Accelerating Translational Medicine using Heterogeneous Data: A Case for Better Metadata

Purvesh Khatri
Institute for Immunity, Transplantation and Infection
Center for Biomedical Informatics Research
Department of Medicine
Stanford University

Email: pkhatri@stanford.edu
@purveshkhatri

“Reproducibility Crisis”

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine 2005

Estimating the reproducibility of psychological science

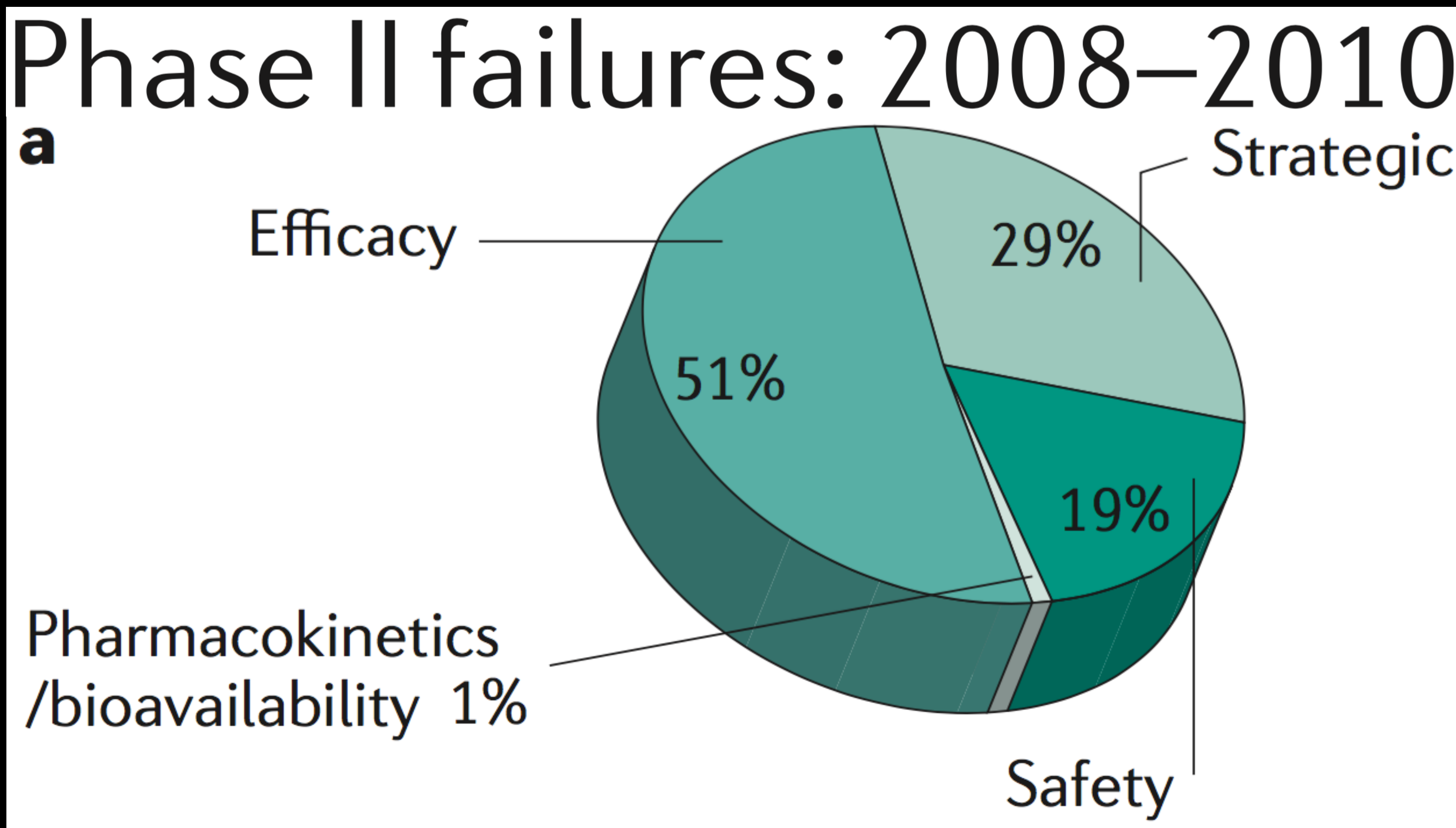
Open Science Collaboration*

Science 2015

Raise standards for preclinical cancer research

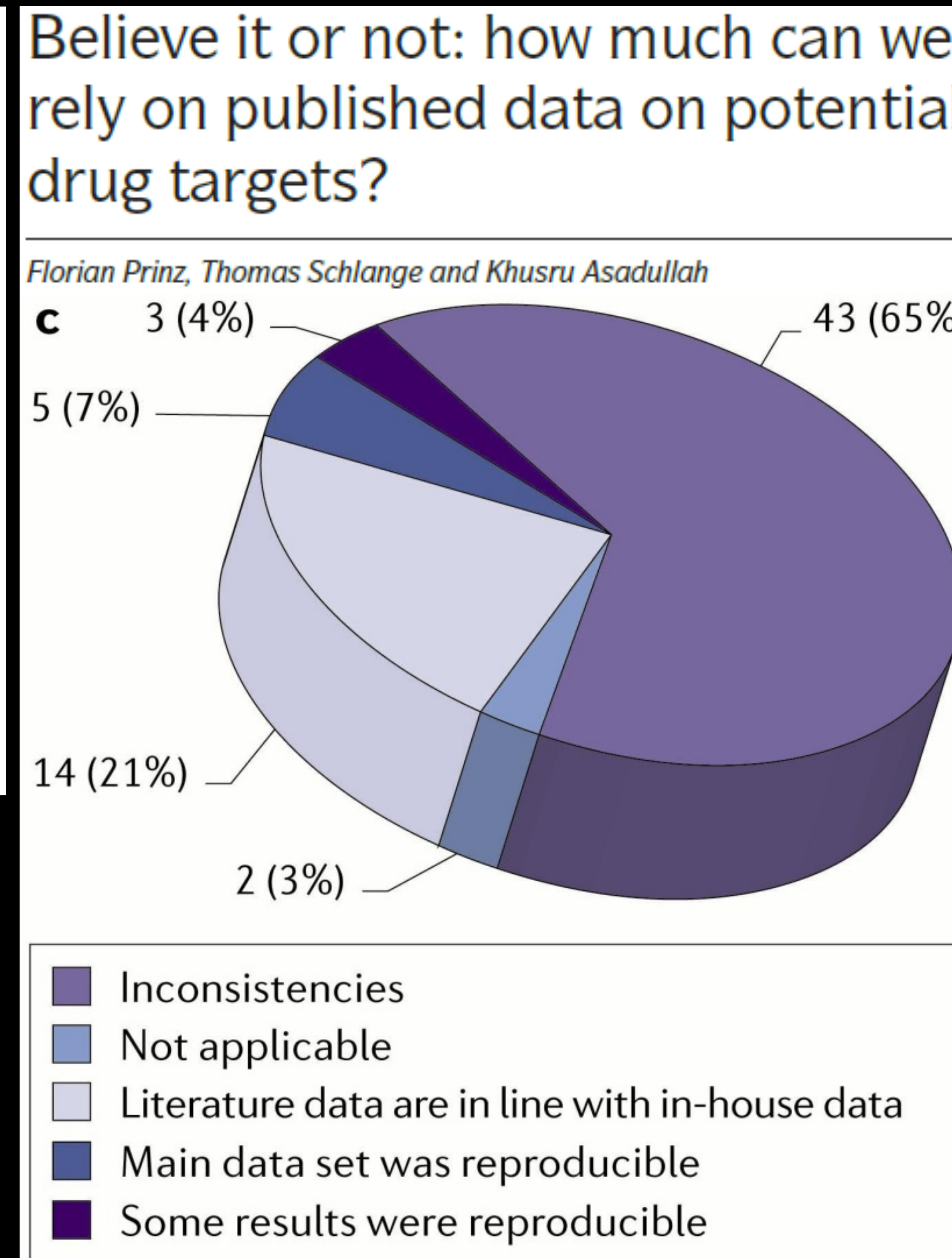
Nature 2012

Lack of reproducibility affects translation



Nat Rev Drug Discovery 2011

- Phase II success rate reduced from 28% to 18%



Nat Rev Drug Discovery 2011

Reasons for lack of reproducibility - Methods and Models

STATISTICAL ERRORS

Nature 2014

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

Of Mice and Not Men: Differences between Mouse and Human Immunology

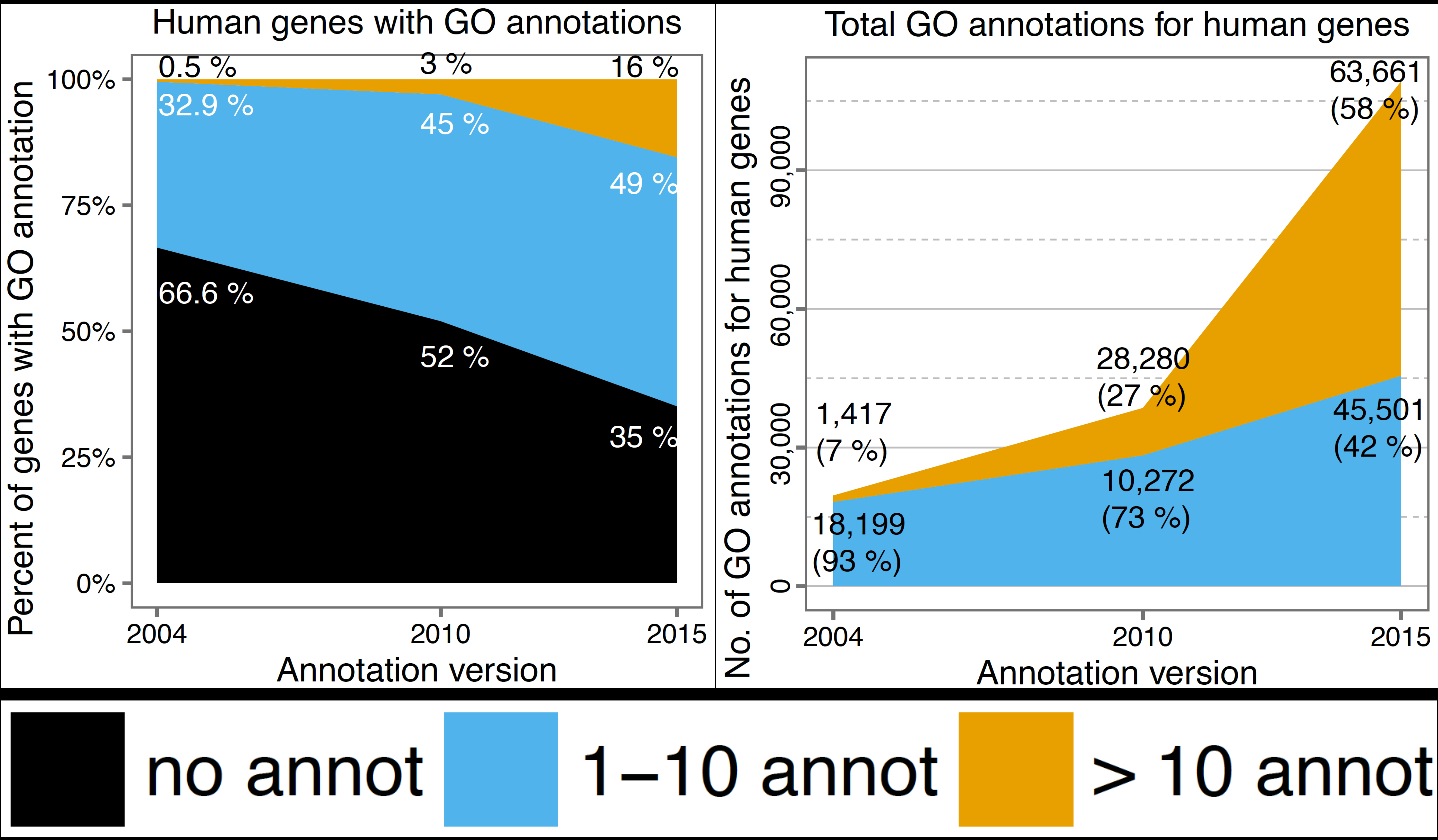
Javier Mestas and Christopher C. W. Hughes[1]

