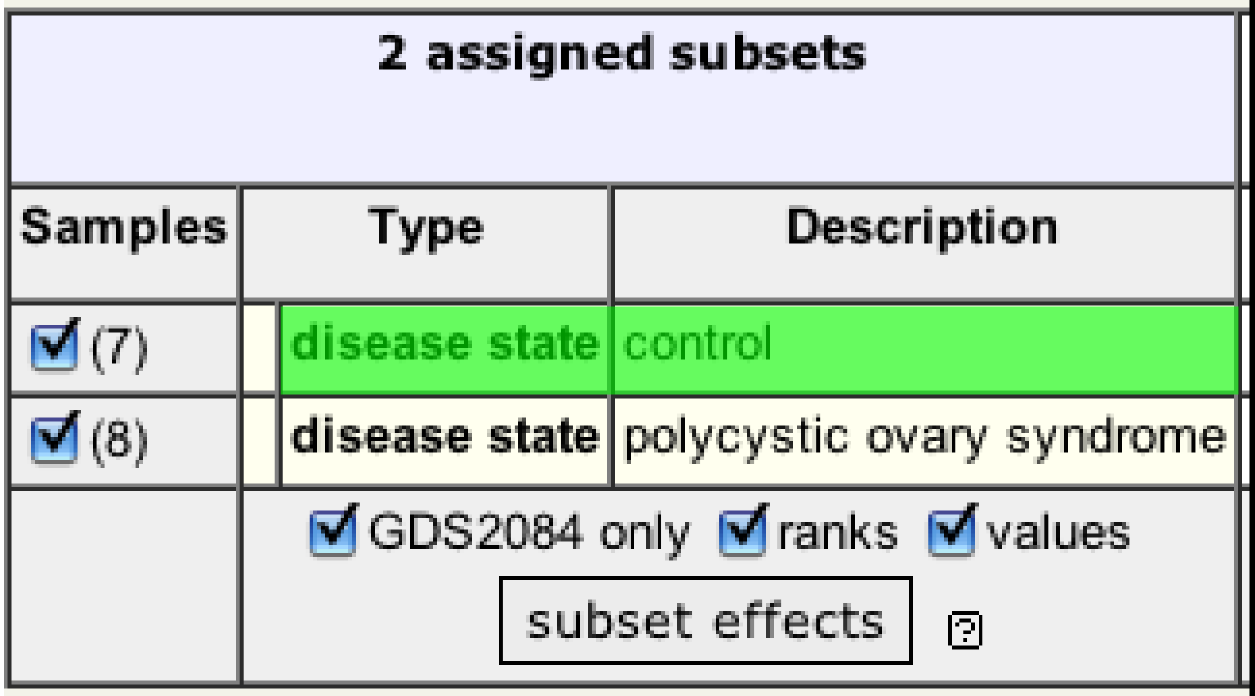
12 assigned subsets

Samples		Type	Description
☑ (14)	☑	tissue	spleen
☑ (14)		tissue	thymus
☑ (4)	☑	strain	NOD
☑ (4)		strain	Idd3
☑ (4)		strain	Idd5
☑ (4)		strain	Idd3+Idd5
☑ (4)		strain	Idd9
☑ (4)		strain	B10.H2g7
☑ (4)		strain	B10.H2g7 Idd3
☑ (4)	☑	disease state	diabetic
☑ (16)		disease state	diabetic-resistant
☑ (8)		disease state	nondiabetic

Free Text!



Accession:	GDS2084 ☞ View Expression (GEO profiles)		
Title:	Polycystic ovary syndrome: adipose tissue		
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)		
Summary:	Analysis of omental adipose tissues of morbidly obese patients with polycystic ovary syndrome (PCOS). PCOS is a common hormonal disorder among women of reproductive age, and is characterized by hyperandrogenism and chronic anovulation. PCOS is associated with obesity.		
Platform:	GPL96: Affymetrix GeneChip Human Genome U133 Array Set HG-U133A		
Citations:	Cortón M, Botella-Carretero JI, Benguría A, Villuendas G et al. Differential gene expression profile in omental adipose tissue in women with polycystic ovary syndrome. *J Clin Endocrinol Metab* 2007 Jan;92(1):328-37. PMID: 17062763		
Sample organism:	Homo sapiens	**Platform organism:**	Homo sapiens
Feature count:	22283	**Value type:**	count
Series:	GSE5090	**Series published:**	06/17/2006
Last GDS update:	03/21/2007		



2 assigned subsets

Samples	Type	Description
☑ (7)	disease state	control
☑ (8)	disease state	polycystic ovary syndrome

☑ GDS2084 only ☑ ranks ☑ values

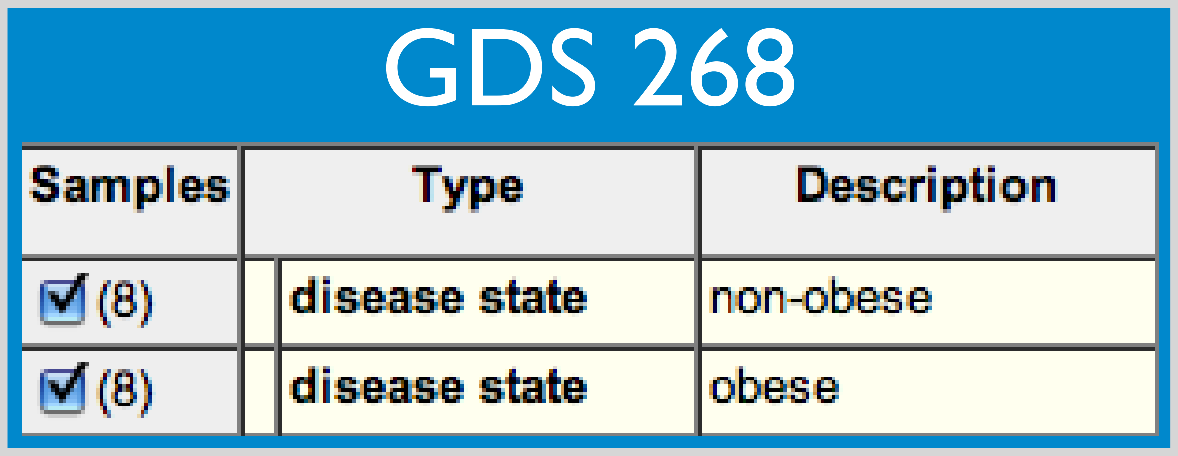
subset effects ?

→ C0032460

↓

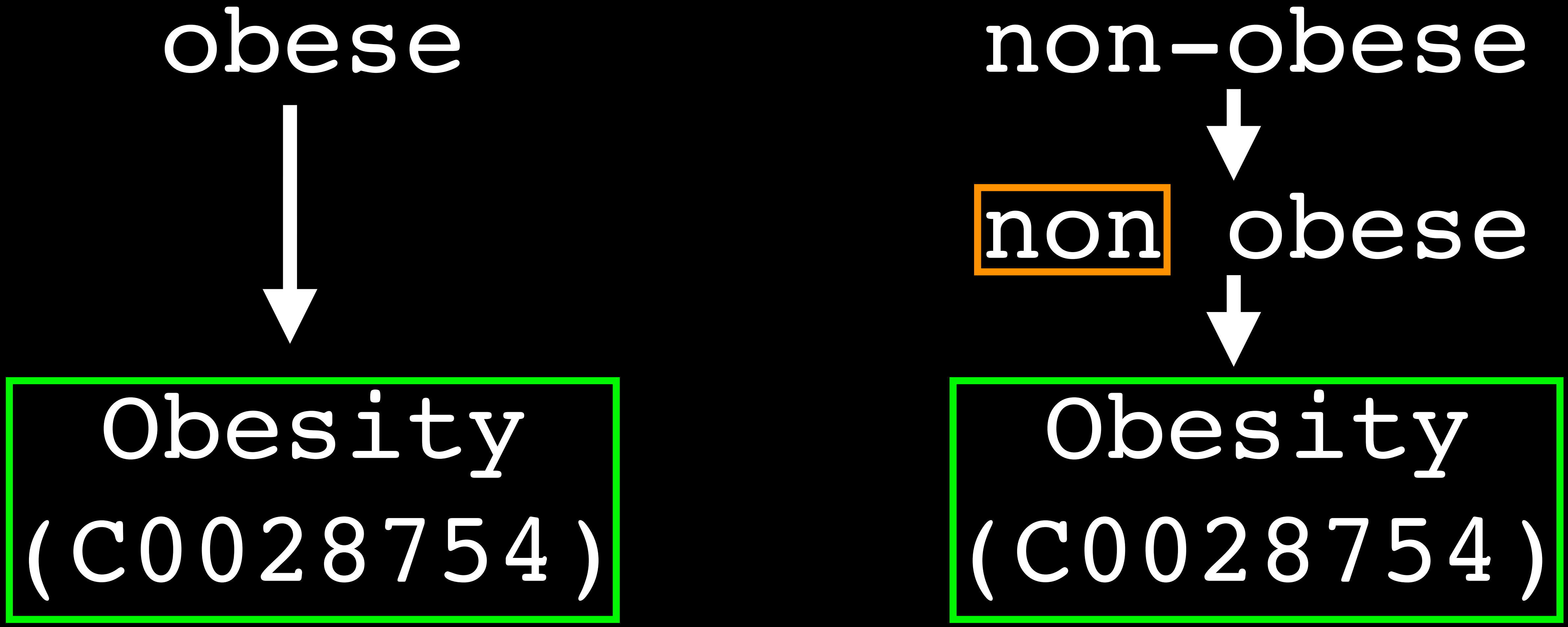
Disease or Syndrome (T047)

