Illuminating the Druggable Genome: Recent Advances

Tudor I. Oprea

University of New Mexico

UNM SCHOOL of MEDICINE | DEPARTMENT OF INTERNAL MEDICINE
Translational Informatics Division

IDG Consortium website: targetcentral.ws
IDG KMC portal: pharos.nih.gov
Harmonizome: amp.pharm.mssm.edu/Harmonizome
TIN-X app: newdrugtargets.org
Funding: U54 CA189205 (NIH)

Joint NIH NCATS Council and CAN Review Board Meeting
September 15, 2016, Bethesda, MD
Copyright © Tudor I. Oprea, 2016. All rights reserved

COMMENT

ANTHROPOLOGY Call for unity in the science of human beings **p.166**

GENETICS Reviewed: two primers on personal genomics **p.169**

POLICY Sanitation, not vaccination, is most important in Haiti **p.175**

OBITUARY Jack Oliver, key player in proof of plate tectonics, remembered **p.176**

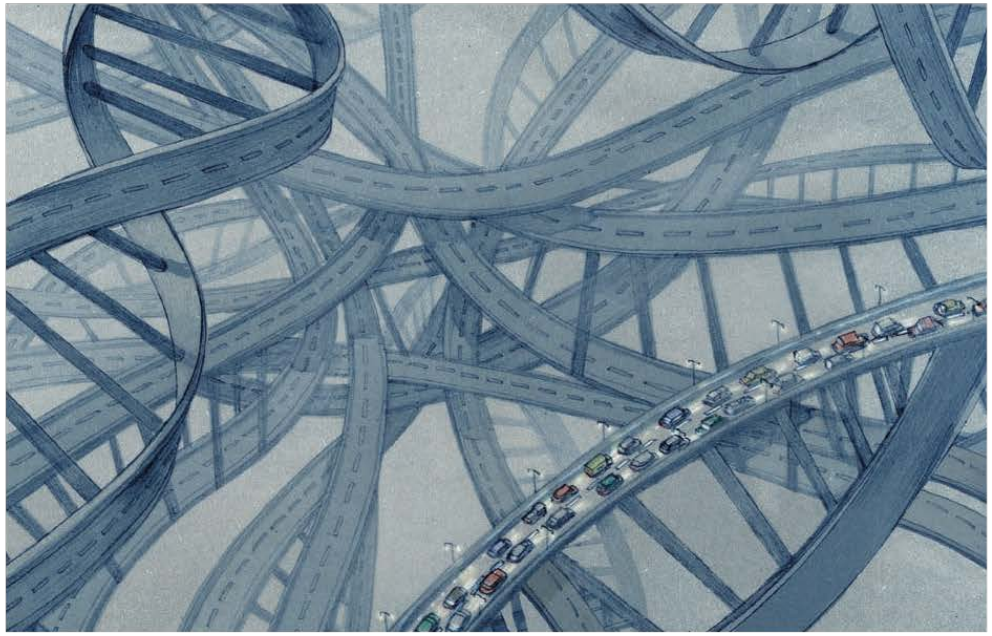


ILLUSTRATION BY JONATHAN BURTON

Too many roads not taken

Most protein research focuses on those known before the human genome was mapped. Work on the slew discovered since, urge **Aled M. Edwards** and his colleagues.

When a draft of the human genome was announced in 2000, funders, governments, industry and researchers made grand promises about how genome-based discoveries would revolutionize science. They promised that it would transform our understanding of human biology and disease, and provide new targets for drug discovery. Yet more than 75% of protein research still focuses on the 10% of proteins that were known before the genome was mapped — even though many more have been genetically linked to disease.

We performed a bibliometric analysis to assess how research activity has altered over time for three protein families that are central in disease and drug discovery: kinases, ion channels and nuclear receptors. For all three, we found very little change in the pattern of research activity — which proteins are associated with the highest number of publications — over the past 20 years[1]. Even those proteins that have been directly associated with disease remain 'hidden in plain sight', with scientists proving very reluctant to study them.

NATURE.COM
Protein mapping gains a human focus:
go.nature.com/vbqctf

Where there has been a shift in research activity, it was often spurred by the emergence of tools to study a particular protein, not by a change in the protein's perceived importance. We believe that ensuring high-quality tools are developed for all the proteins discovered may be all that is needed to drive research into the unstudied parts of the human genome — even within funding and peer-review systems that are inherently conservative.