J of Immunology 2004

Genomic responses in mouse models poorly mimic human inflammatory diseases

PNAS 2013

Our biological knowledge is <u>incomplete</u> and <u>biased</u>



Aurelie Tomczak et al.
(under review)

Lack of biological and technological heterogeneity is a significant problem

Traditional approach - <u>reduce</u> heterogeneity

- Single cohort
  - Clinical homogeneity
  - Minimize technical variance
  - Internal validation

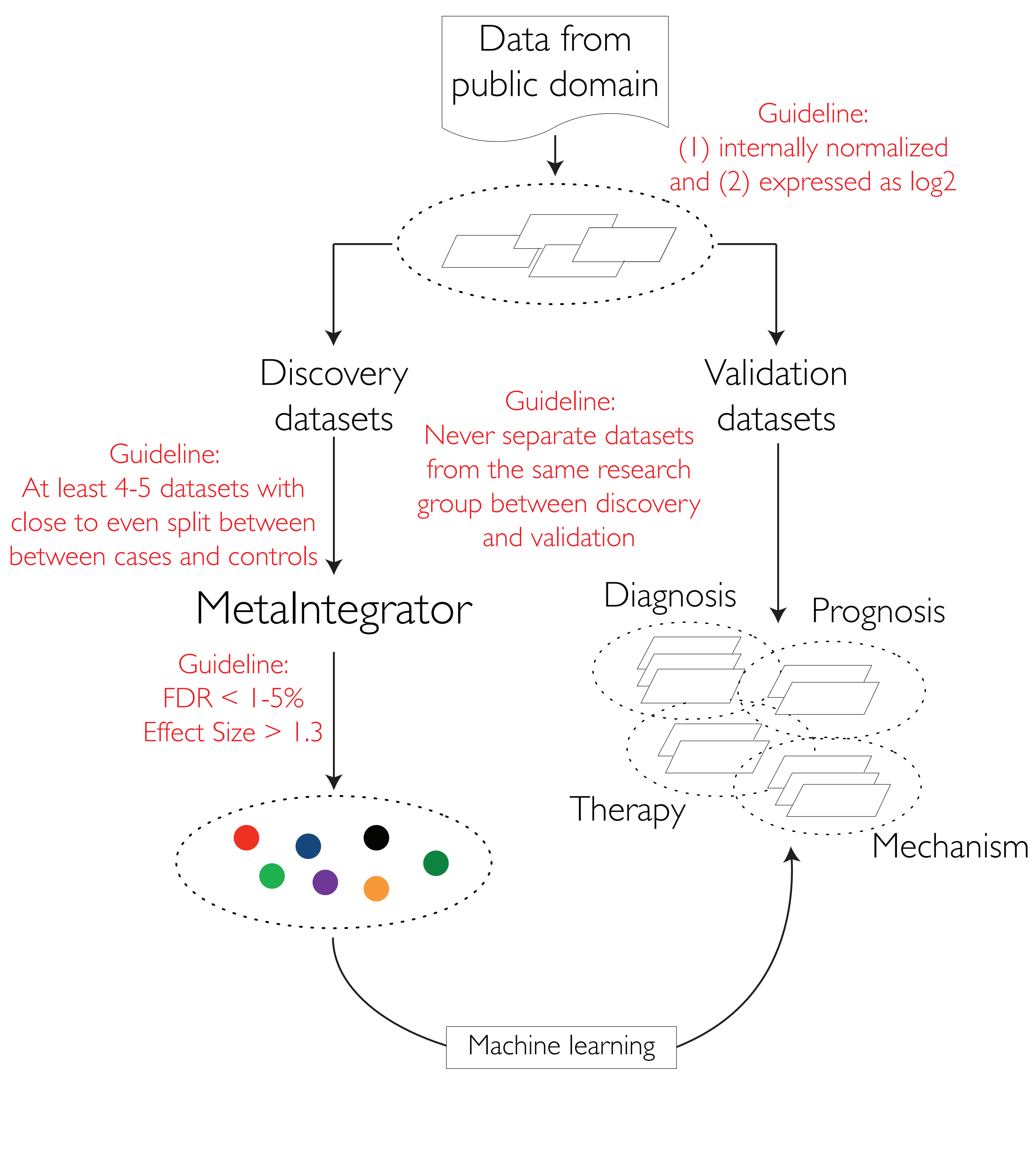
- Does not capture heterogeneity of a disease
- Results are difficult to generalize

Embrace heterogeneity

- Public data - multiple datasets asking the same question
    - Clinical heterogeneity
    - Different treatments
    - Different technologies

- Generalizable results
- Unexpected results are more "believable"
- *"Dirty data" - integration is challenging*

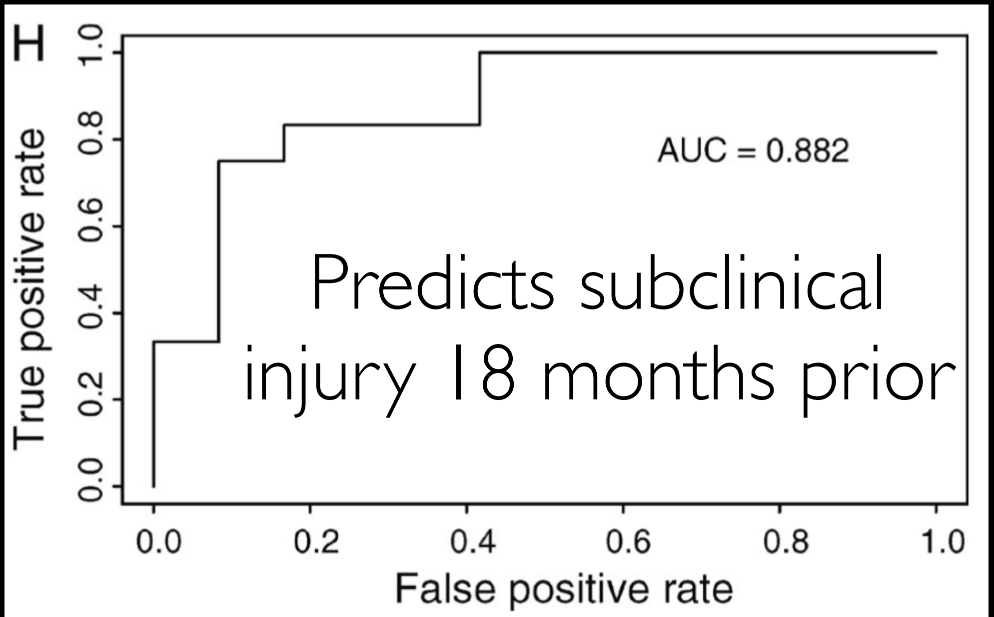
Framework for leveraging heterogeneity



Sweeney et al. NAR 2016

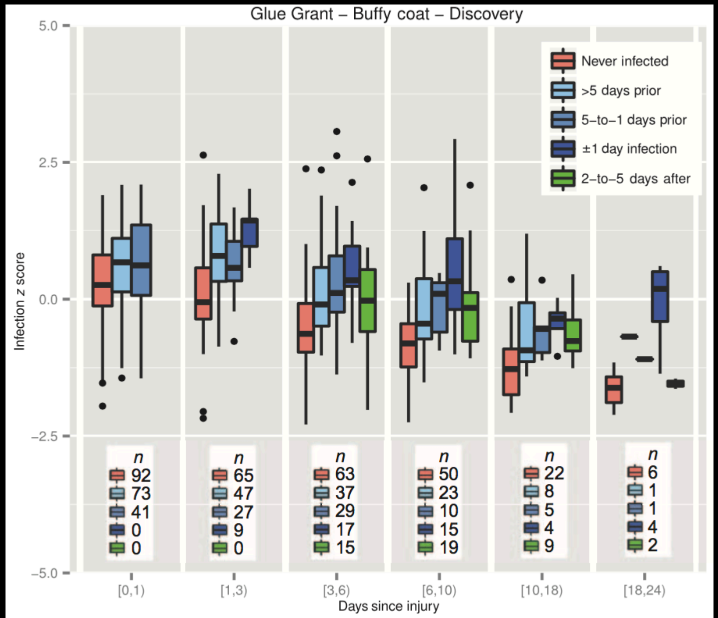
Diagnostic and Prognostic Markers using Heterogeneous Data