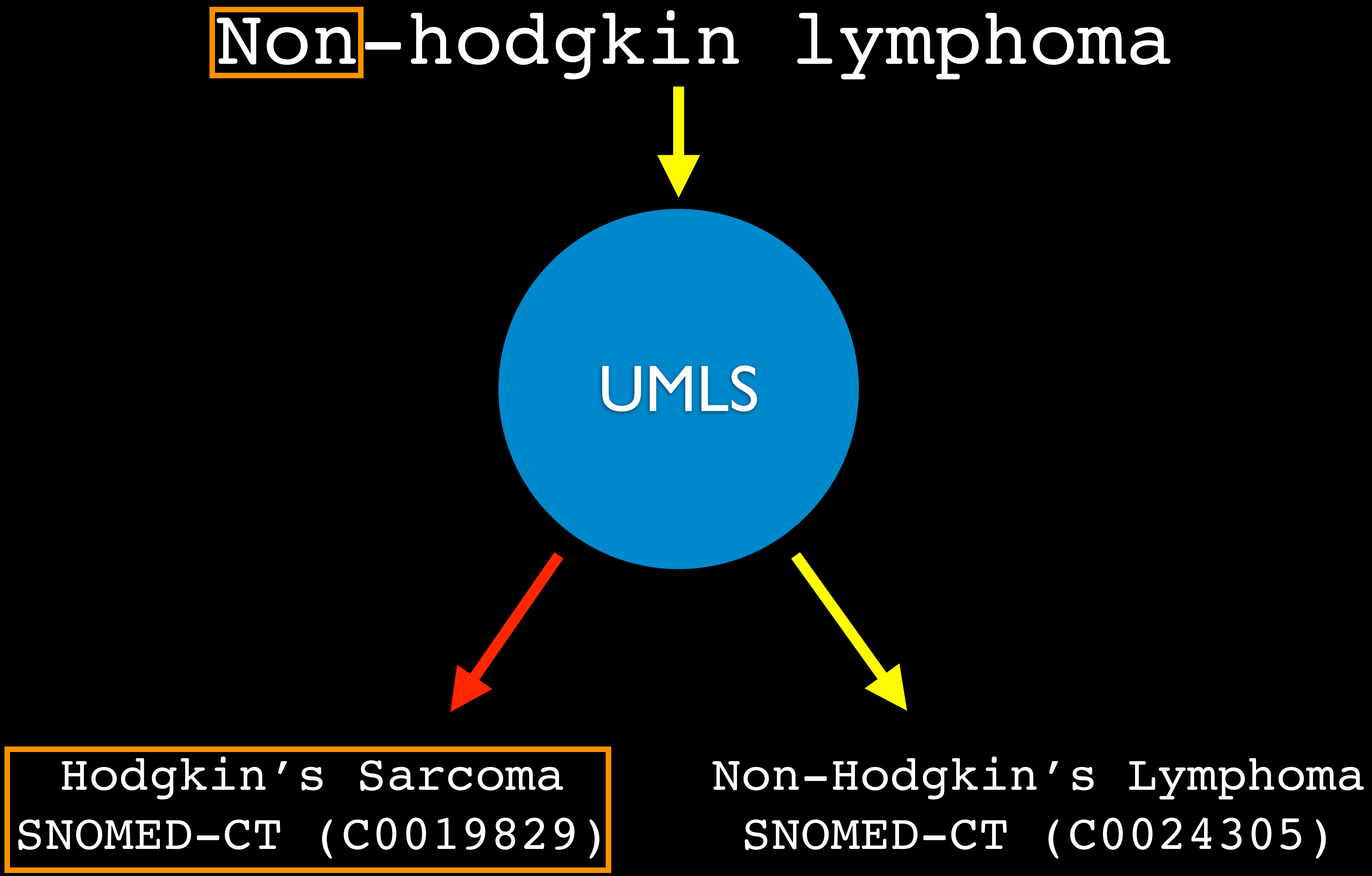
Method B: Identifying Control Subsets Defined Using a Negation Scheme



GDS 268

Samples	Type	Description
☑ (8)	disease state	non-obese
☑ (8)	disease state	obese





Non-hodgkin lymphoma

UMLS

Hodgkin's Sarcoma
SNOMED-CT (C0019829)

Non-Hodgkin's Lymphoma
SNOMED-CT (C0024305)

Dudley J and Butte AJ. Enabling integrative genomic analysis of high-impact human diseases through text mining. Pacific Symposium on Biocomputing (2008) pp. 580-91

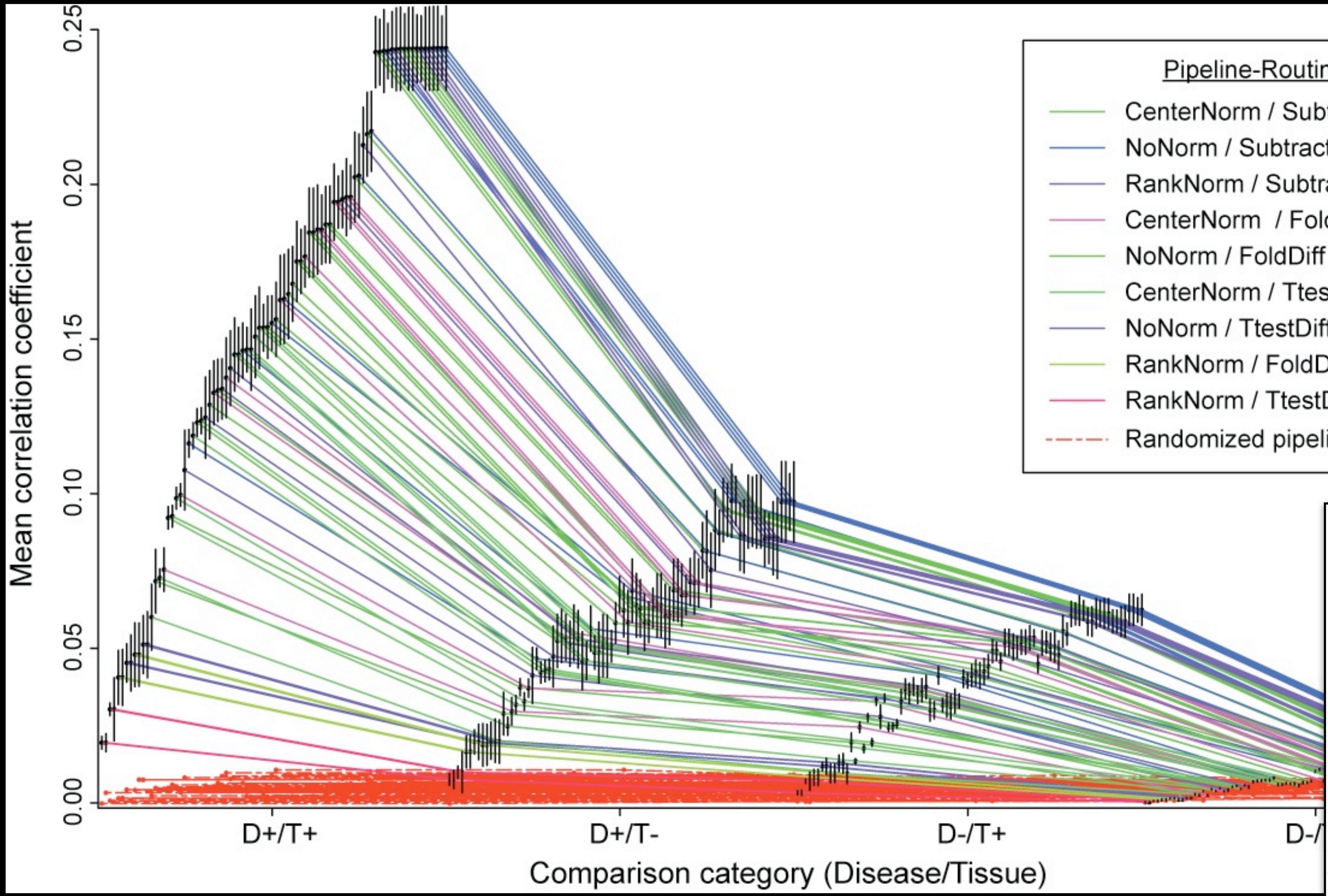
AILUN: Extracting GEO gene lists

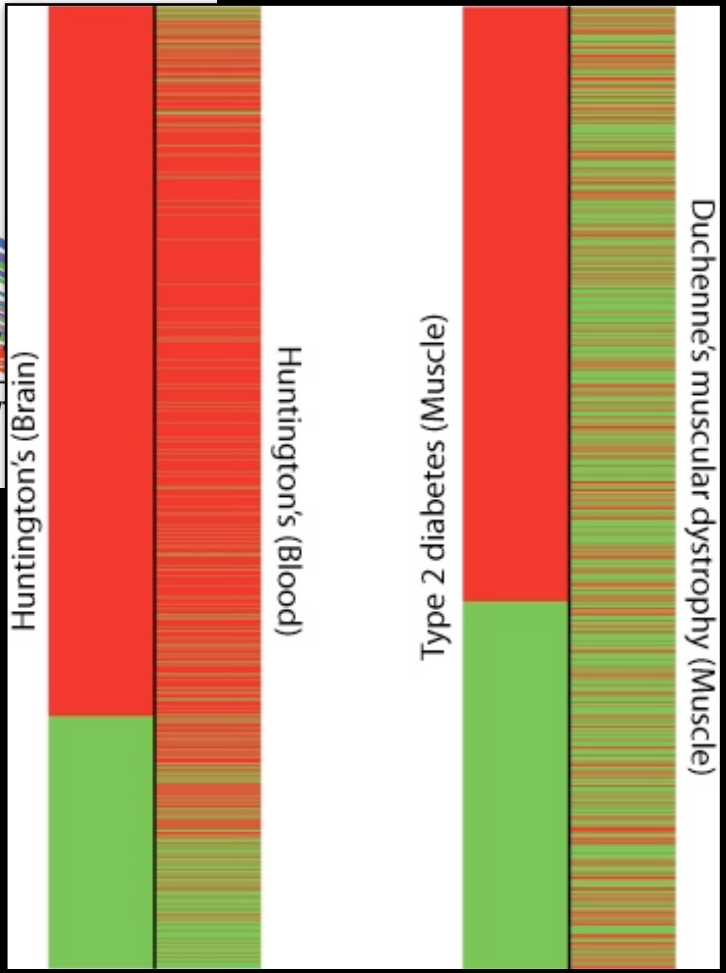
- GEO has 12.6+ billion measurements across ~4000 platforms
- Decoding measured gene is a challenge
  - Varied use of identifiers
  - Identifiers change meaning
- We have ~100 million mappings to NCBI Gene ids
- We mapped 67% of GEO platforms to NCBI identifiers

Chen R, Butte AJ. *Nature Methods*, November 2007.