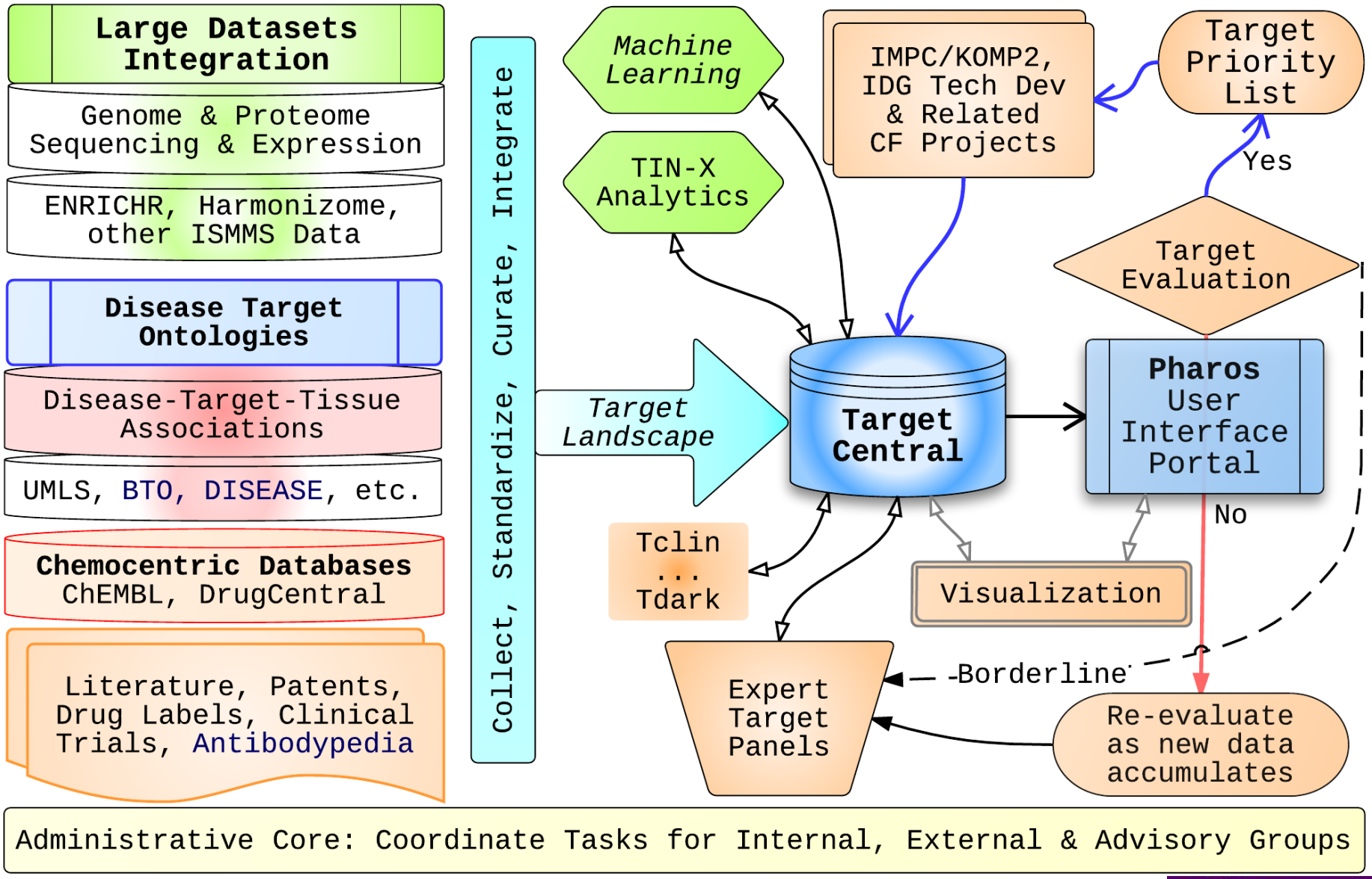
We searched for mention of every human ▶

© 2011 Macmillan Publishers Limited. All rights reserved

75% of protein research still focused on 10% genes known before human genome was mapped