Common rejection module across all solid organs



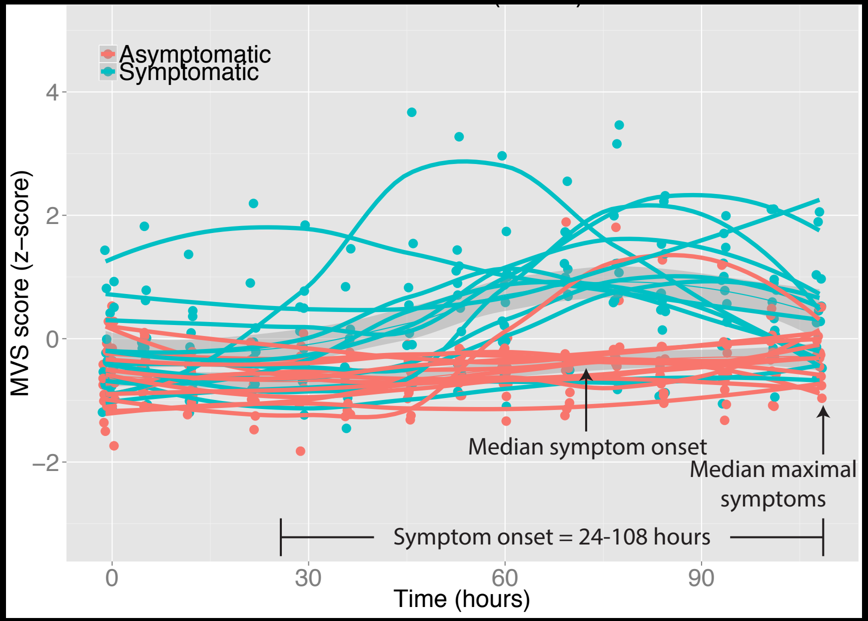
Khatri *et al.*
J Exp Med 2013

Sepsis diagnosis 1-to-5 days prior



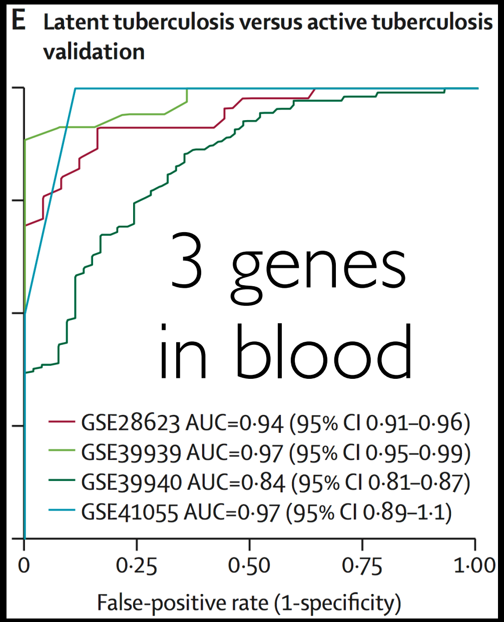
Sweeney *et al.*
Sci Trans Med 2015

Common host response to viral infections



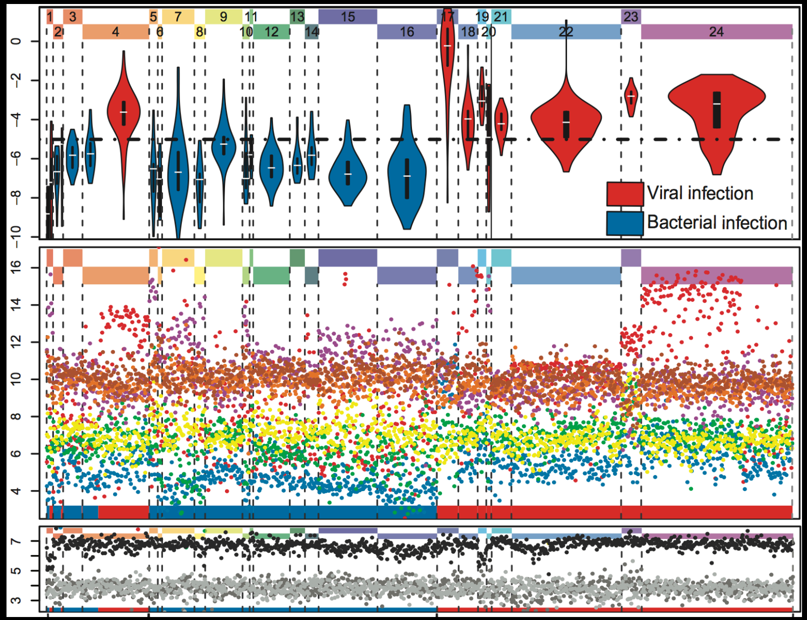
Andres-Terre *et al.*
Immunity 2015

TB - satisfies WHO TPP



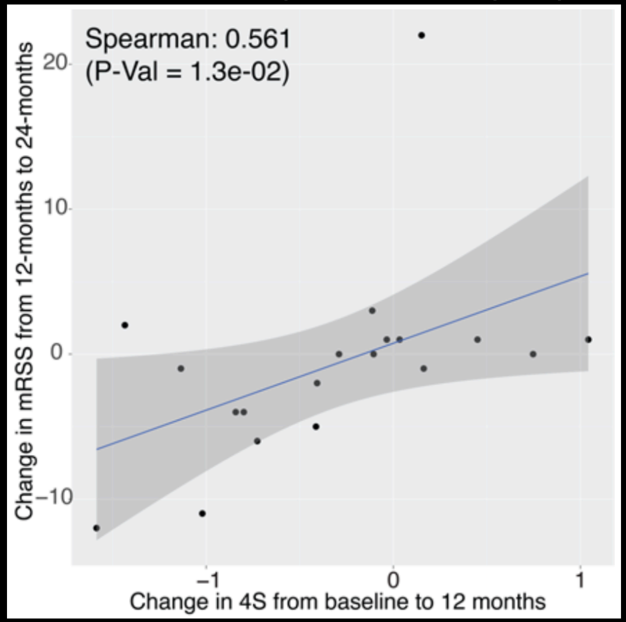
Sweeney *et al.*
Lancet Resp Med 2016

Bacterial vs viral infection diagnosis



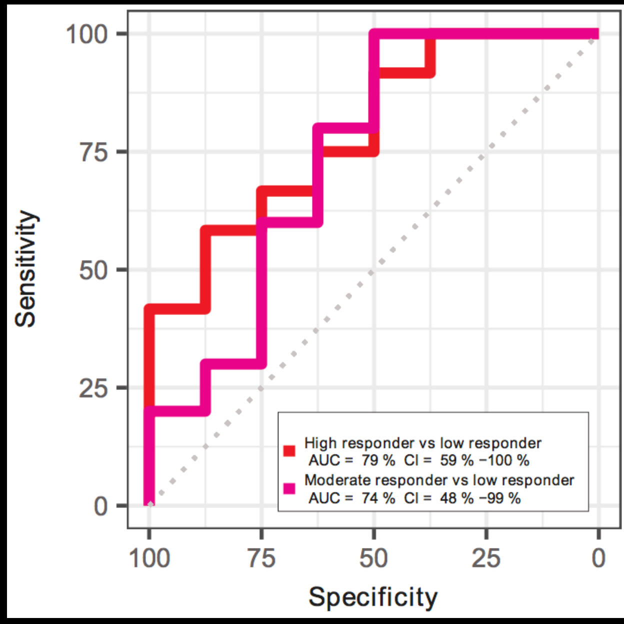
Sweeney *et al.*
Sci Trans Med 2016

Scleroderma - predicts treatment response 1 yr prior



Lofgren *et al.*
JCI Insight 2016

Predict response to vaccine at baseline

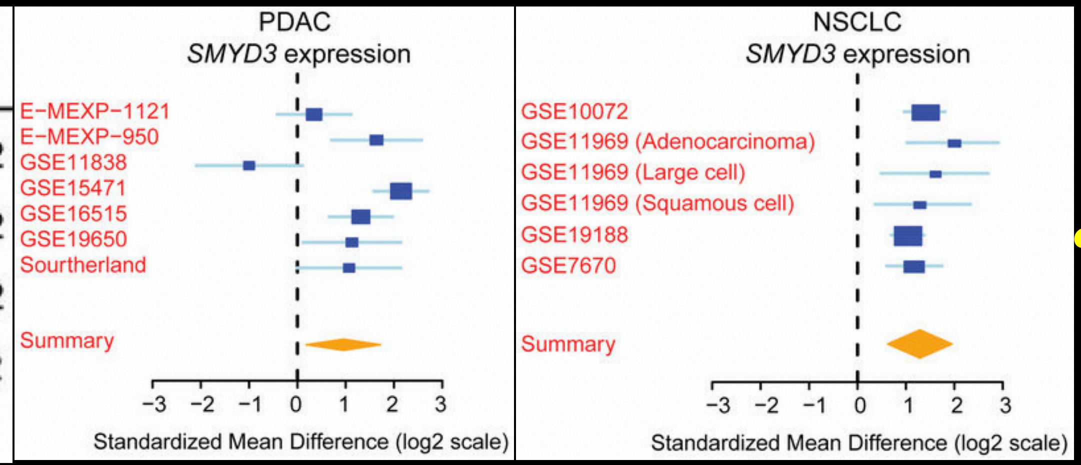


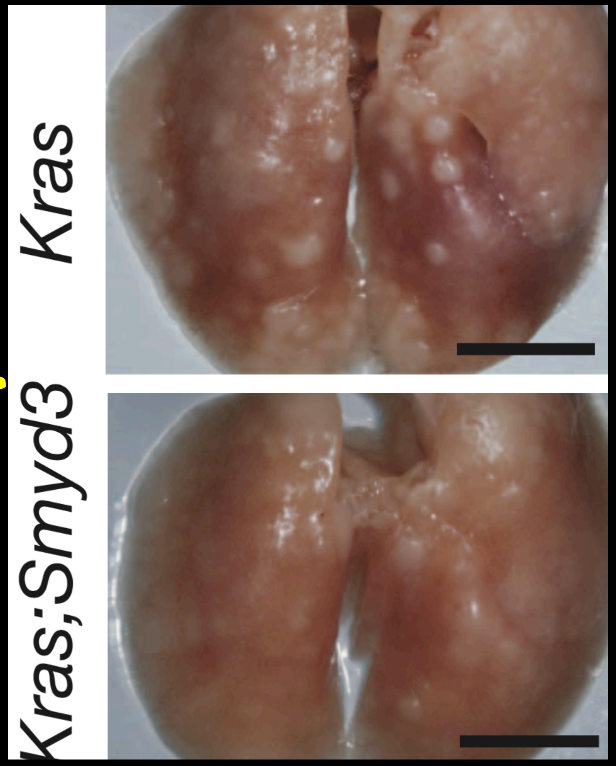
HIPC-CHI
Sci Immunology 2017

Target Discovery using Heterogeneous Data

Mazur *et al. Nature* 2014

KMT	Meta effect size	p-value	FDR
SMYD3	0.952	1.59E-02	5.88E-02
MLL5	0.475	6.90E-03	3.25E-02
EZH2	0.833	7.38E-03	3.41E-02
SETD5	0.338	4.08E-03	2.21E-02
WHSC1L1	0.471	3.49E-02	0.103

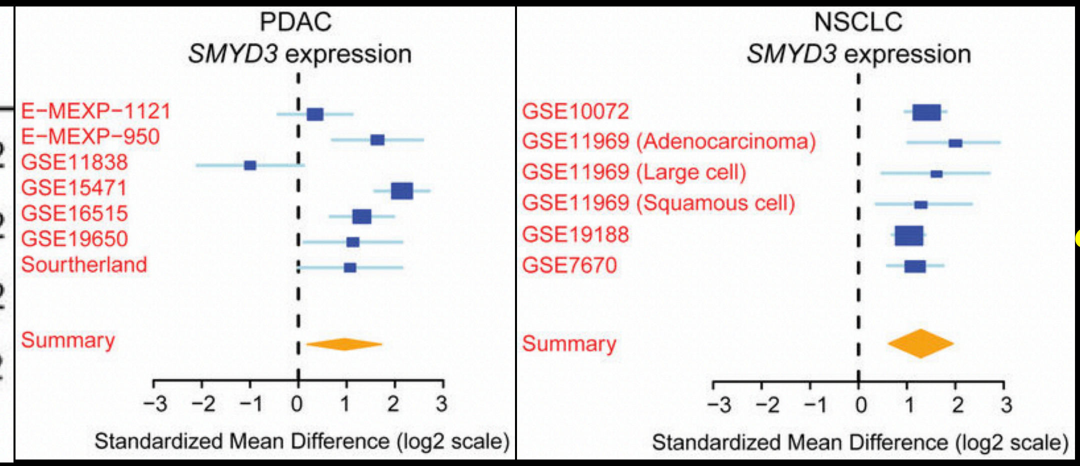


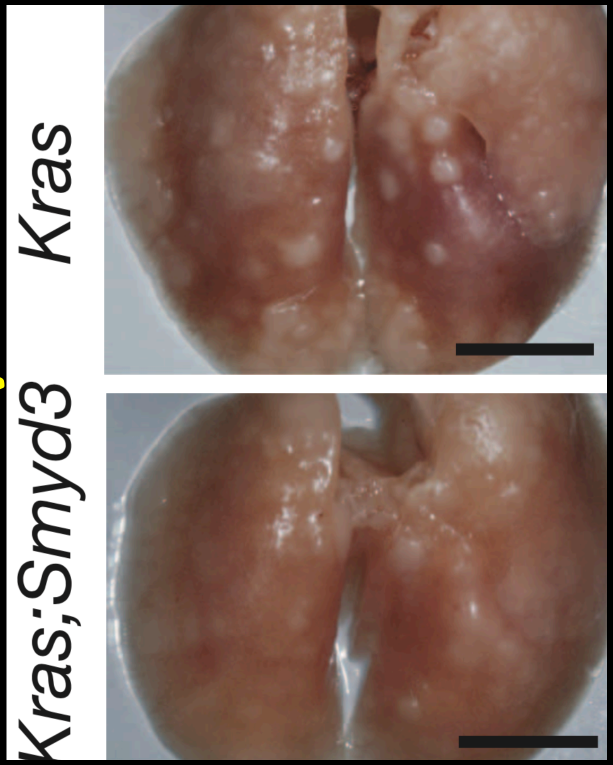


Target Discovery using Heterogeneous Data

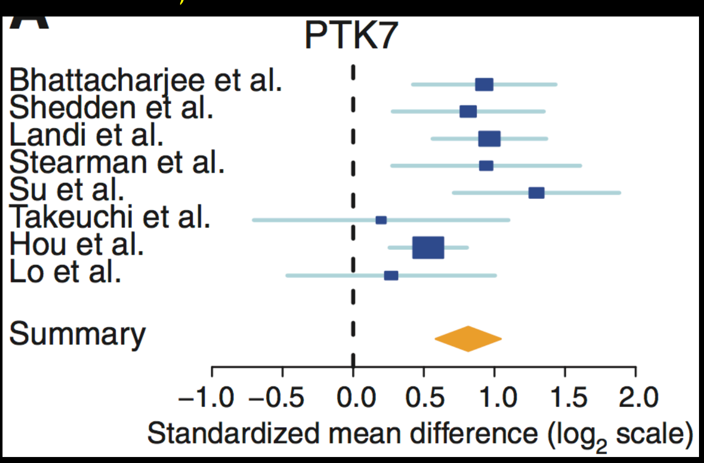
Mazur *et al. Nature* 2014

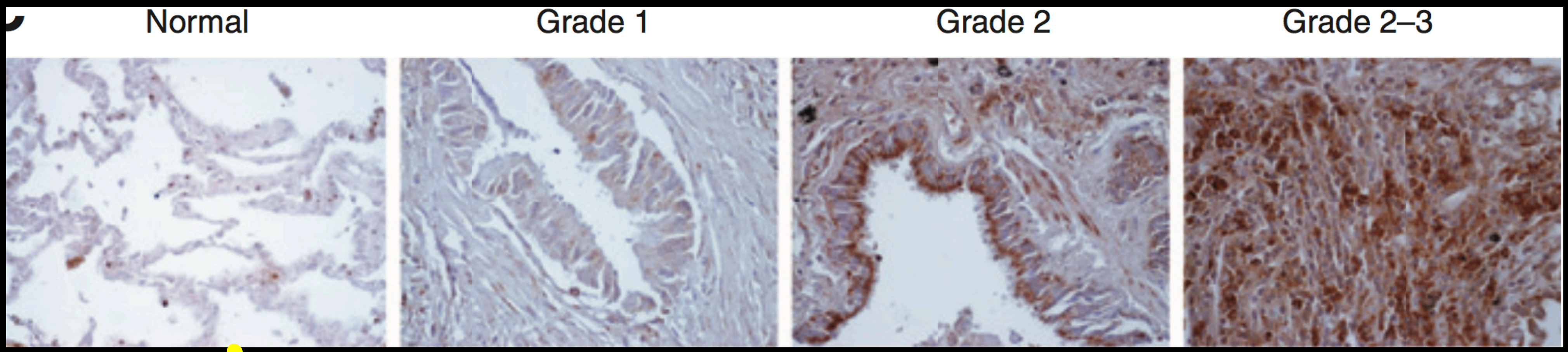
KMT	Meta effect size	p-value	FDR
SMYD3	0.952	1.59E-02	5.88E-02
MLL5	0.475	6.90E-03	3.25E-02
EZH2	0.833	7.38E-03	3.41E-02
SETD5	0.338	4.08E-03	2.21E-02
WHSC1L1	0.471	3.49E-02	0.103