Gene Identifier	Gene Identifier Vocabulary
AI262683	GenBank
NM_000015	GenBank
Hs.2	UniGene
NP_000006	Protein
P11245	Protein
NAT2	NCBI Gene official symbols
AAC2	NCBI Gene all symbols
IMAGE:1870937	IMAGE clone
UI-H-FG1-bgl-g-02-0-UI	University of Iowa clone
IMAGp998I184581_	Institute of Molecular Biology and Genetics Ukraine clone
10286060	GenBank GI
TC110817	OriGene Technologies Clone
HIE06837r	Gunma University Clone
CMPD10049	University of Padova Clone
3NHC3746	Institute of Medical Science Japan Clone

Human N-acetyltransferase 2

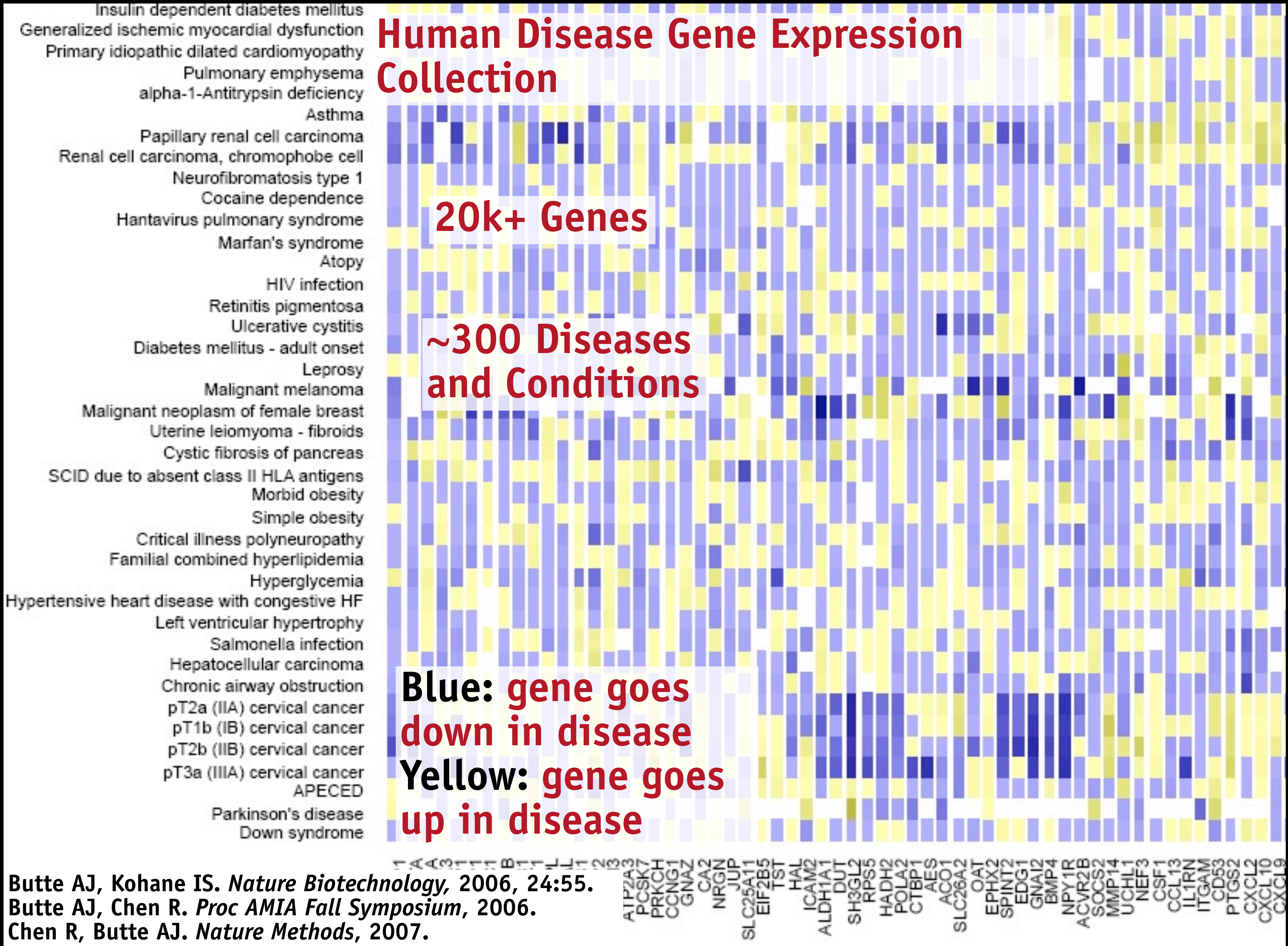




What is the quality of the public gene expression data?

Dudley et al. Disease signatures are robust across tissues and experiments. *Molecular systems biology* (2009) vol. 5 pp. 307

Human Disease Gene Expression Collection



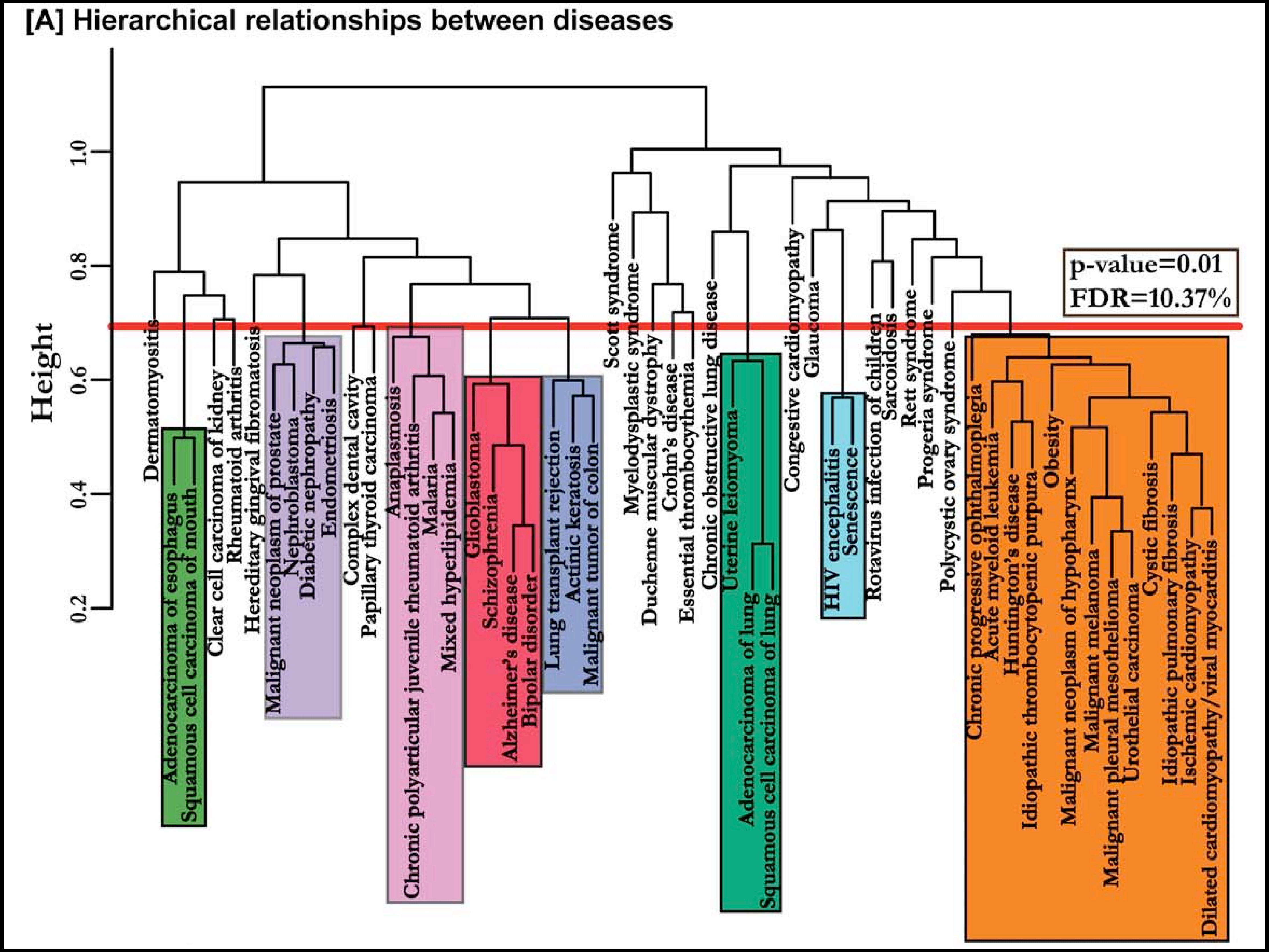
20k+ Genes

~300 Diseases and Conditions

Blue: gene goes down in disease

Yellow: gene goes up in disease

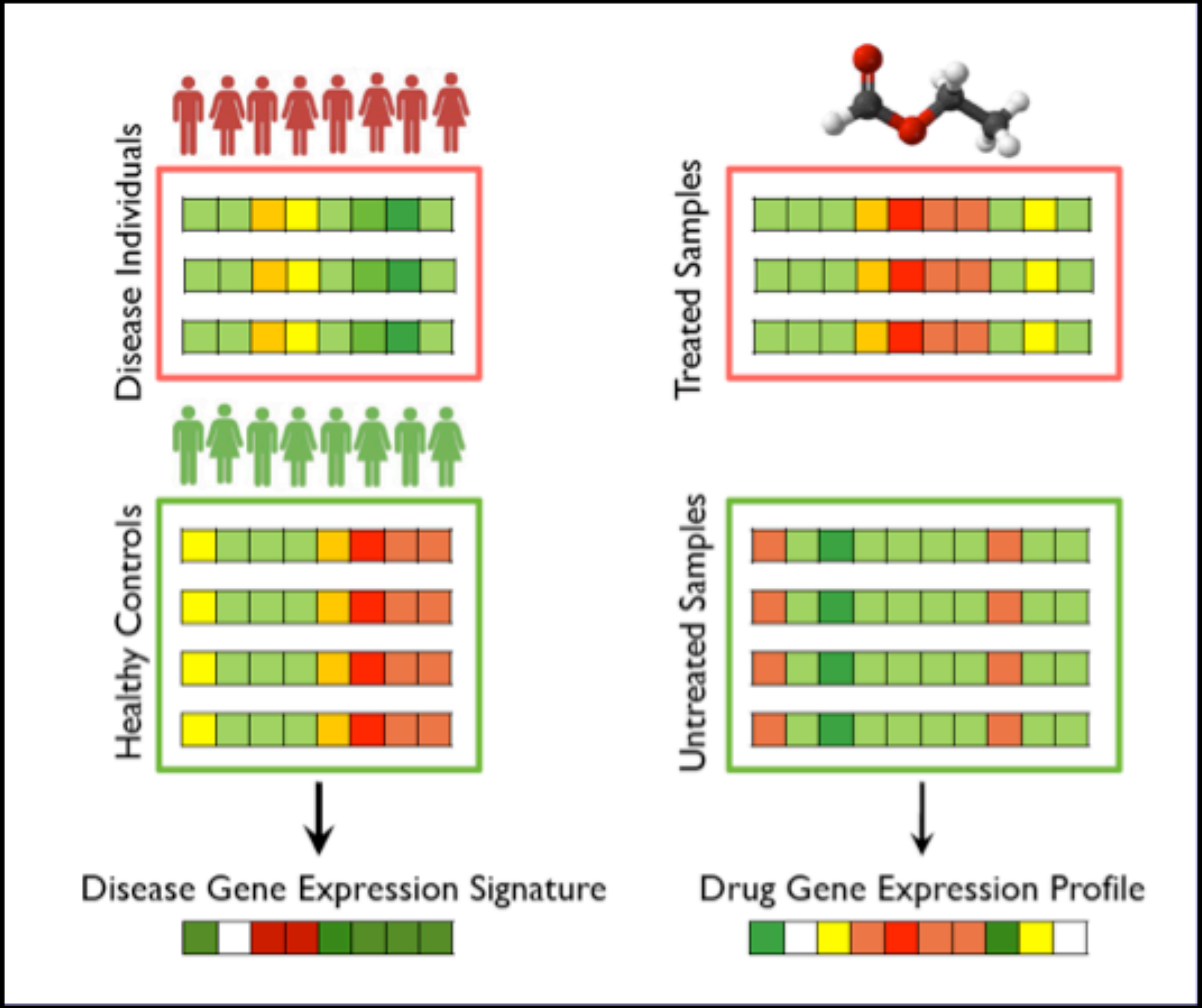
Butte AJ, Kohane IS. *Nature Biotechnology,* 2006, 24:55.
Butte AJ, Chen R. *Proc AMIA Fall Symposium*, 2006.
Chen R, Butte AJ. *Nature Methods*, 2007.