AM Edwards et al, *Nature*, 2011

IDG KMC Workflow



IDG KMC portal: pharos.nih.gov

3/20/15 revision

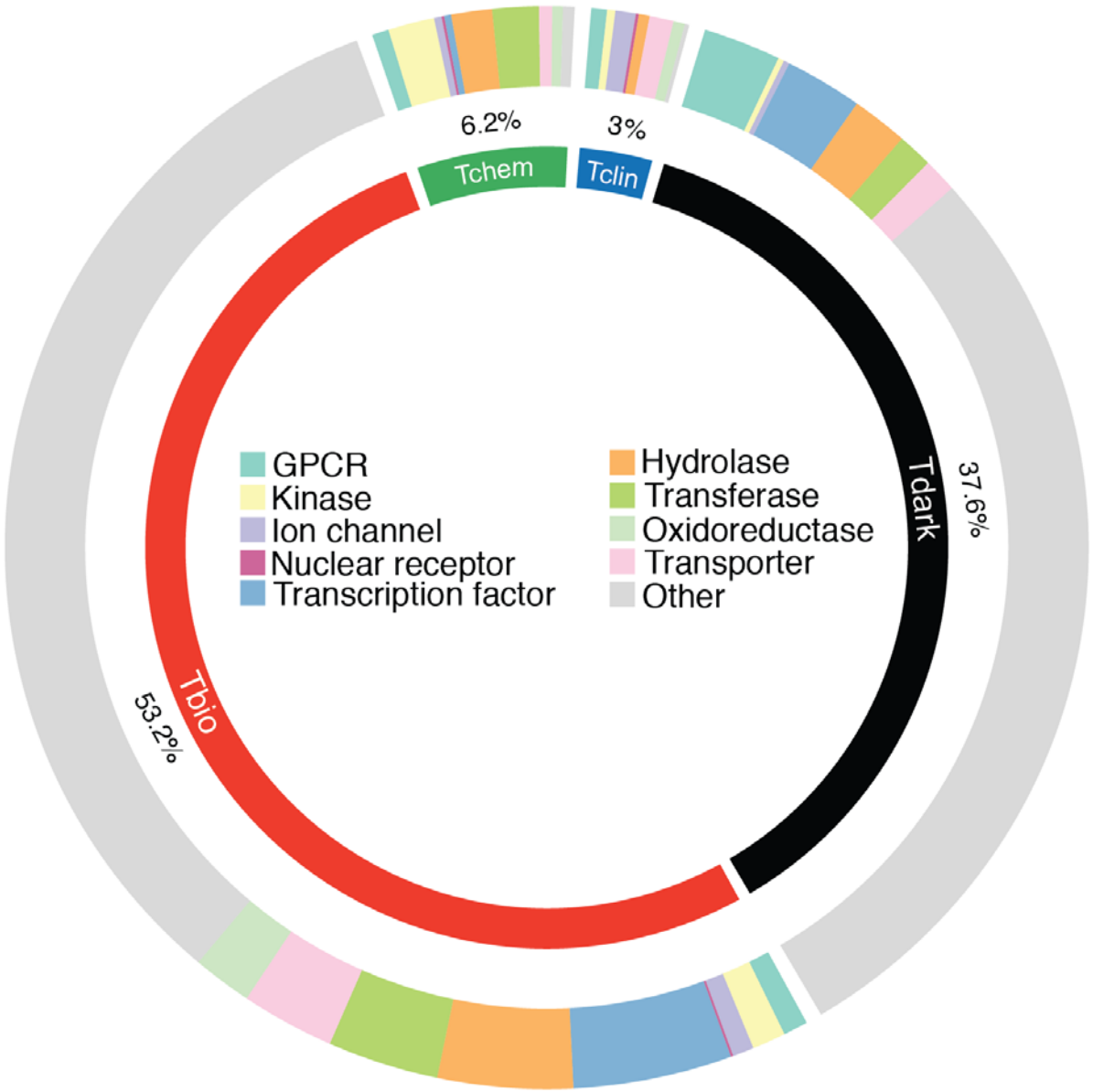
What is a Drug Target?

- **A material entity with a quantifiable mass**

  …typically a macromolecule
  - It physically interacts with the therapeutic drug;
  - It is typically native to the biological system on which the drug acts ("native" can be in a disease state)
  - the physical Drug-Target interaction causes detectable effects in living systems
- A drug target is not a pathway or other concept
- However, the clinical outcome may be due to down-stream / ripple effects
- Amenable to classification/ontology

IDG

Target Development Level



8/31/16 revision

DT Development Level 1

- **Tclin** proteins are associated with drug Mechanism of Action (MoA)