Chen*, Khatri* *et al. Cancer Research* 2014

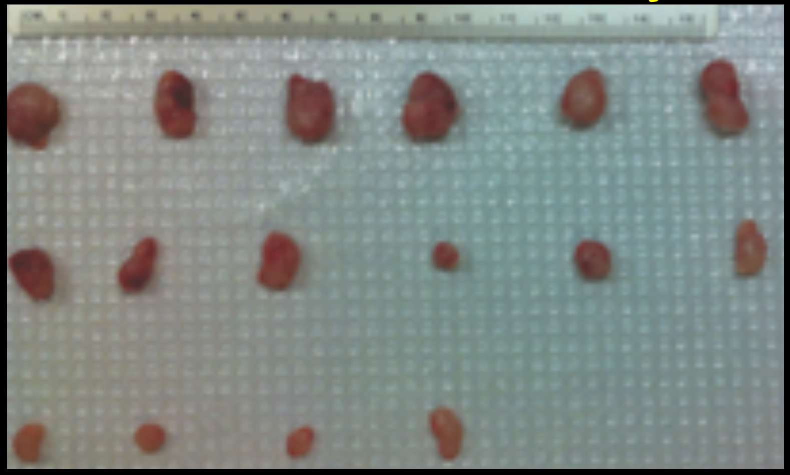




shCtrl

shPTK7

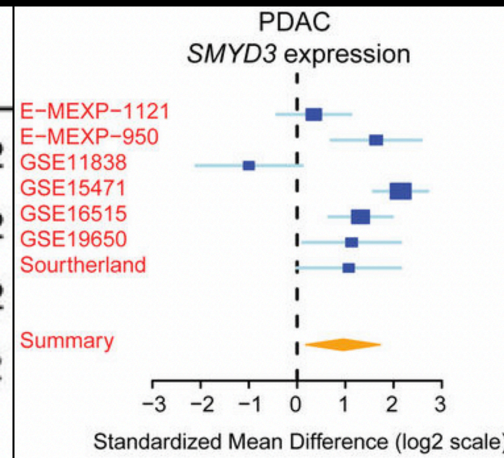
shPTK7

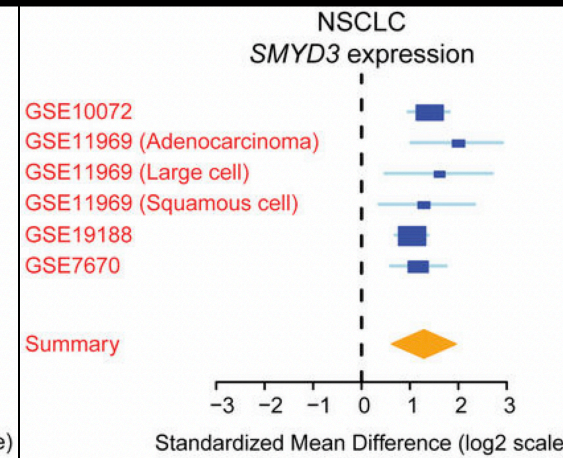


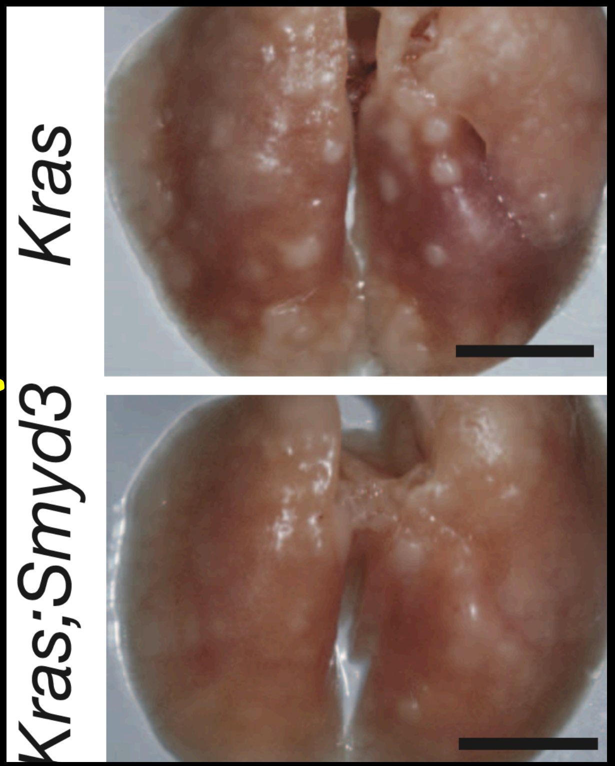
Target Discovery using Heterogeneous Data

Mazur et al. Nature 2014

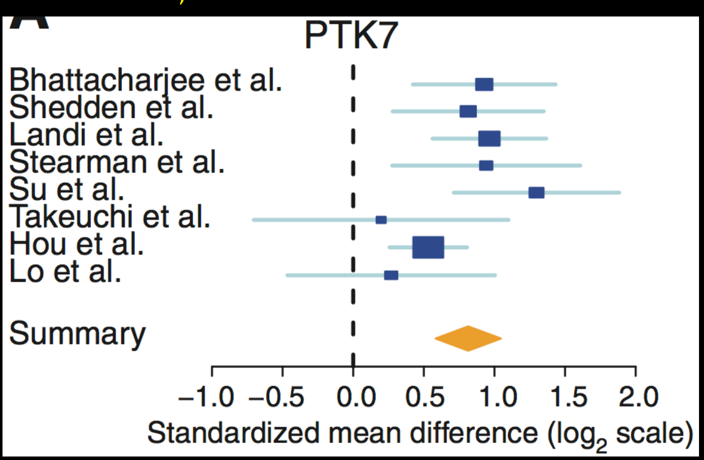
KMT	Meta effect size	p-value	FDR
SMYD3	0.952	1.59E-02	5.88E-02
MLL5	0.475	6.90E-03	3.25E-02
EZH2	0.833	7.38E-03	3.41E-02
SETD5	0.338	4.08E-03	2.21E-02
WHSC1L1	0.471	3.49E-02	0.103

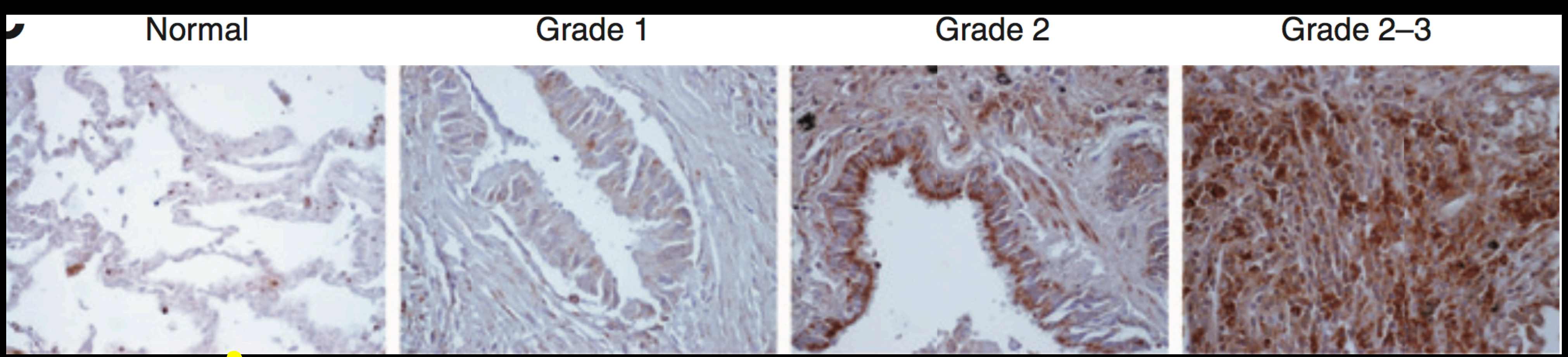






Chen, Khatri* et al. Cancer Research 2014*

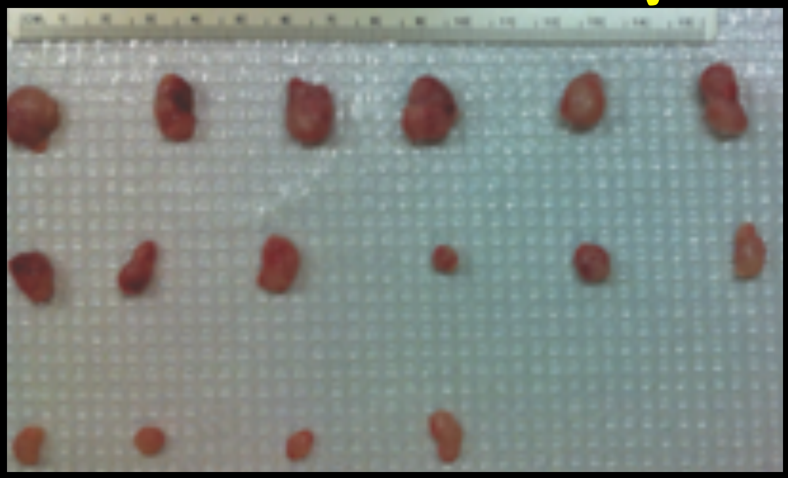




shCtrl

shPTK7

shPTK7



SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

CANCER

A PTK7-targeted antibody-drug conjugate reduces tumor-initiating cells and induces sustained tumor regressions

Marc Damelin,[1*†] Alexander Bankovich,[2*] Jeffrey Bernstein,[2] Justin Lucas,[1] Liang Chen,[1] Samuel Williams,[2] Albert Park,[2] Jorge Aguilar,[2] Elana Ernstoff,[1] Manoj Charati,[1] Russell Dushin,[3] Monette Aujay,[2] Christina Lee,[2] Hanna Ramoth,[2] Milly Milton,[2] Johannes Hampl,[2] Sasha Lazetic,[2] Virginia Pulito,[1] Edward Rosfjord,[1] Yongliang Sun,[3] Lindsay King,[3] Frank Barletta,[1] Alison Betts,[3] Magali Guffroy,[1] Hadi Falahatpisheh,[1] Christopher J. O'Donnell,[3] Robert Stull,[2] Marybeth Pysz,[2] Paul Escarpe,[2] David Liu,[2] Orit Foord,[2] Hans Peter Gerber,[1] Puja Sapra,[1†] Scott J. Dylla[2†]

A comment from an NIH grant reviewer

Weaknesses

- PI completely inexperienced in scleroderma – seems to like bright objects and flits from one shiny project to another without focus.