[A] Hierarchical relationships between diseases

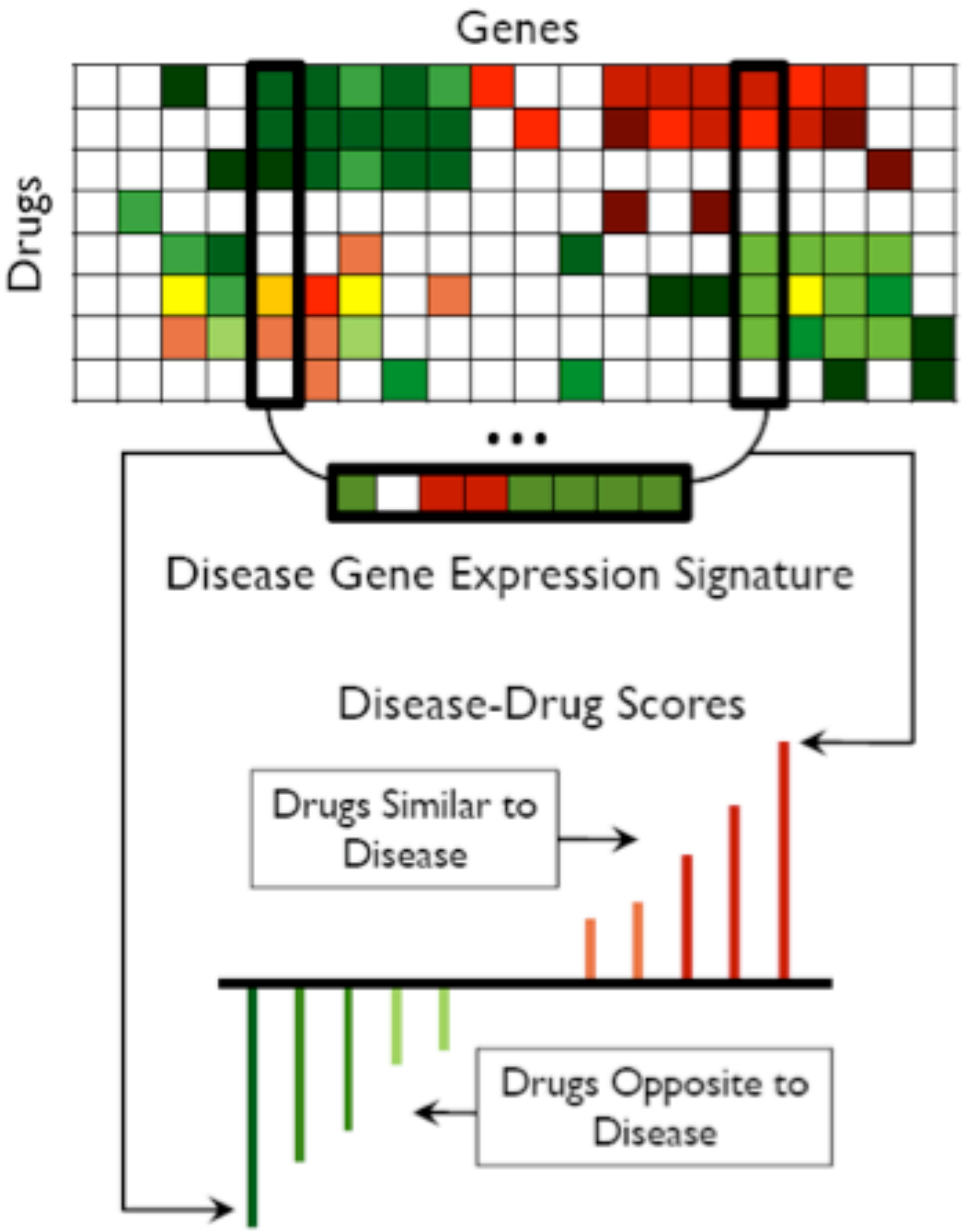
Suthram S, Dudley J et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology* (2010) vol. 6 (2)

Mining Public Data for Drug Repositioning



Dudley JT, Sirota M et al. Discovery and validation of drug indications using compendia of public gene expression data (in revision)