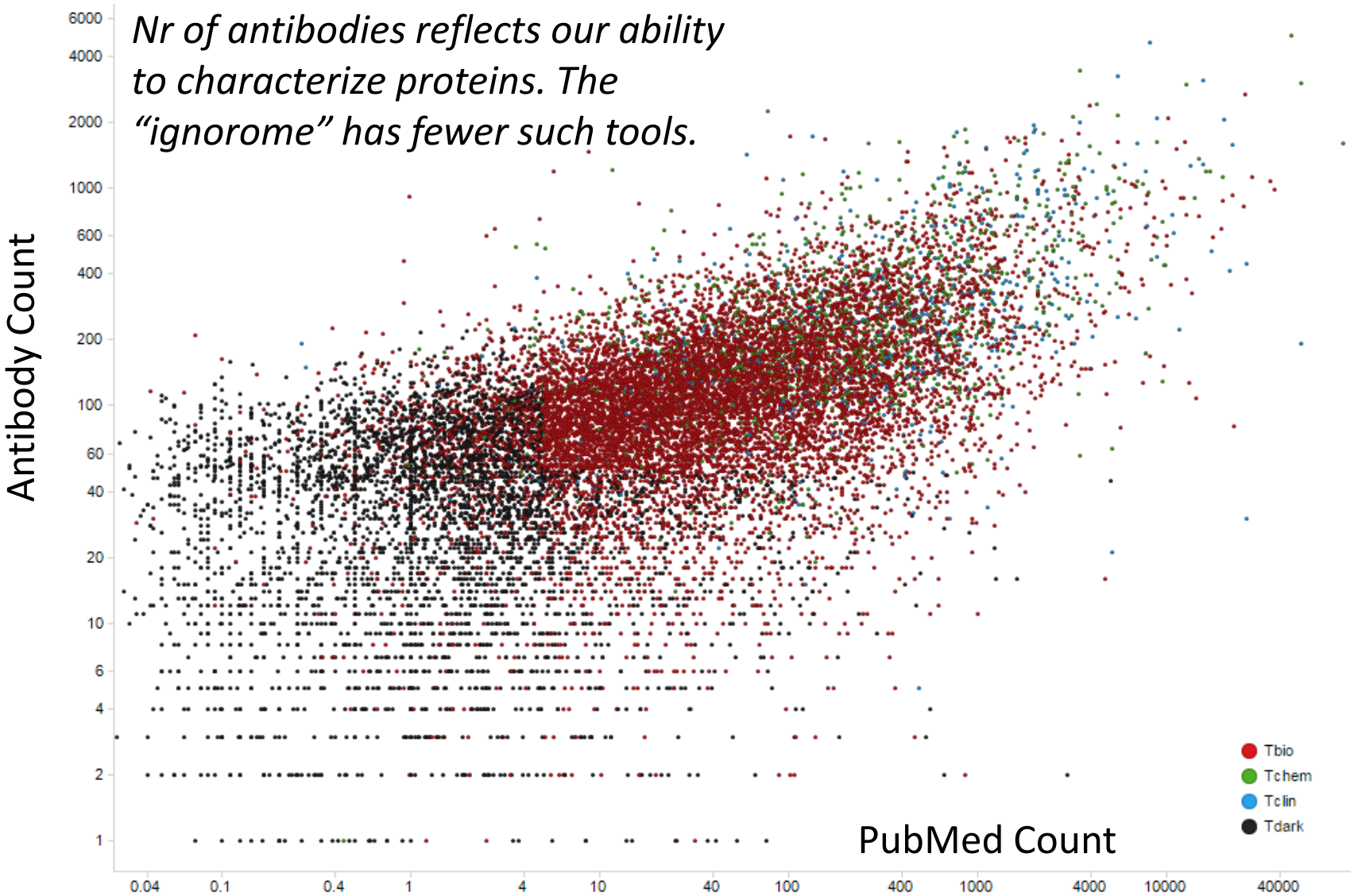
- **Tchem** proteins have bioactivitis in ChEMBL and DrugCentral, + human curation for some targets
  - Kinases: <= 30nM
  - GPCRs: <= 100nM
  - Nuclear Receptors: <= 100nM
  - Ion Channels: <= 10µM
  - Non-IDG Family Targets: <= 1µM

Note: Bioactivity cut-off values are subject to revision

DT Development Level 2

- **Tbio** proteins lack small molecule annotation cf. Tchem criteria, and satisfy one of these criteria:
  - protein is above the cutoff criteria for T_{dark}
  - protein is annotated with a GO Molecular Function or Biological Process leaf term(s) with an Experimental Evidence code
  - protein has confirmed OMIM phenotype(s)
- **Tdark** ("ignorome") have little information available, and satisfy these criteria:
  - PubMed text-mining score from Jensen Lab < 5
  - <= 3 Gene RIFs
  - <= 50 Antibodies available according to antibodypedia.com