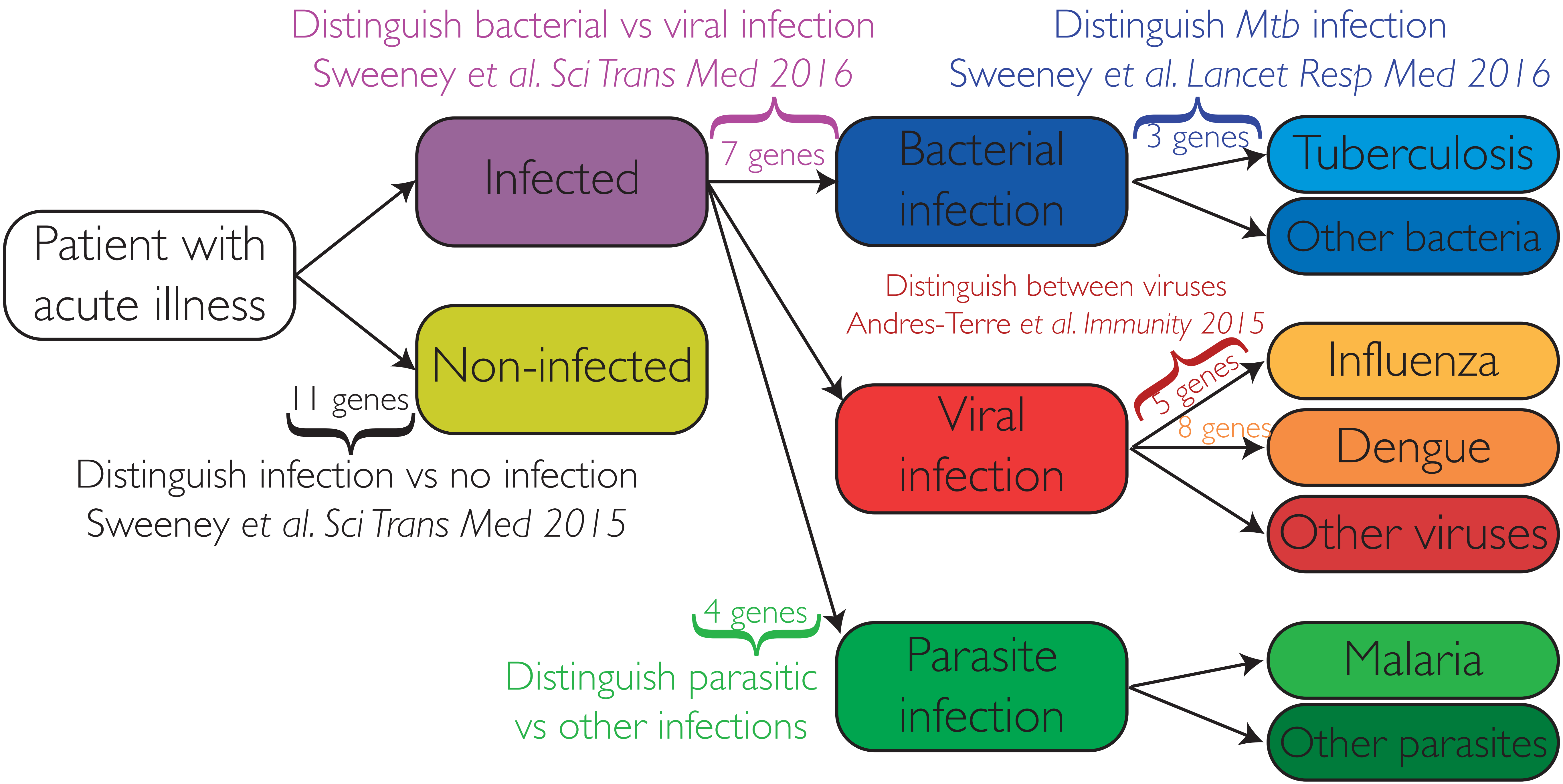
A comment from an NIH grant reviewer

Weaknesses

- PI completely inexperienced in scleroderma – seems to like bright objects and flits from one shiny project to another without focus.

But…there is a method to my ADD!

"Reading the immune response" to build phylogeny of host response to infectious diseases



Meeting Report

High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting

28–29 April 2014

Geneva, Switzerland

World Health Organization

Executive summary

- a point-of-care non-sputum-based test capable of detecting all forms of TB by identifying characteristic biomarkers or biosignatures (known as the biomarker test);
- a point-of-care triage test, which should be a simple, low-cost test that can be used by first-contact health-care providers to identify those who need further testing (the triage test);
- a point-of-care sputum-based test to replace smear microscopy for detecting pulmonary TB (the smear-replacement test);
- a rapid drug-susceptibility test that can be used at the microscopy-centre level of the health-care system to select first-line regimen-based therapy (the rapid DST test).

	Year	Reference	Platform	Use	Country	Age	HIV status	Active tuberculosis culture or smear	Healthy controls	Latent tuberculosis	Other disease	Active tuberculosis	Treatment	Total	Miscellaneous
GSE19491	2010	Berry[8]	GPL6947	Discovery	South Africa, UK, USA	Adults	Negative	Positive	86	69	193	31	..	409	Other disease breakdown: 28 ASLE, 82 PSLE, 31 Still's, 52 *Streptococcus* and/or *Staphylococcus* infection; post-treatment samples not used.
GSE25534	2010	Maertzdorf[30]	GPL1708	Validation	South Africa	Adults	Negative	Positive	6	19	..	19	..	44	Two-colour array (on-chip comparisons between healthy controls, latent tuberculosis, and active tuberculosis)
GSE28623	2011	Maertzdorf[22]	GPL4133/ GPL6480	Validation	The Gambia	Adults	Negative	Positive	37	25	..	46	..	108	..
Cliff Combined Dataset	2013	Cliff[13]	GPL570	Validation	South Africa	Adults	Negative	Positive	..	..	..	36	117	153	Treatment measured at 1, 2, 4, and 26 weeks
GSE34608	2012	Maertzdorf[24]	GPL4133/ GPL6480	Validation	Germany	Adults	Negative	Positive	18	..	18	8	..	44	Other diseases all sarcoid
GSE37250	2014	Kaforou[7]	GPL10558	Discovery	Malawi, South Africa	Adults	Positive and negative	Positive	..	167	175	195	..	537	See reference for other disease distributions; 194 patients with other diseases reported but only 175 available with microarrays.
GSE39939	2014	Anderson[6]	GPL10558	Validation	Kenya	Child-ren	Positive and negative	Positive and negative	..	14	64	44 negative, 35 positive	..	157	Other diseases breakdown: 33 pneumonia, 5 sepsis, 7 malnutrition, 19 other
GSE39940		Anderson[6]		Validation	Malawi, South Africa	Child-ren	Positive and negative	Positive	..	54	169	111	..	334	Other diseases breakdown: 86 pneumonia, 8 CLD, 11 URI, 34 other infections, 12 malignancy, 18 other
GSE40553	2012	Bloom[9]	GPL10558	Validation	South Africa, UK	Adults	Negative	Positive	..	..	..	36	130	166	Treatment measured at 0·5, 2, 4, 6, and 12 months. Two cohorts followed. Latent tuberculosis not used; overlaps with GSE19491
GSE41055	2013	Verhagen[10]	GPL5175	Validation	Venezuela	Child-ren	Negative	Positive and negative	9	9	..	7 negative; 2 positive	..	27	..
GSE42834	2014	Bloom[9]	GPL10558	Discovery	UK, France	Adults	Negative	Positive	118	..	123	40	..	281	Other diseases breakdown: 83 sarcoidosis, 24 pneumonia, 16 cancer
GSE56153	2012	Ottenhoff[23]	GPL6883	Validation	Indonesia	Adults	Negative	Positive	18	..	..	18	35	71	Treatment measured at 8 and 28 weeks
GSE62147	2015	Tientcheu[29]	GPL6480	Validation	The Gambia	Adults	Negative	Positive	..	..	..	26	26	52	*M africanum* and *M tuberculosis*
GSE74092	2015	Maertzdorf[12]	RT-PCR array GPL21040	Validation	India	Adults	Negative	Positive	76	..	..	113	..	189	*KLF2* not present in these data

ASLE=adult systemic lupus erythematosus. PSLE=paediatric systemic lupus erythematosus. CLD=chronic lung disease. URI=upper respiratory infection.

***Table:* Summary table of all datasets that matched inclusion criteria (whole blood, clinically active pulmonary tuberculosis)**