Reference Database of Drug Gene Expression



Genes

Drugs

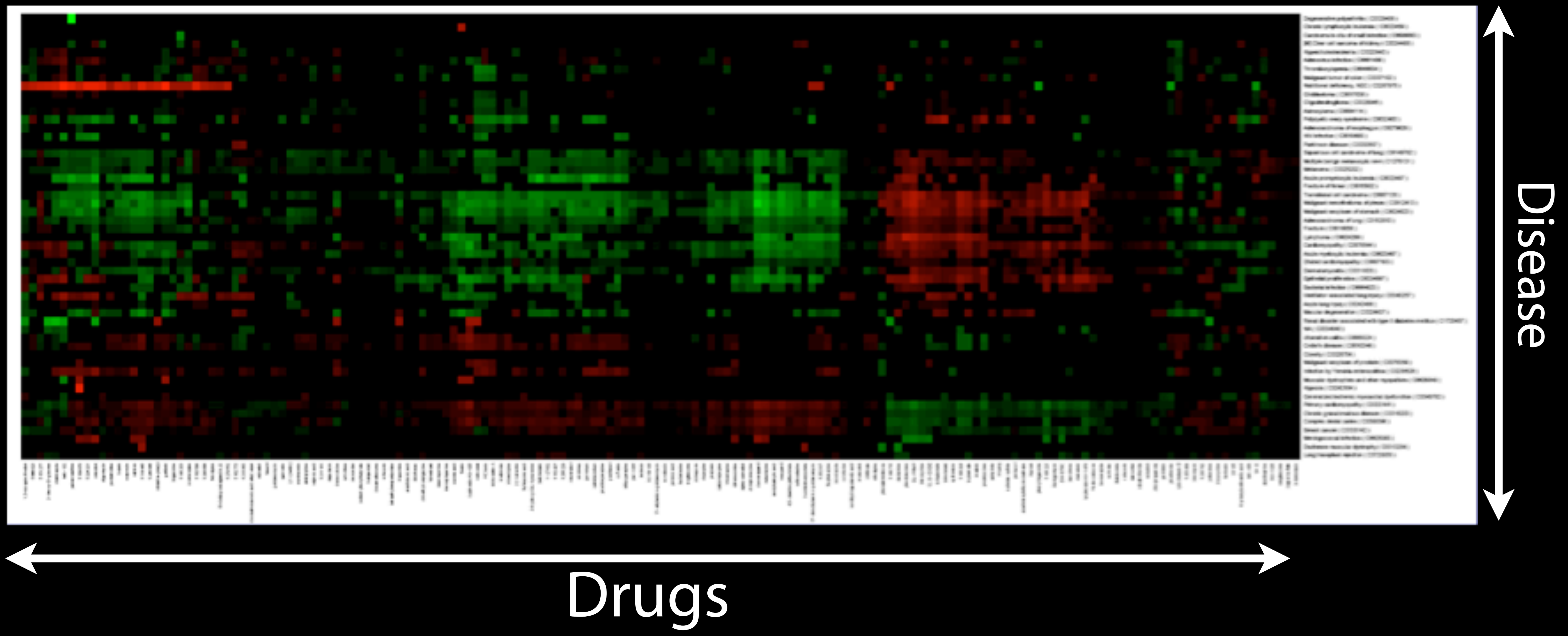
• • •

Disease Gene Expression Signature

Disease-Drug Scores

Drugs Similar to Disease

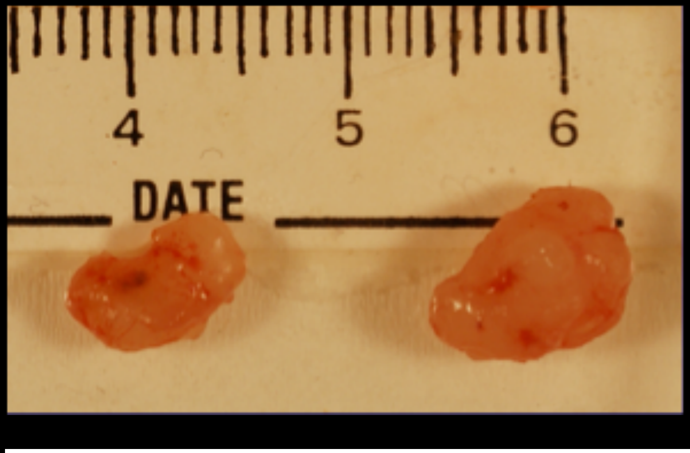
Drugs Opposite to Disease

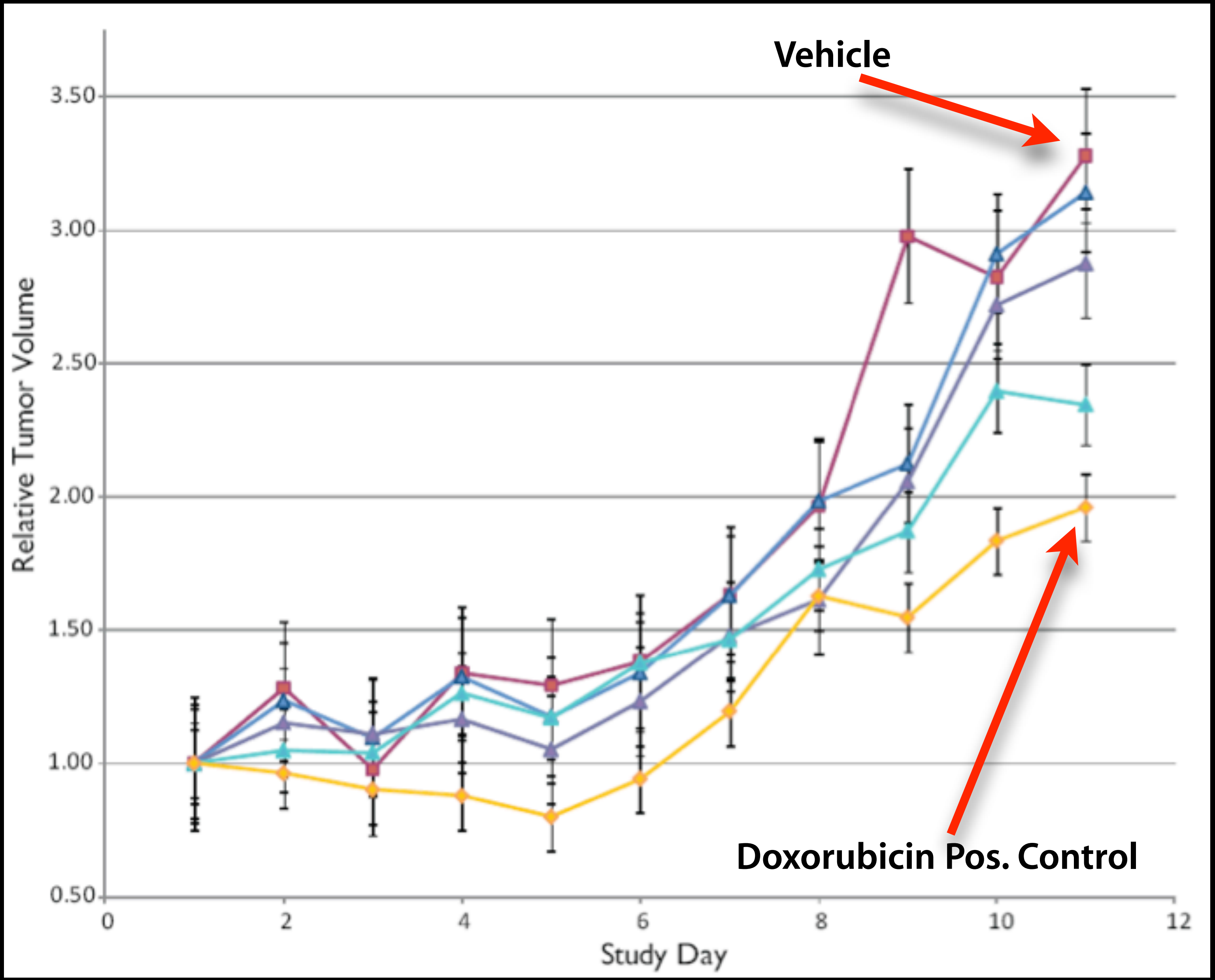


Disease

Drugs

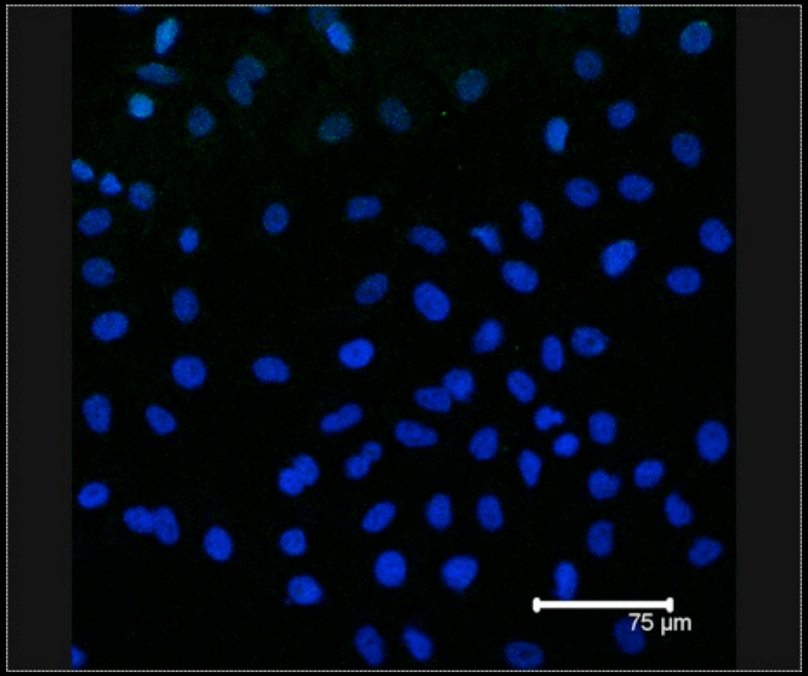
Anti-ulcer drug inhibits lung adenocarcinoma *in vitro* and *in vivo*