Antibodies vs Publications

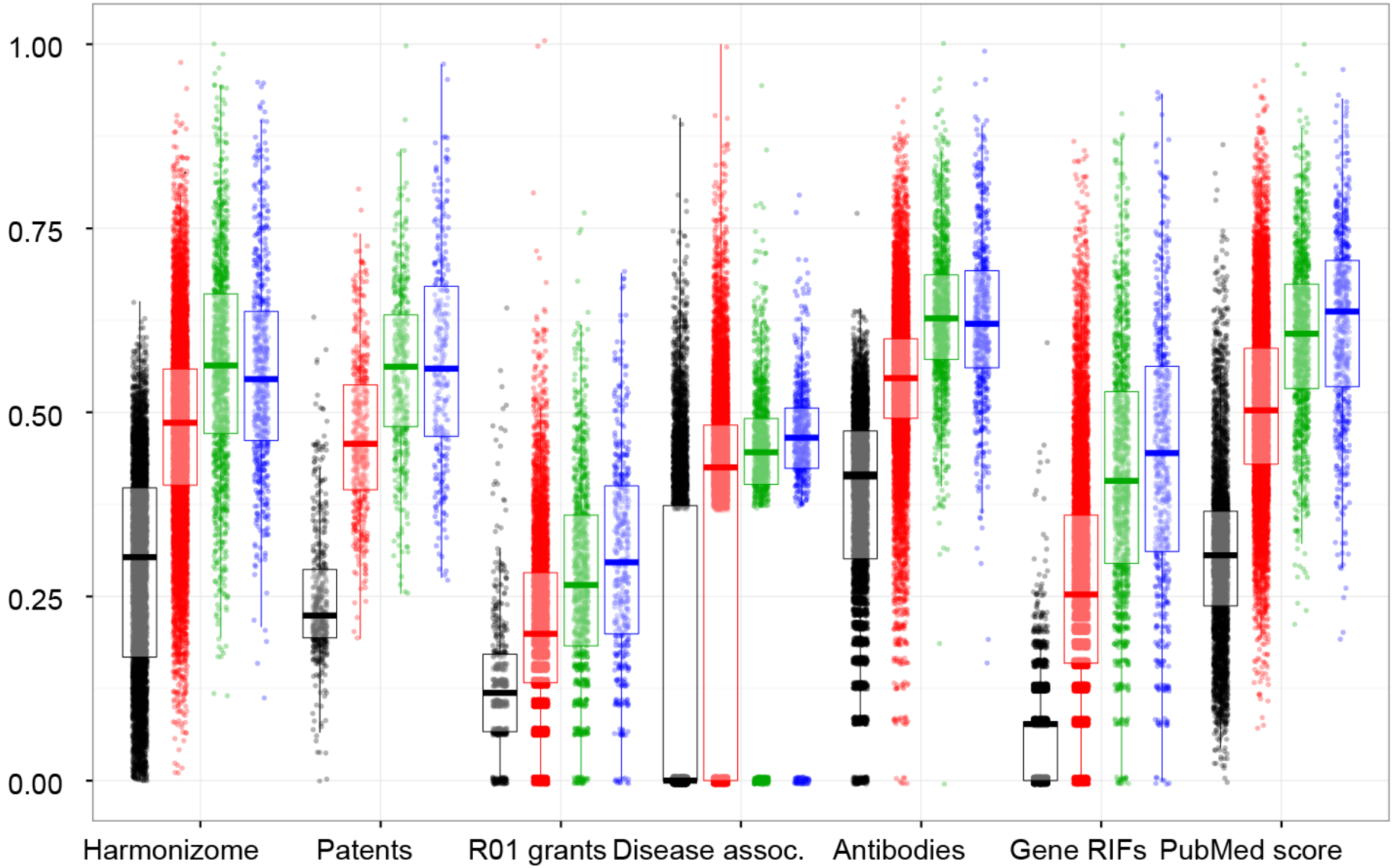


Human proteome (20,186 proteins). Spearman R = 0.68. Axes in log scale.