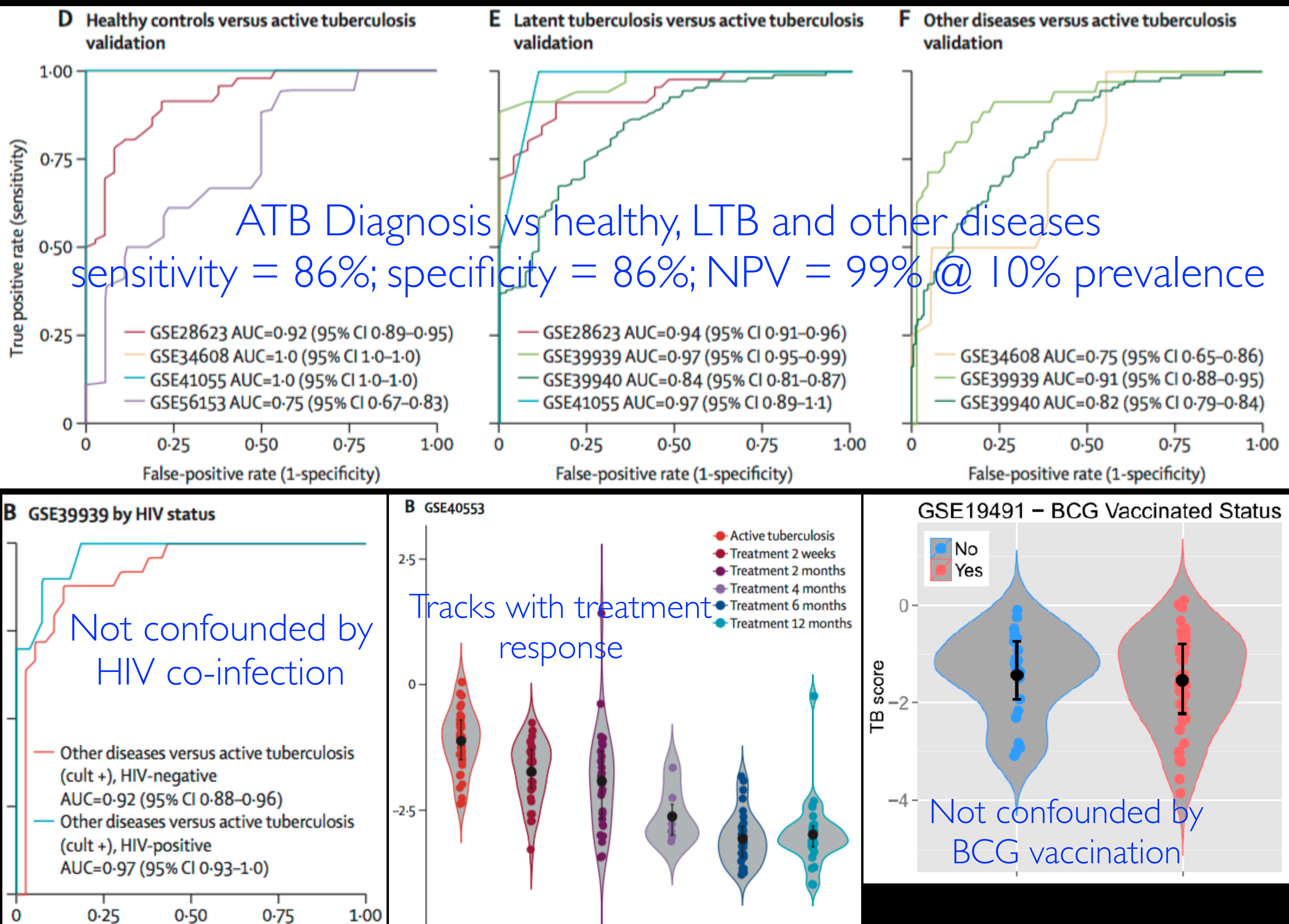
11 countries

14 cohorts

2,572 samples

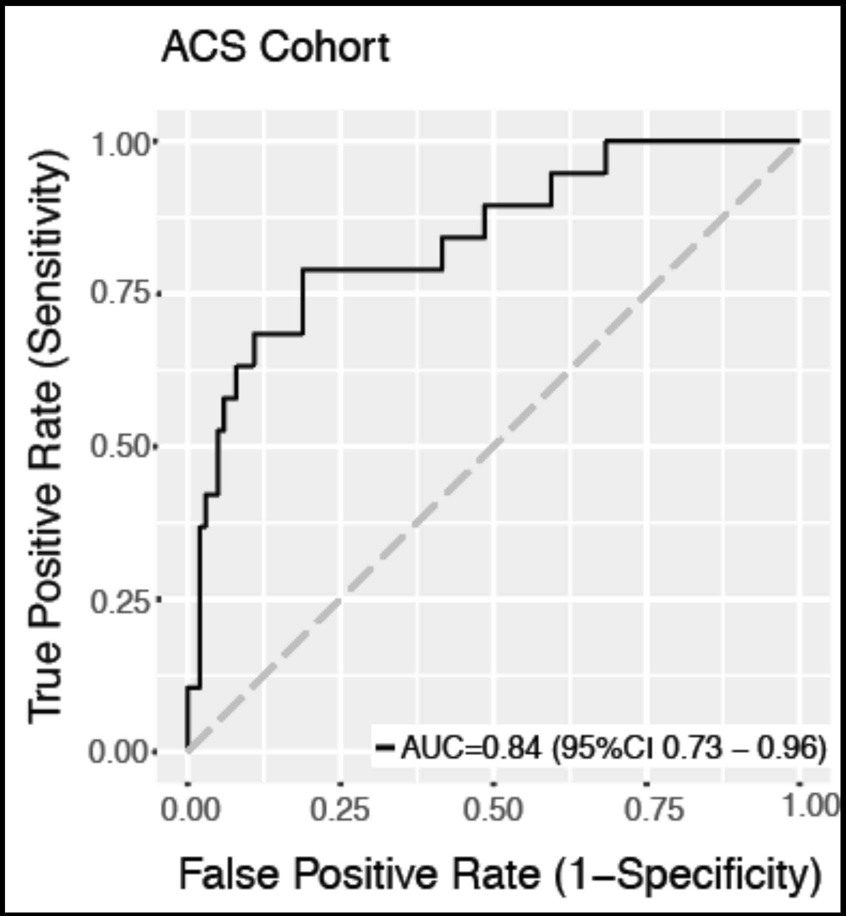
3 genes

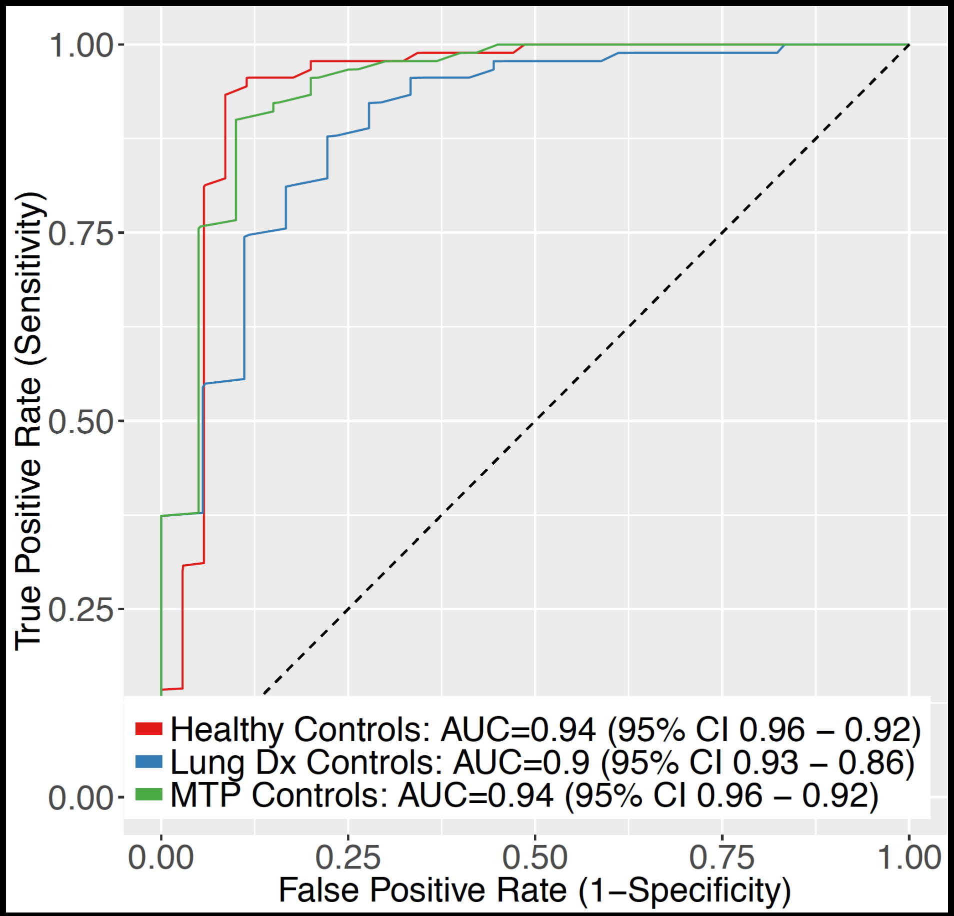
(DUSP3, GBP5, KLF2)

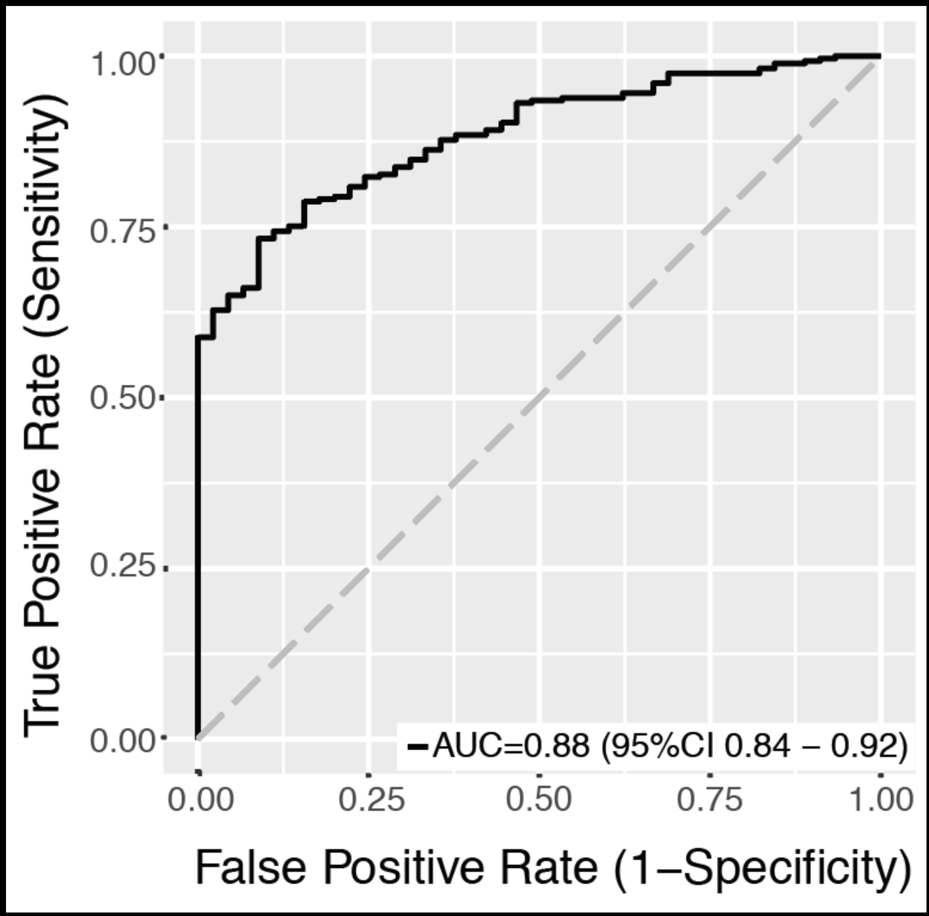


Sweeney et al. Lancet Resp Med 2016

3-gene signature distinguishes ATB in prospective cohorts





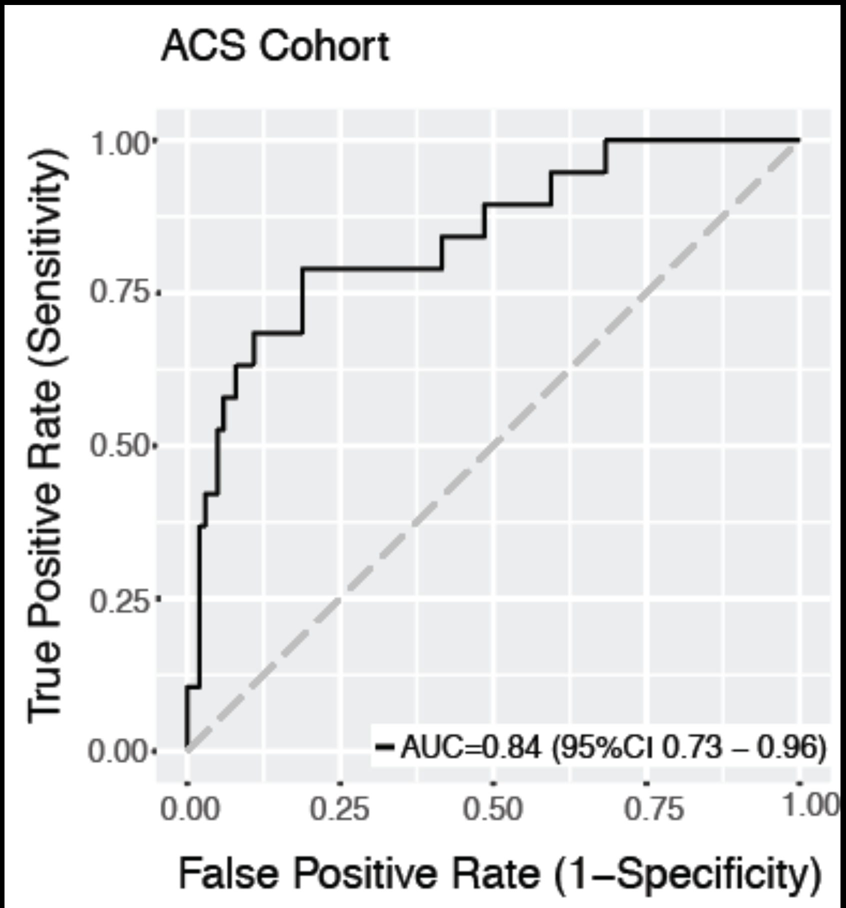


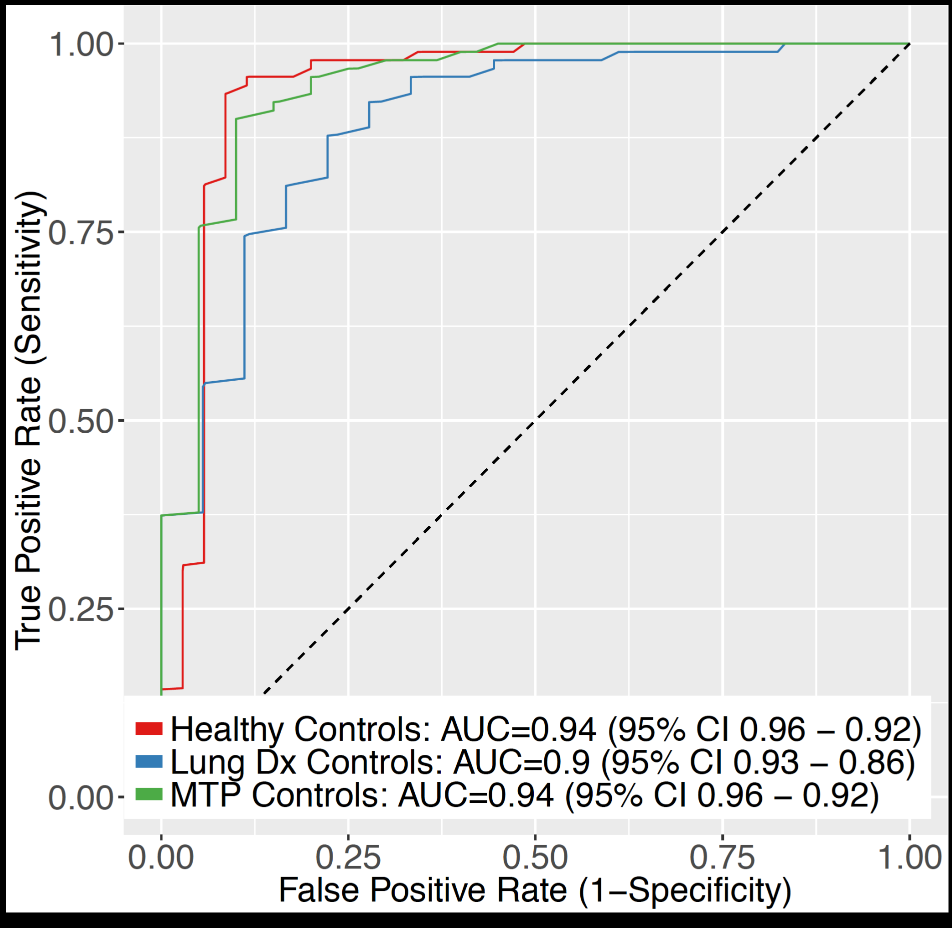
Zak et al. Lancet 2016
Adolescents
LTB vs ATB
RNAseq