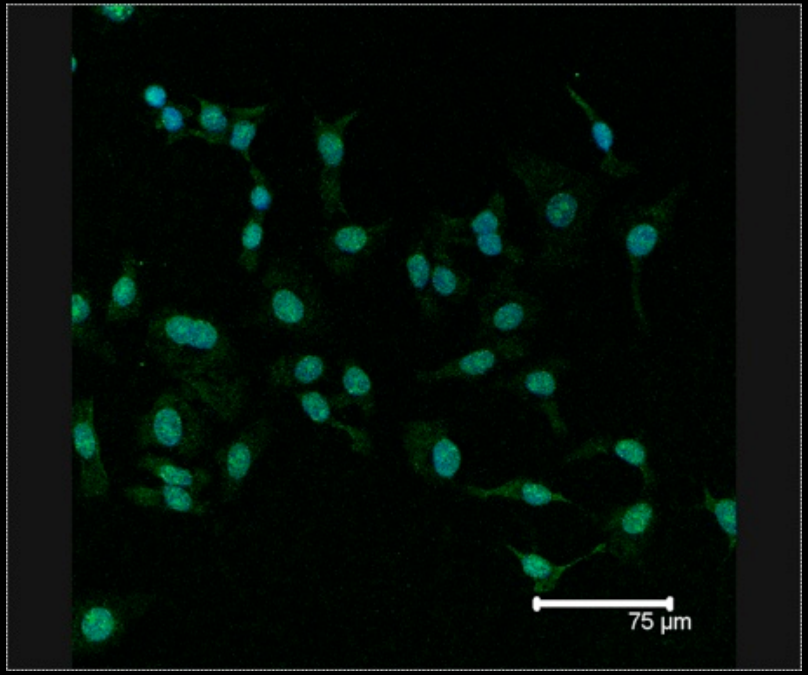


Vehicle

Doxorubicin Pos. Control

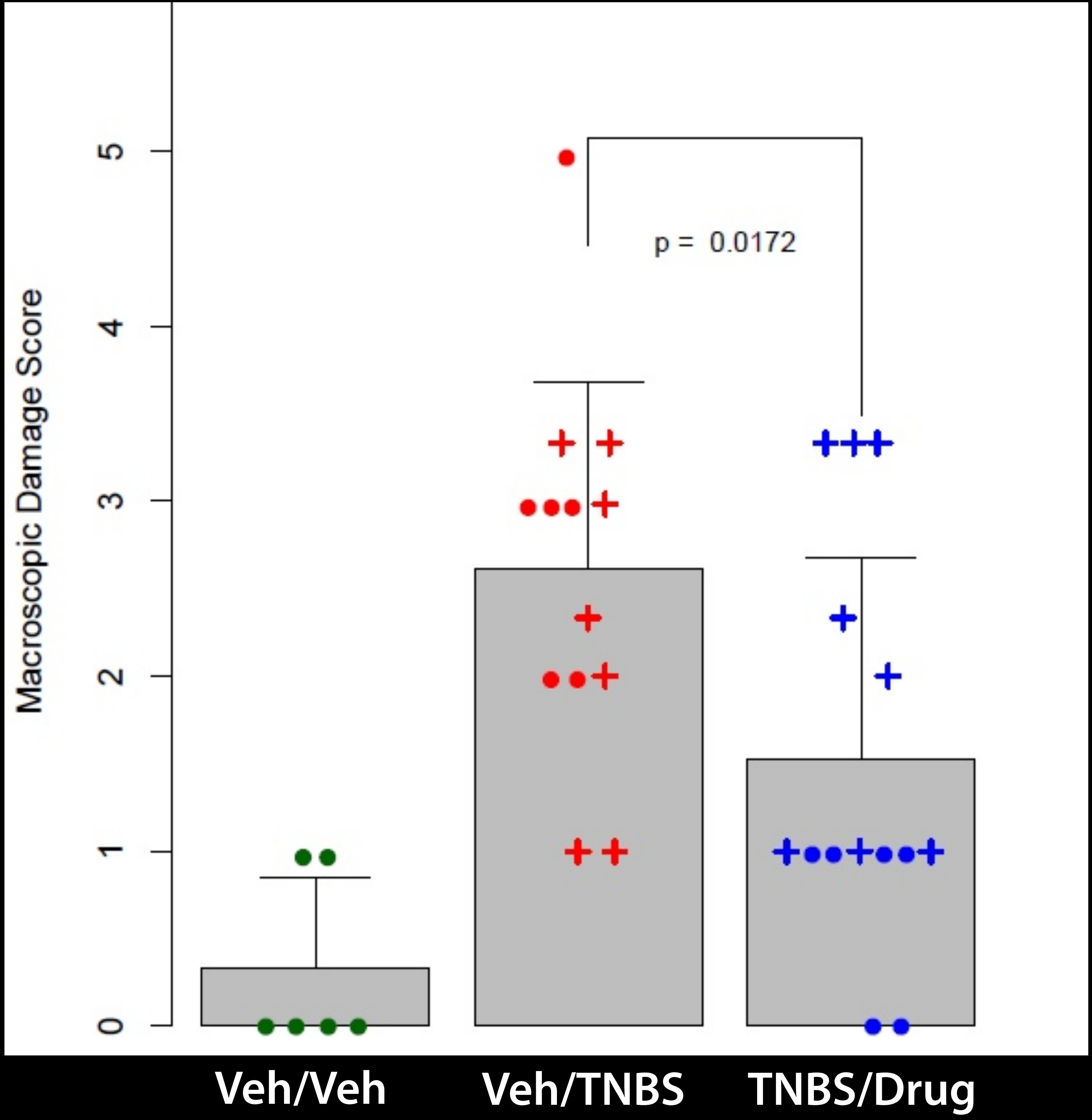


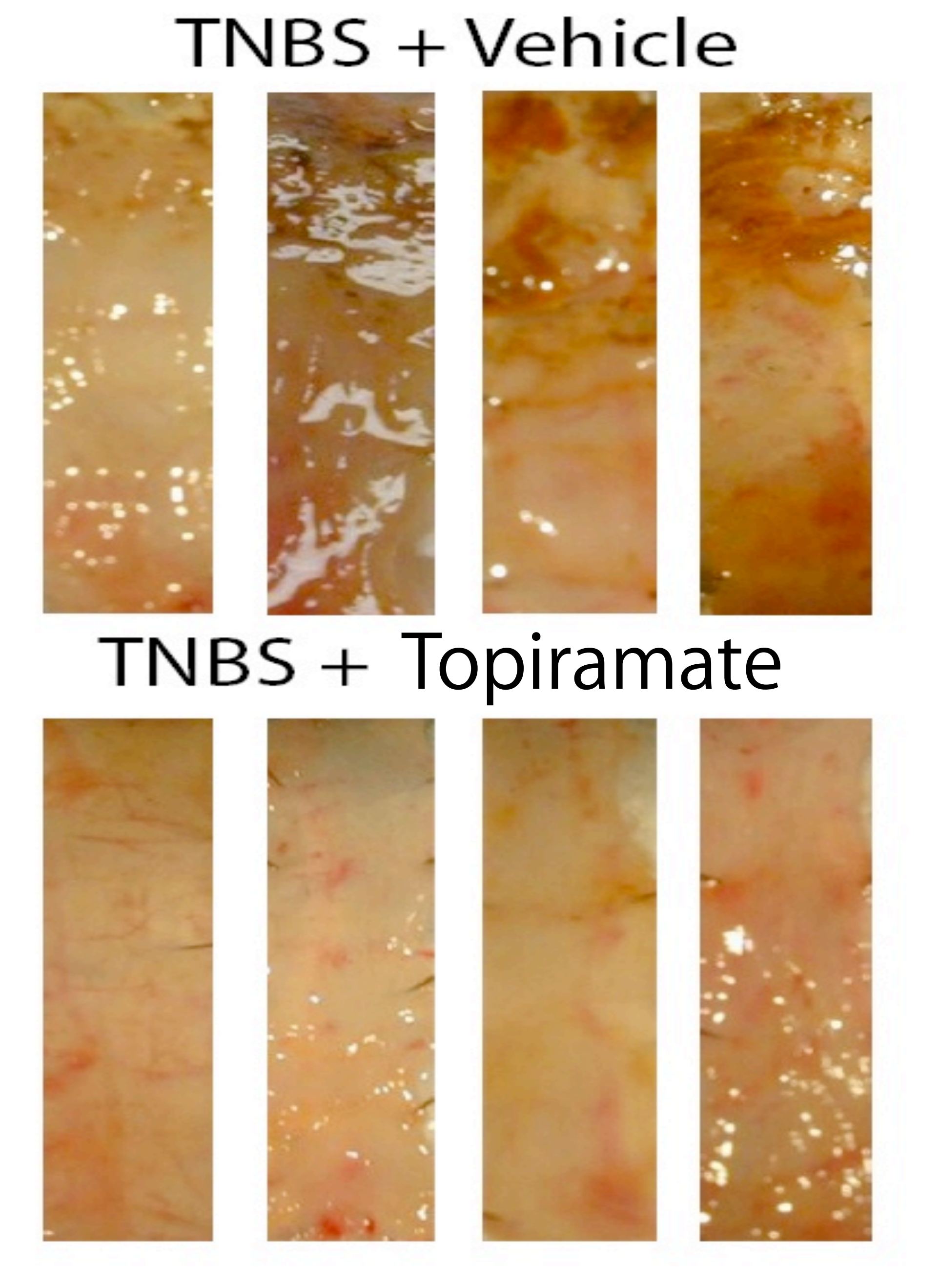
Vehicle



Treated

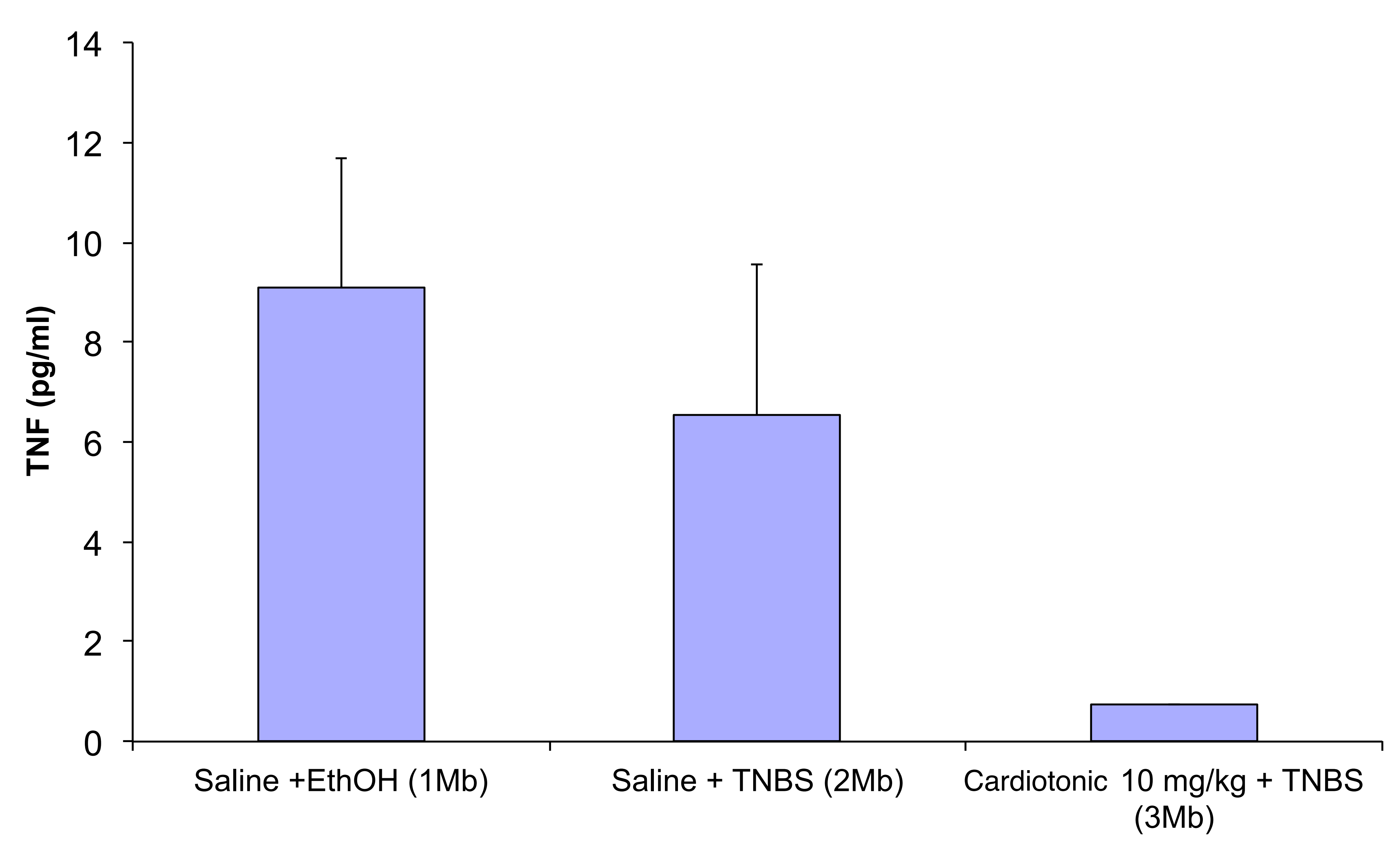
Anti-seizure drug works against a rat model of inflammatory bowel disease