Antibodypedia.com

8/31/16 revision

TDL: Independent Validation

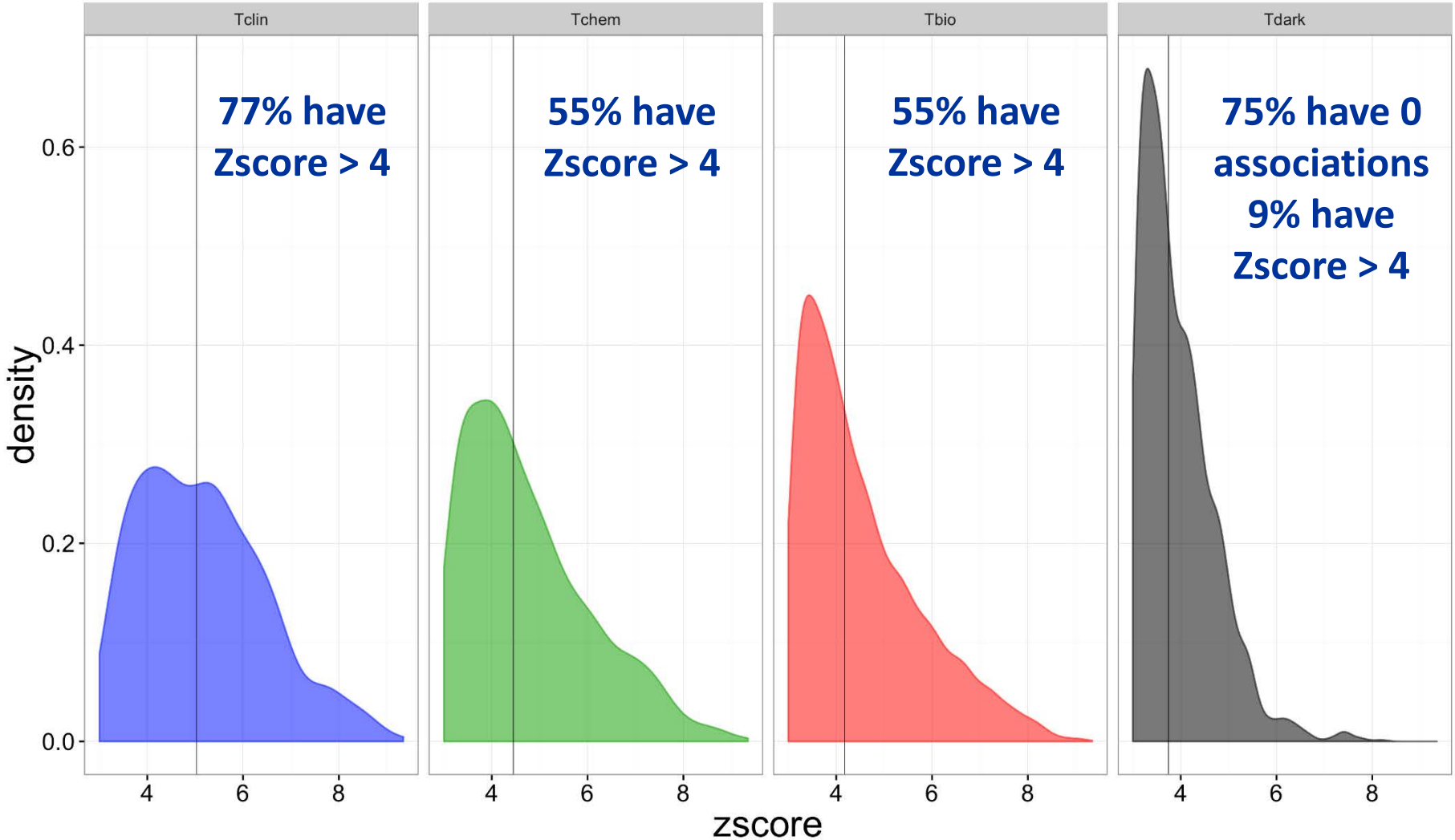


8/31/16 revision

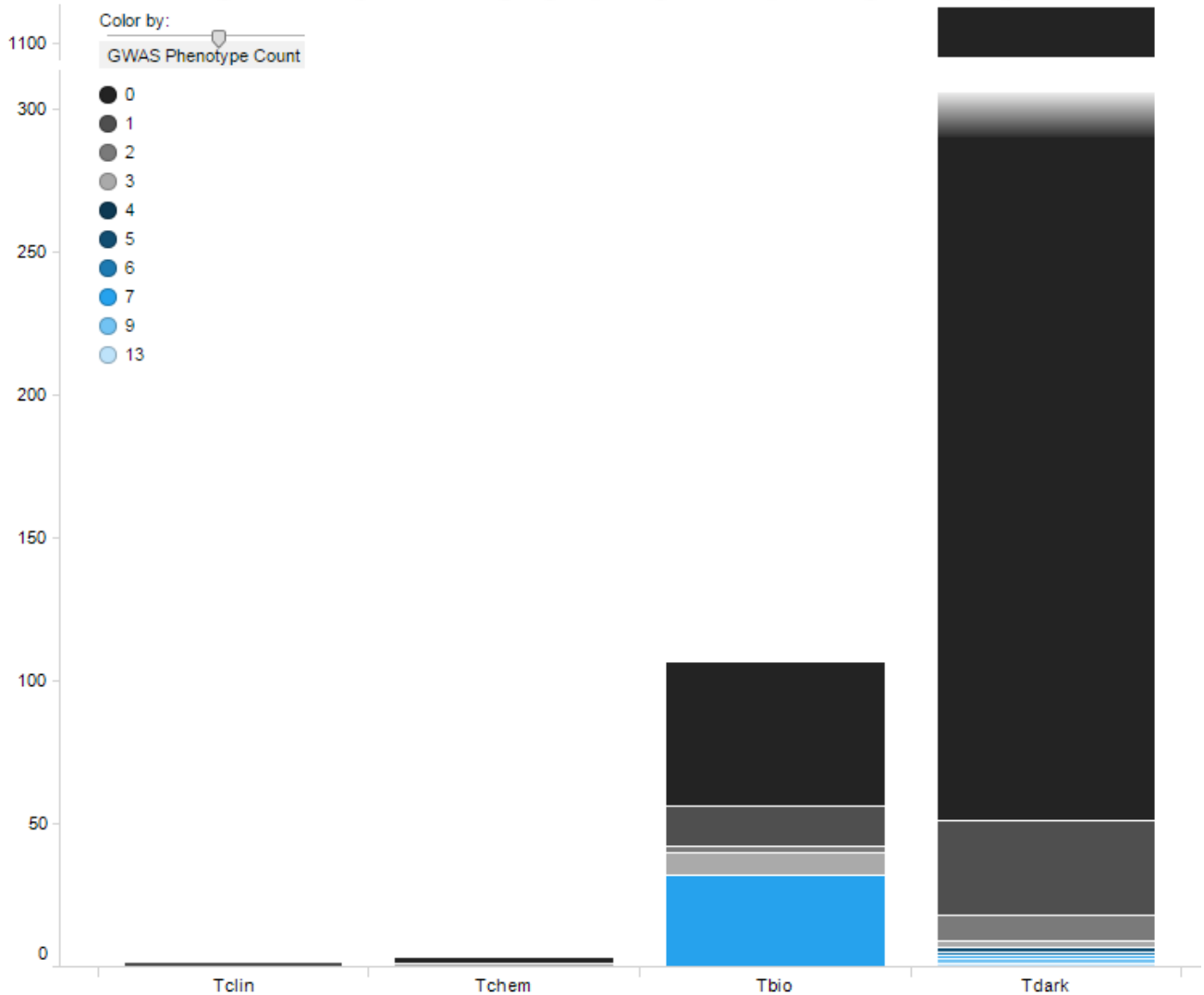
Tdark: Searching for the Light

- Avi Ma'ayan's *Harmonizome* examines experimental information density per protein, processed from 70 genomic datasets. ***Tdark*** *proteins have less data compared to the other 3 categories*.
- "Patents" examines the distribution of text-mined granted patents per protein from SureChEMBL. ***Tdark*** *proteins are subject to a significantly lower number of patents.*
- "R01 grants" examines the distribution of text-mined R01 grant counts per protein, using NIH RePORTER data. *Most* ***Tdark*** *proteins are not funded via the R01 mechanism*.
- "Disease associations" examines the distribution of text-mined disease associations per protein. *~90% of* ***Tdark*** *proteins have a score of zero.*
- This uneven distribution is reproduced across multiple instances, e.g., from a different literature corpus (patents), and when using experimental data (Harmonizome). **Thus, there appears to be a Knowledge Deficit concerning "dark" proteins.**

Target Disease Associations



The Darkest of the Dark



Presence (color) or absence (black) of GWA studies for TDL (1,251 human proteins), for which there <u>is no Tissue Expression data</u> (aggregated from multiple sources). Of these proteins lacking GWAS/expression data, 1,090 (5.4%) are Tdark.