Zak et al. Tuberculosis 2017
Adults
ATB vs controls
RNAseq

Warsinske *et al.*
Active screen in adults
ATB vs controls
PCR

3-gene signature distinguishes ATB in prospective cohorts





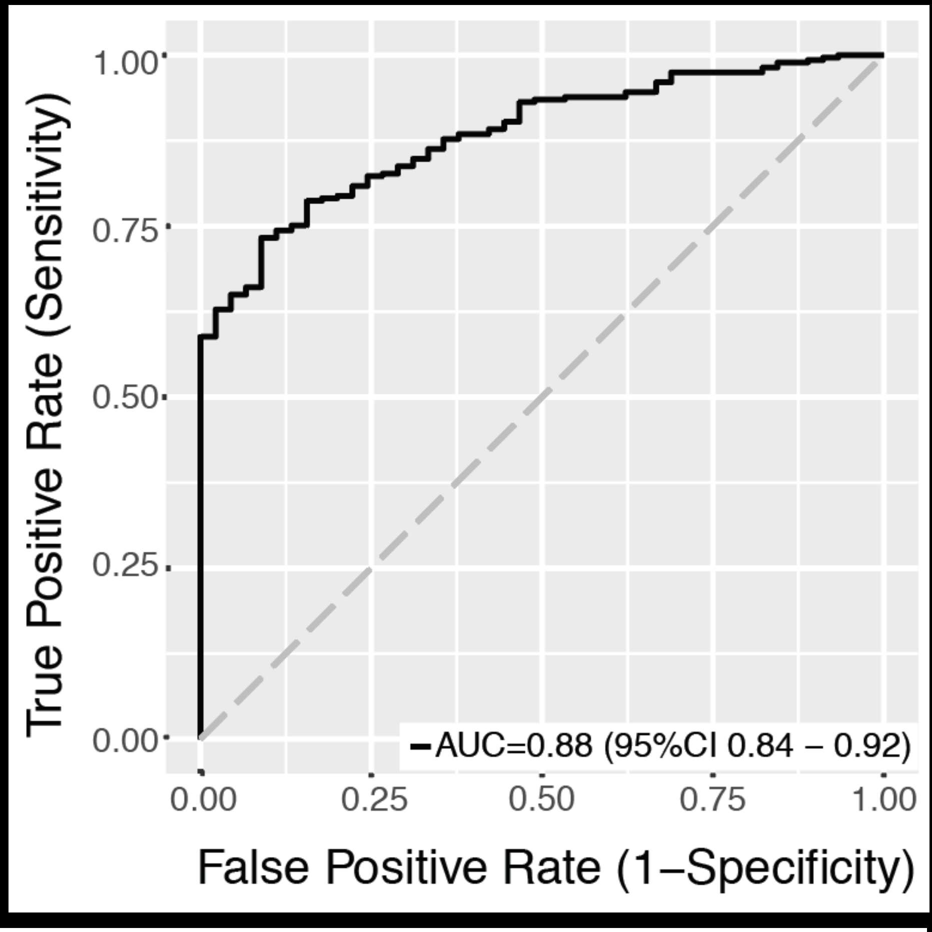
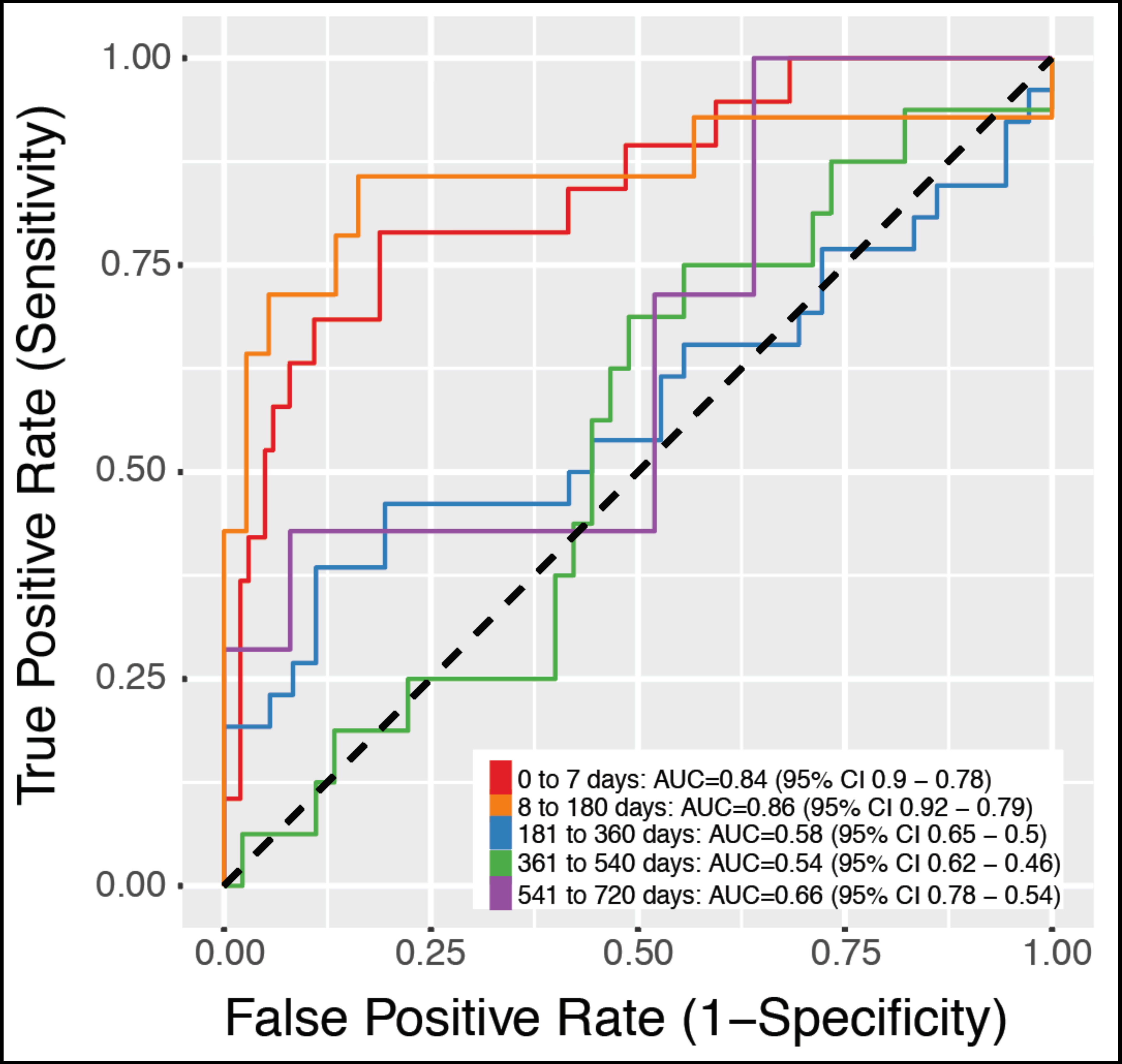


Table 3 Maximized sensitivity values obtained from the ROC analysis of *GBP5*, *DUSP3* and *KLF2* combinations in WB cohort test.

Francisco et al. J of Infection 2017

	GBP5	*DUSP3*	*KLF2*	*GBP5,DUSP3*	*GBP5,KLF2*	*DUSP3,KLF2*	*GBP5,DUSP3,KLF2*
ATB vs HC							
AUC	0.85	0.73	0.62	0.84	0.86	0.77	0.85
95%CI	0.81-0.90	0.67-0.78	0.56-0.68	0.80-0.89	0.82-0.91	0.72-0.82	0.81-0.89
Sensitivity	80.6%	61.8%	31.3%	77.8%	77.8%	66.0%	85.5%
Specificity	90.9%	78.0%	96.7%	89.5%	87.1%	82.3%	70.8%