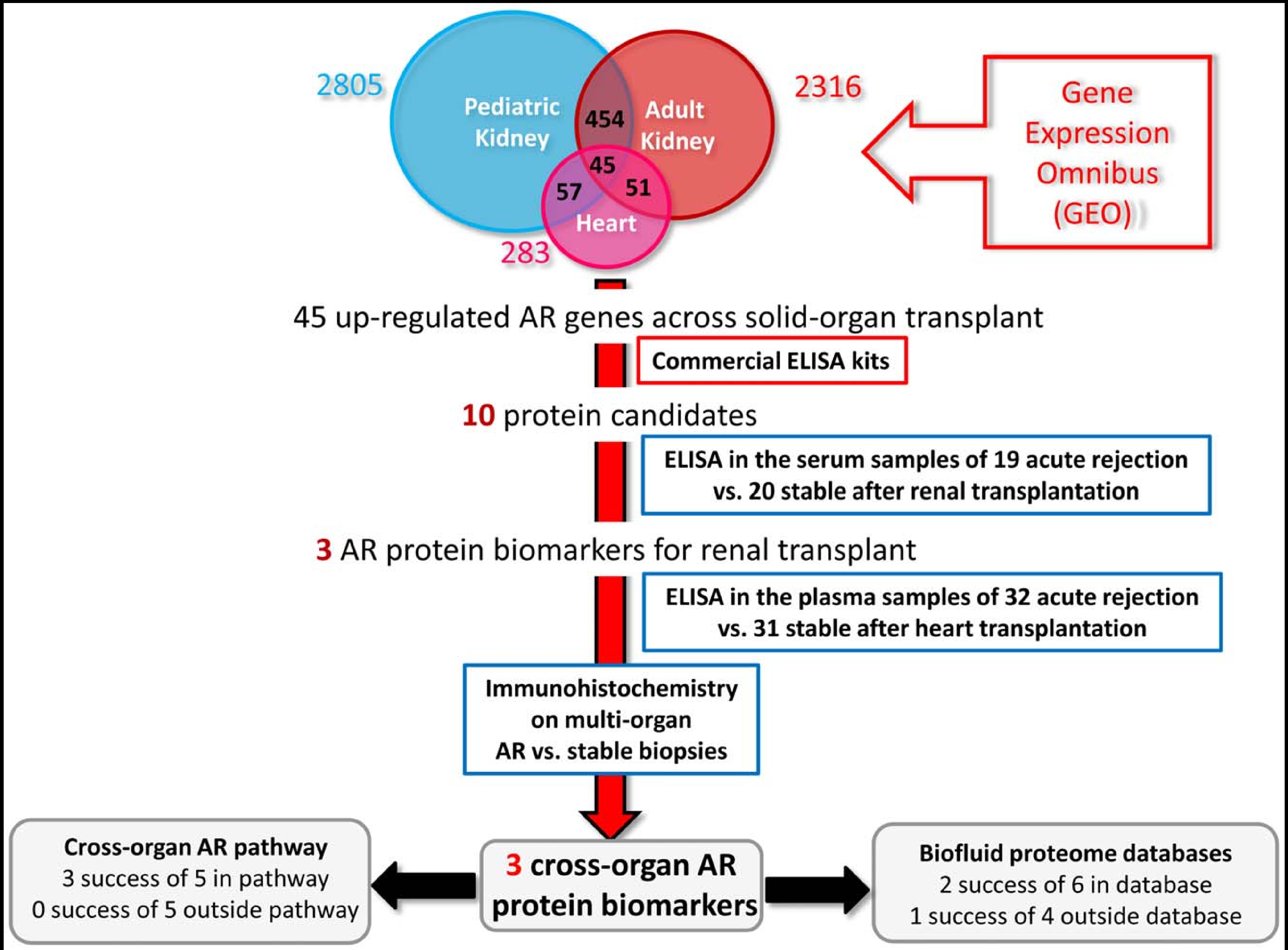


Dudley JT*, Sirota M* et al. (in revision)

Cardiotonic drug inhibits ameliorates inflammatory cytokine TNF-alpha



Discovery of peripheral biomarkers for transplant rejection through integration of public data



45 up-regulated AR genes across solid-organ transplant

Commercial ELISA kits

10 protein candidates

ELISA in the serum samples of 19 acute rejection vs. 20 stable after renal transplantation

3 AR protein biomarkers for renal transplant

ELISA in the plasma samples of 32 acute rejection vs. 31 stable after heart transplantation

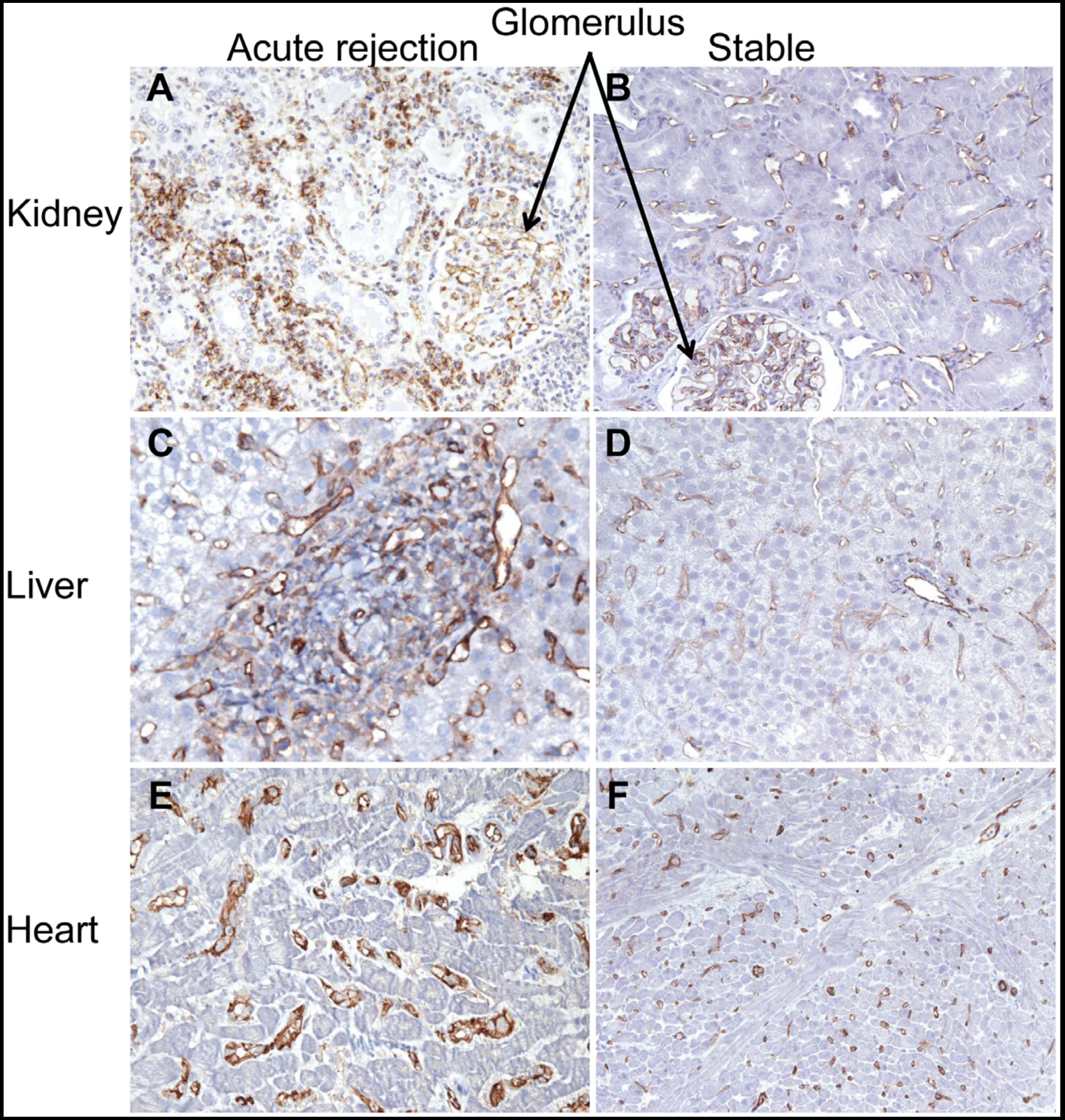
Immunohistochemistry on multi-organ AR vs. stable biopsies

Cross-organ AR pathway
3 success of 5 in pathway
0 success of 5 outside pathway

3 cross-organ AR protein biomarkers

Biofluid proteome databases
2 success of 6 in database
1 success of 4 outside database

Chen R et al. Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions.. *PLoS Computational Biology* (2010) e1000940


Glomerulus
Acute rejection
Stable
Kidney
Liver
Heart
A
B
C
D
E
F

Many more examples of new medicine from public data

- New large-effect genetic risk variant for Type 2 diabetes
- New drug target for Type 2 diabetes
- Biomarker for medulloblastoma
- Biomarker for pancreatic cancer
- Biomarker for lung cancer
- Biomarker for atherosclerosis

We can do this because we have the computational firepower, but what about others?

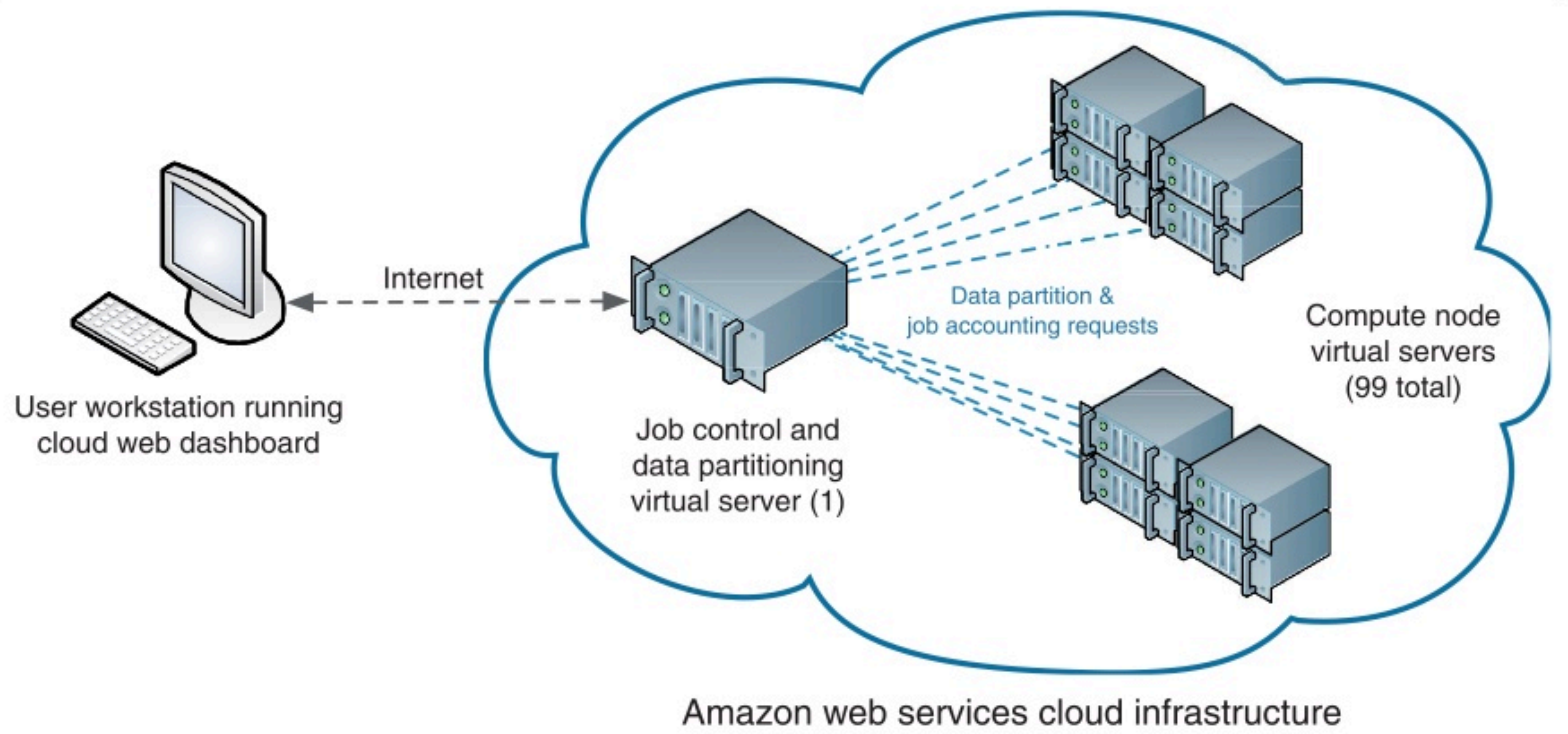

Internet
User workstation running cloud web dashboard
Job control and data partitioning virtual server (1)
Data partition & job accounting requests
Compute node virtual servers (99 total)
Amazon web services cloud infrastructure