3-gene signature <u>predicts</u> progression from LTB to ATB



Where we are today

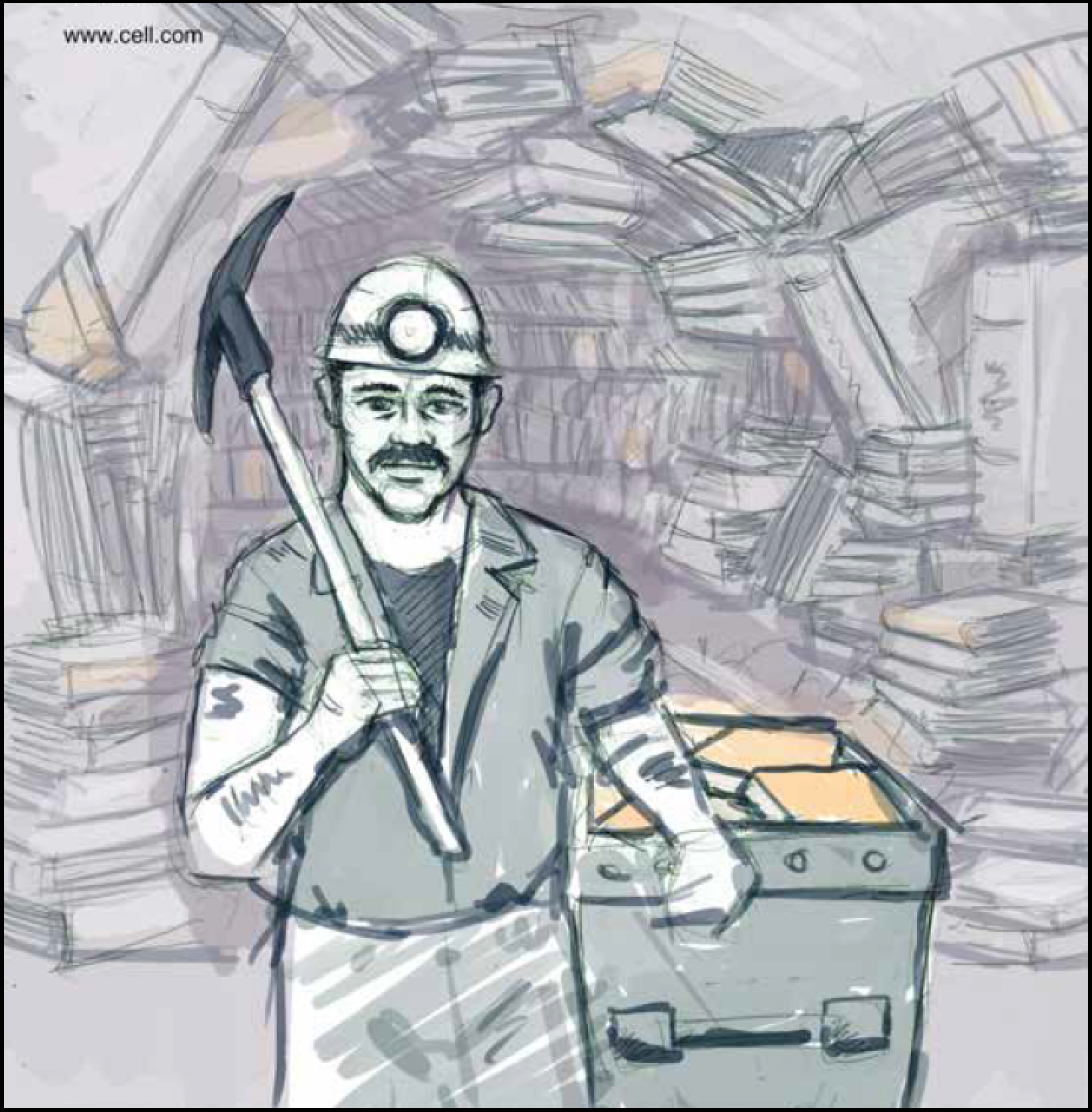
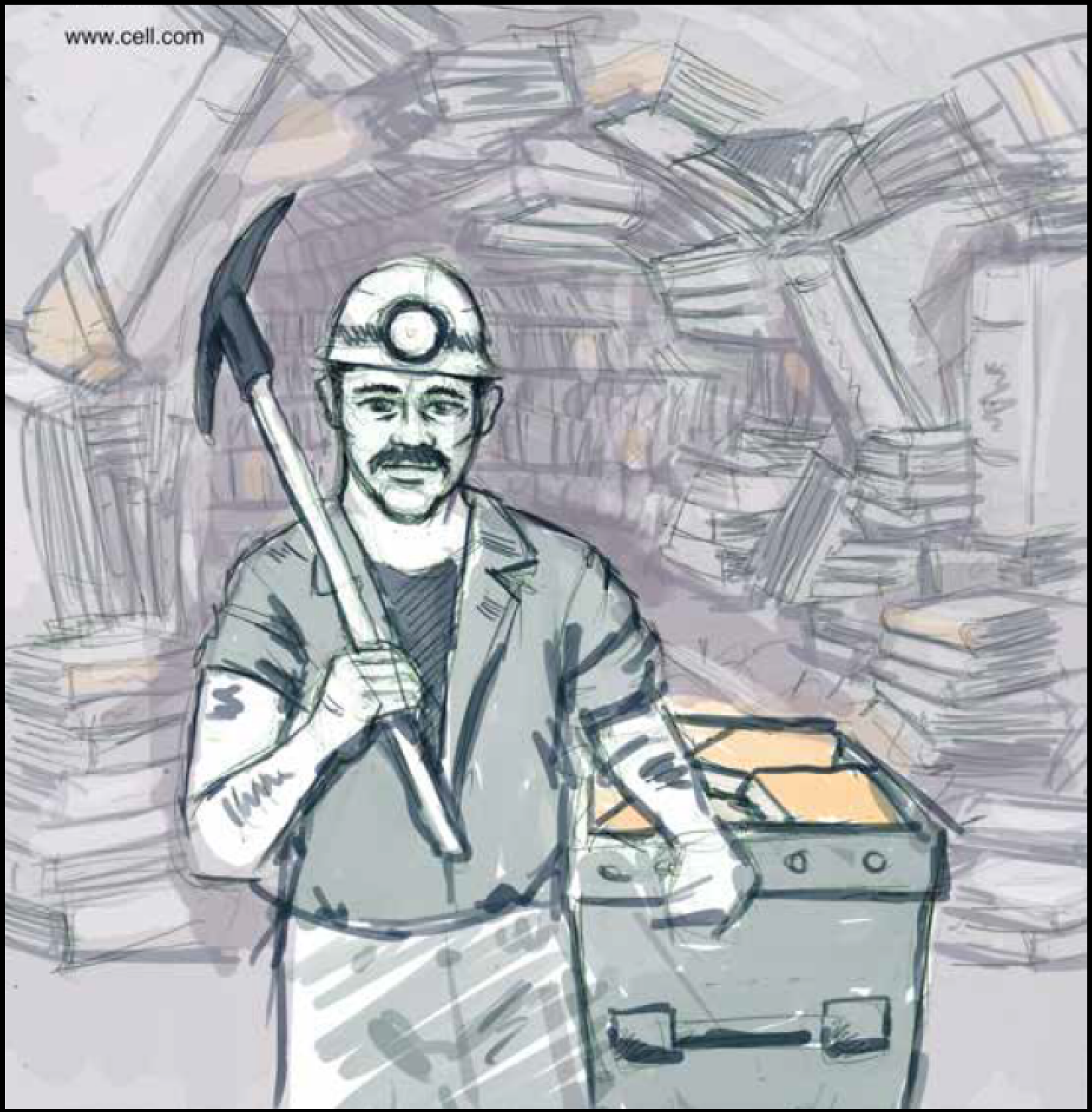


Image courtesy:
Chloe McDougall

Where we are today



Where we want to go

Image courtesy:
Chloe McDougall

Where we are today

Better metadata

Where we want to go

Image courtesy:
Chloe McDougall

Where we are today

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

Better metadata

Where we want to go

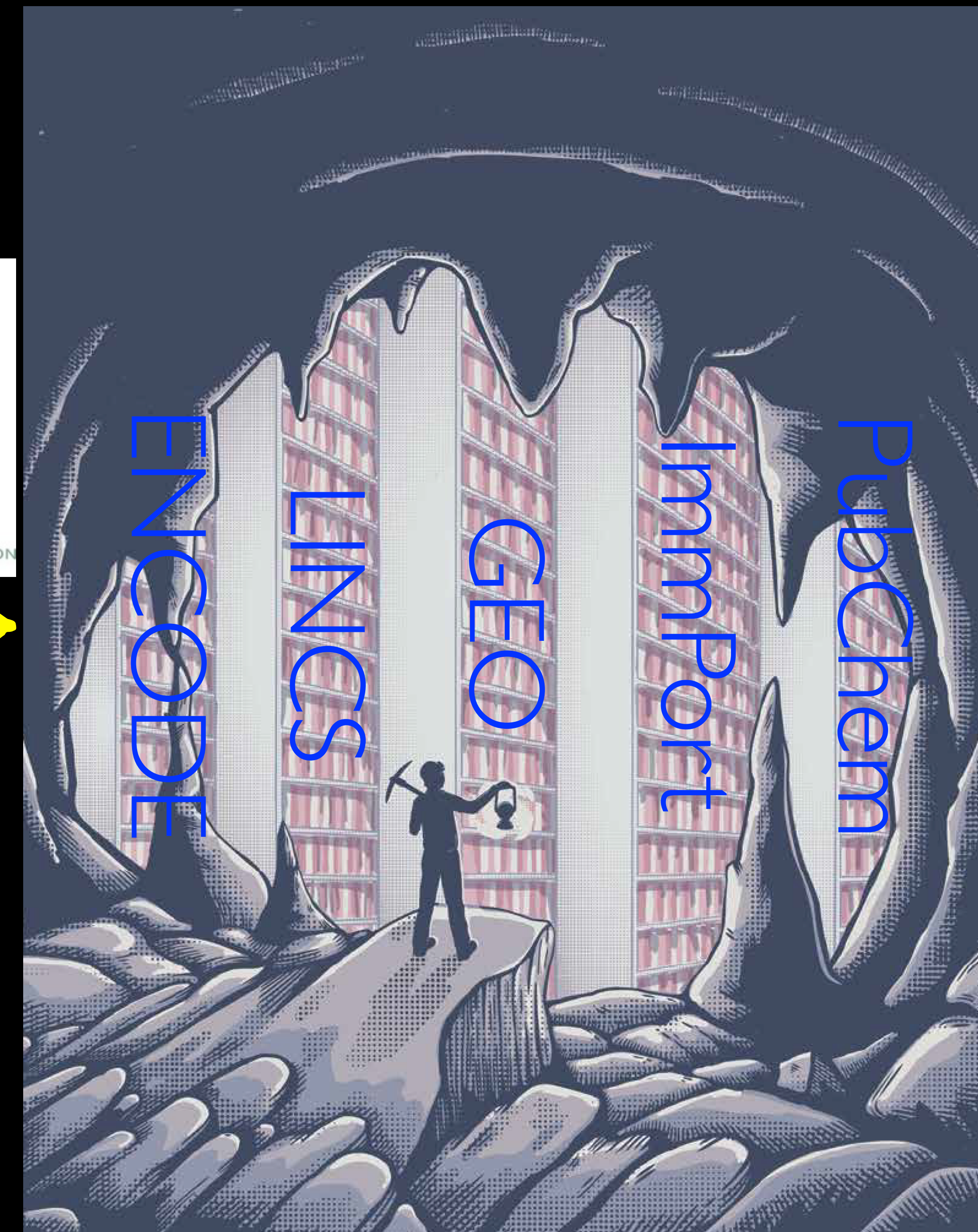
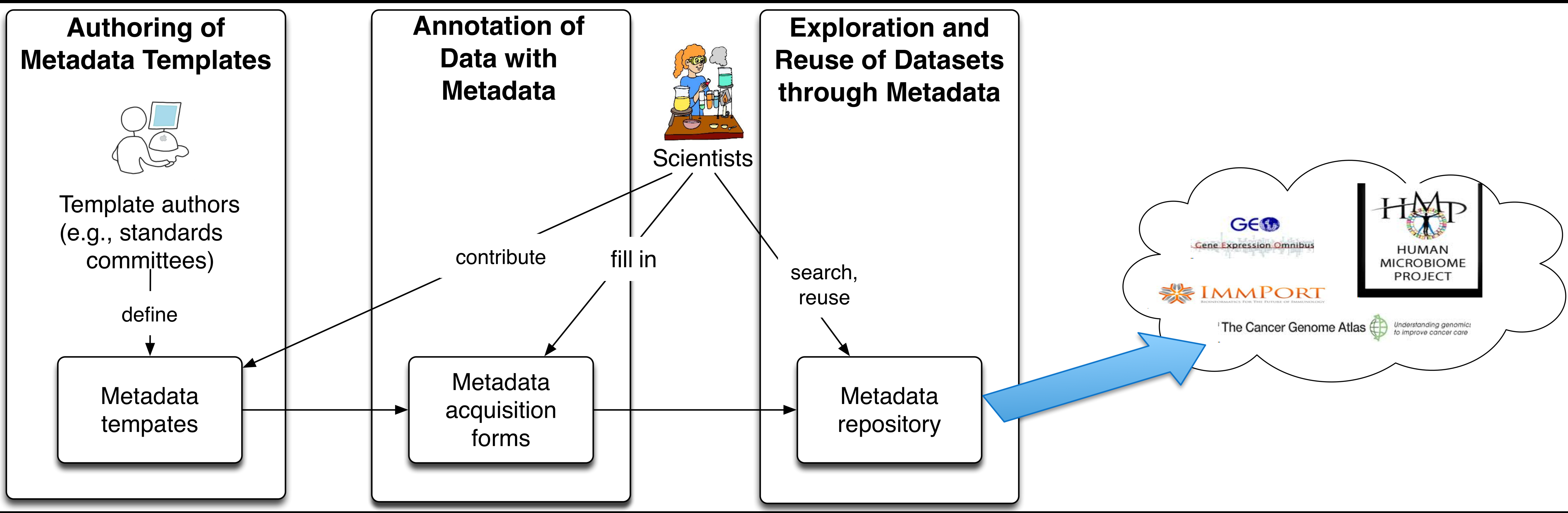


Image courtesy: Chloe McDougall

The CEDAR approach to better metadata



- Template editor
- Metadata editor
- Metadata repository

Courtesy: Mark Musen

Summary

- Heterogeneity: a blessing in disguise
- Leverage biological and technical heterogeneity
    - Increase reproducibility
    - Accelerate translational medicine
- “Data reproducibility” versus “reporting reproducibility”
- Need for better metadata
    - CEDAR

Acknowledgements

Khatri Lab

Tim Sweeney
Shane Lofgren
Marta Andres-Terre
Winn Haynes
Michele Donato
Steven Schaffert
Francesco Vallania
Erika Bongen
Aurelie Tomczak
Ravi Shankar
Tej Deepak Azad
Brandon Turner
Matthew Daniel Li
Madeleine Scott
Andrew Liu
Lindsay Braviak
Caroline Braviak

SIMR students

Andrew Tam
Charles Liu
Sophia Luo
Jeffrey Cheng

PJ Utz
Alex Kuo
Peggie Cheung

Shirit Einav
Yuan Jin Tan
Elena Bekerman

Mark Davis
Cristina Tato
Helen McGuire

Nigam Shah
Katie Quinn

Clinical collaborators

Jason Andrews
Julio Croda
Benjamin Tang
Angela Rogers
Ashham Mansur
Lyle Moldawer
Hector Wang
Patrick Caroll
Gabriel Escobar
Jeffrey Freeman

NIH

BILL & MELINDA GATES foundation

VIR