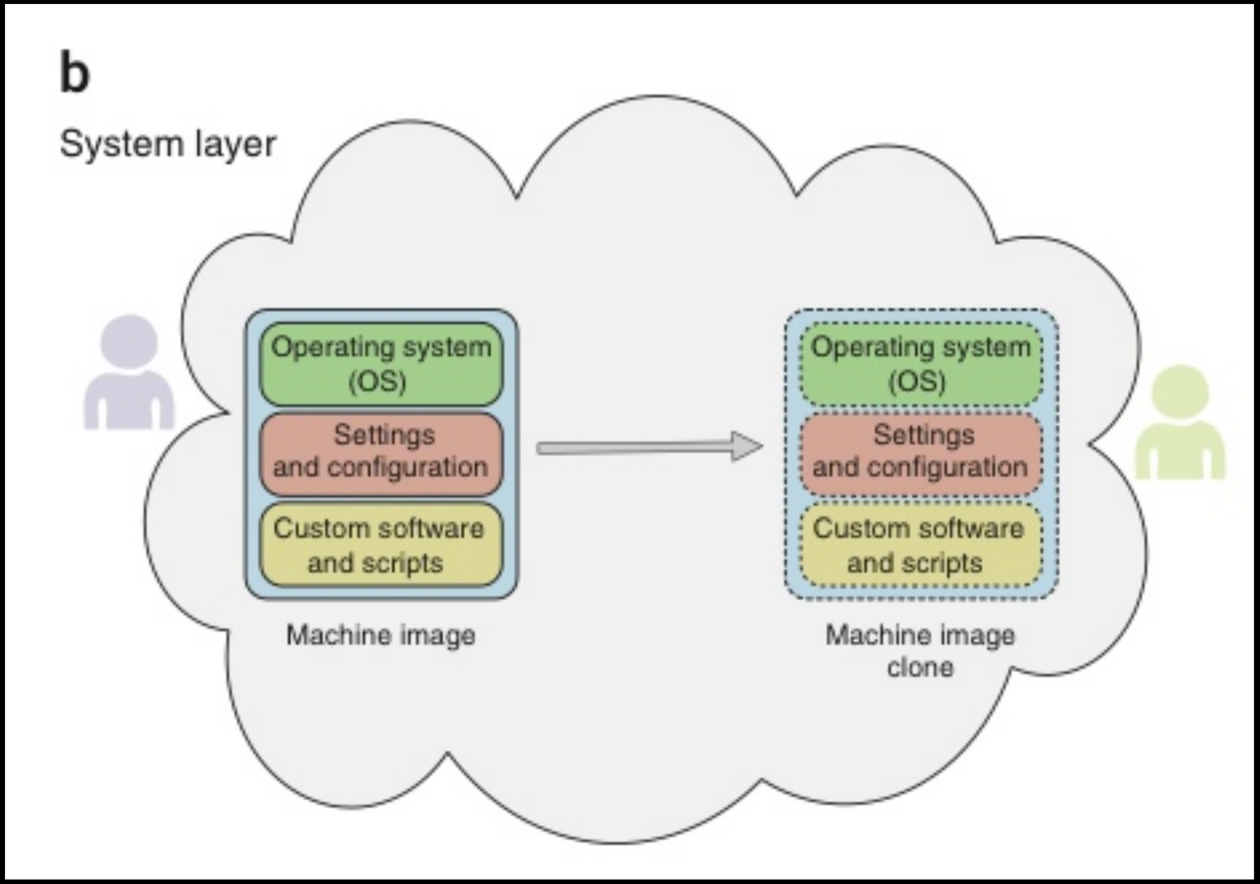
Figure 1 Schematic illustration of the computational strategy utilized for the cloud-based eQTL analysis. One hundred virtual server instances are provisioned using a web-based cloud control dashboard. One of the virtual server instances served as a data distribution and job control server. Upon initialization, the compute nodes would request a subset partition of eQTL comparisons and insert timestamp entries into a job accounting database upon initiation and completion of the eQTL analysis subset it was administered.

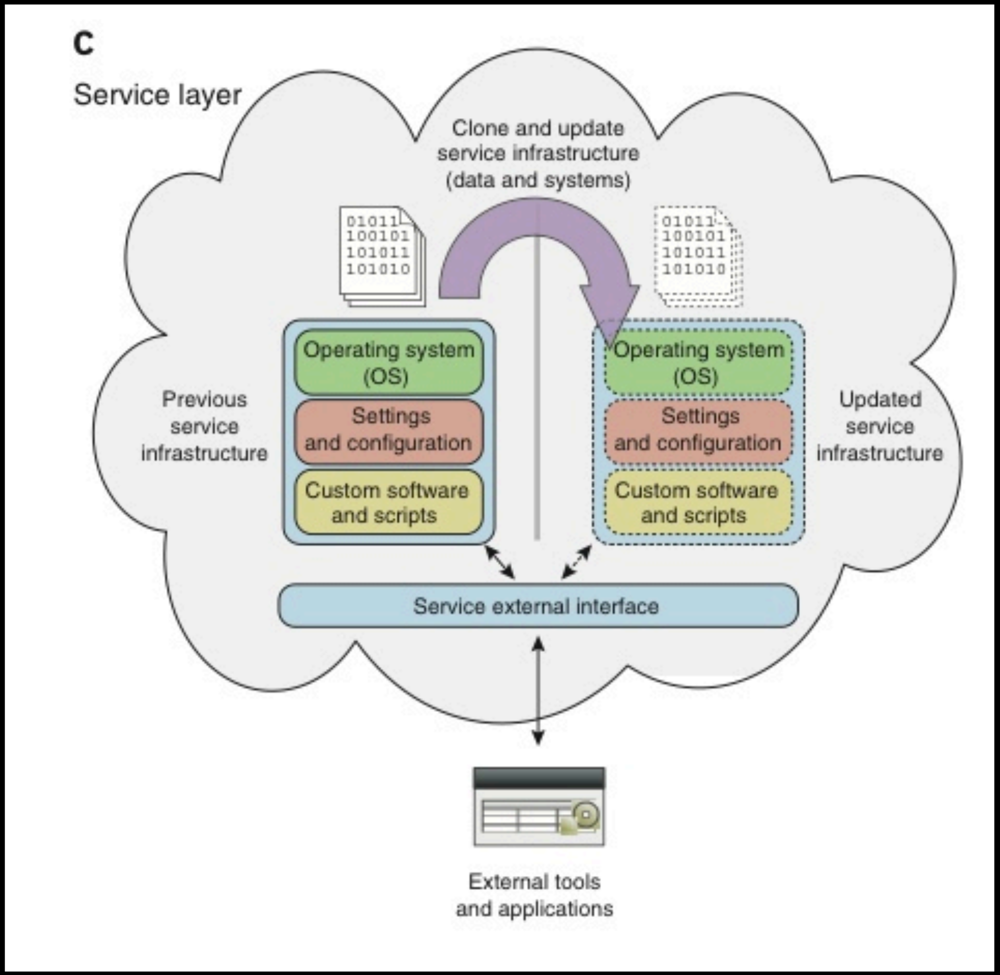
Dudley et al. Translational bioinformatics in the cloud: an affordable alternative. *Genome medicine* (2010) vol. 2 (8) pp. 51

In silico research in the era of cloud computing

Joel T Dudley & Atul J Butte

Snapshots of computer systems that are stored and shared 'in the cloud' could make computational analyses more reproducible.


b
System layer
Operating system (OS)
Settings and configuration
Custom software and scripts
Machine image
Operating system (OS)
Settings and configuration
Custom software and scripts
Machine image clone


c
Service layer
Clone and update service infrastructure (data and systems)
01011 100101 101011 101010
01011 100101 101011 101010
Previous service infrastructure
Operating system (OS)
Settings and configuration
Custom software and scripts
Operating system (OS)
Settings and configuration
Custom software and scripts
Updated service infrastructure
Service external interface
External tools and applications

Dudley and Butte. In silico research in the era of cloud computing. *Nature biotechnology* (2010) vol. 28 (11) pp. 1181-5

Lessons learned from integrating open biomedical data for translational research

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots
- Lightweight integration trumps ontology

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots
- Lightweight integration trumps ontology
- Computation is a major bottleneck
    - Right now there are privileged computational elite

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots
- Lightweight integration trumps ontology
- Computation is a major bottleneck
  - Right now there are privileged computational elite
- Questions first, data second

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots
- Lightweight integration trumps ontology
- Computation is a major bottleneck
  - Right now there are privileged computational elite
- Questions first, data second
- Data really is unreasonably effective

Lessons learned from integrating open biomedical data for translational research

- So far sticks have worked better than carrots
- Lightweight integration trumps ontology
- Computation is a major bottleneck
    - Right now there are privileged computational elite
- Questions first, data second
- Data really is unreasonably effective
- New biology and medicine is possible through "data science"

Thank you for your attention

Funding Support

- Lucile Packard Foundation for Children's Health
- NIH: NLM, NIGMS, NCI, NIAID; NIDDK, NHGRI, NIA, NHLBI
- Howard Hughes Medical Institute
- Hewlett Packard
- California Institute for Regenerative Medicine
- PhRMA Foundation
- Stanford Cancer Center

Contact Me

Email: jdudley@stanford.edu
Twitter: @jdudley
Web: buttelab.stanford.edu

Support Staff

- Susan Aptakar
- Alex Skrenchuk
- Meelan Phalank

Butte Lab