9/14/16 revision

Take Home Message 1

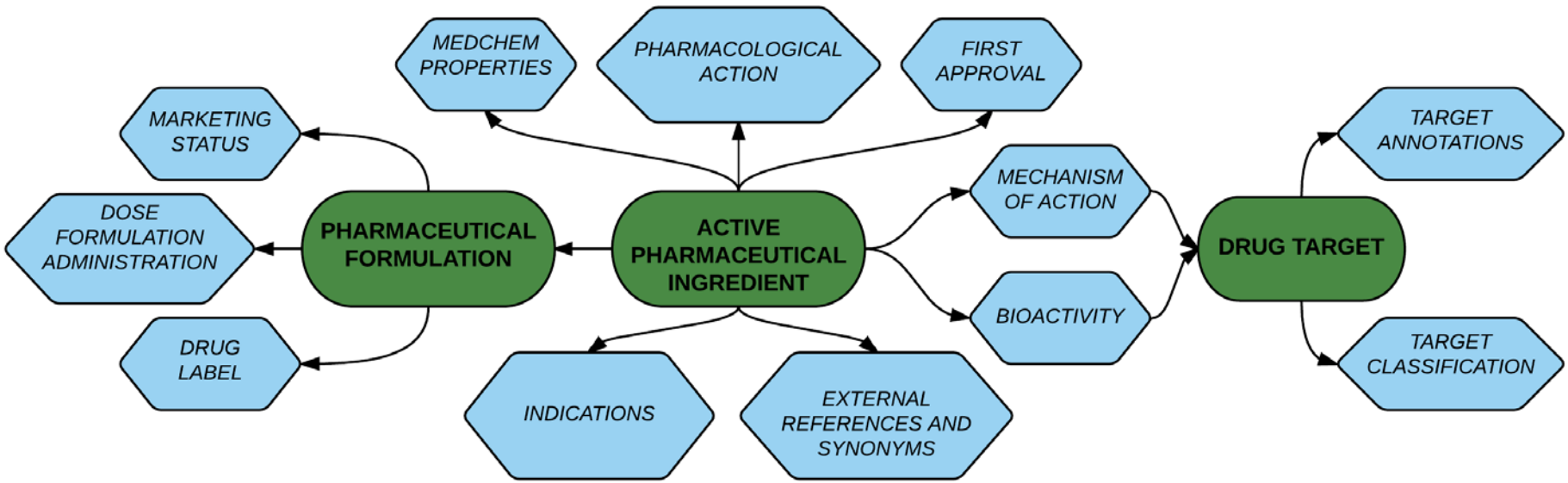
There is a Knowledge Deficit

over 37% of the proteins remain
poorly described (**Tdark**)

~10% of the Proteome (**Tclin** & **Tchem**)
can be targeted by small molecules

These observations are supported by
different methods across multiple datasets

DrugCentral Data Structure



- Initially to answer "how many drugs are out there"…
- Mapped products (what patients and docs call "drugs") onto active ingredients (what scientists call "drugs")
- Also wanted to know how many drug targets there are…….

Oleg Ursu et al., Nucl Acids Res, submitted

8/21/16 revision

DrugCentral Stats: APIs & Targets

	Entities (Annotated APIs)
Active pharmaceutical ingredients	4,444
FDA drugs	2,021
Drugs approved outside US	2,423
Small molecules	3,799
Salts and inorganic molecules	112
Biologics and peptides	239
Other drugs	294
Parent molecules	199 (308)
Drug efficacy targets	837 (1,689)
Human protein targets	600 (1,387)
Infectious agents targets	194 (221)
Metabolites & biopolymers	43 (89)
Protein-drug crystal complex (PDB)	48 (82)
Drug-protein crystal complex (PDB)	1,452 (283)

Oleg Ursu et al., Nucl Acids Res, submitted

8/21/16 revision

Drug/Disease: A Small (Molecule) World

Type	WHO ATC codes	Indications	Contra-indications	Off-label indications
Unique Concepts	4,195	2,224	1,458	847
Unique APIs	2,941	2,247	1,376	646

We introduced controlled vocabularies and identifiers in DrugCentral:

- Xxx disease concepts (331 off-label) addressed by APIs
- *Yyy disease concepts are contra-indications only…*

A Comprehensive Map of Molecular Drug Targets

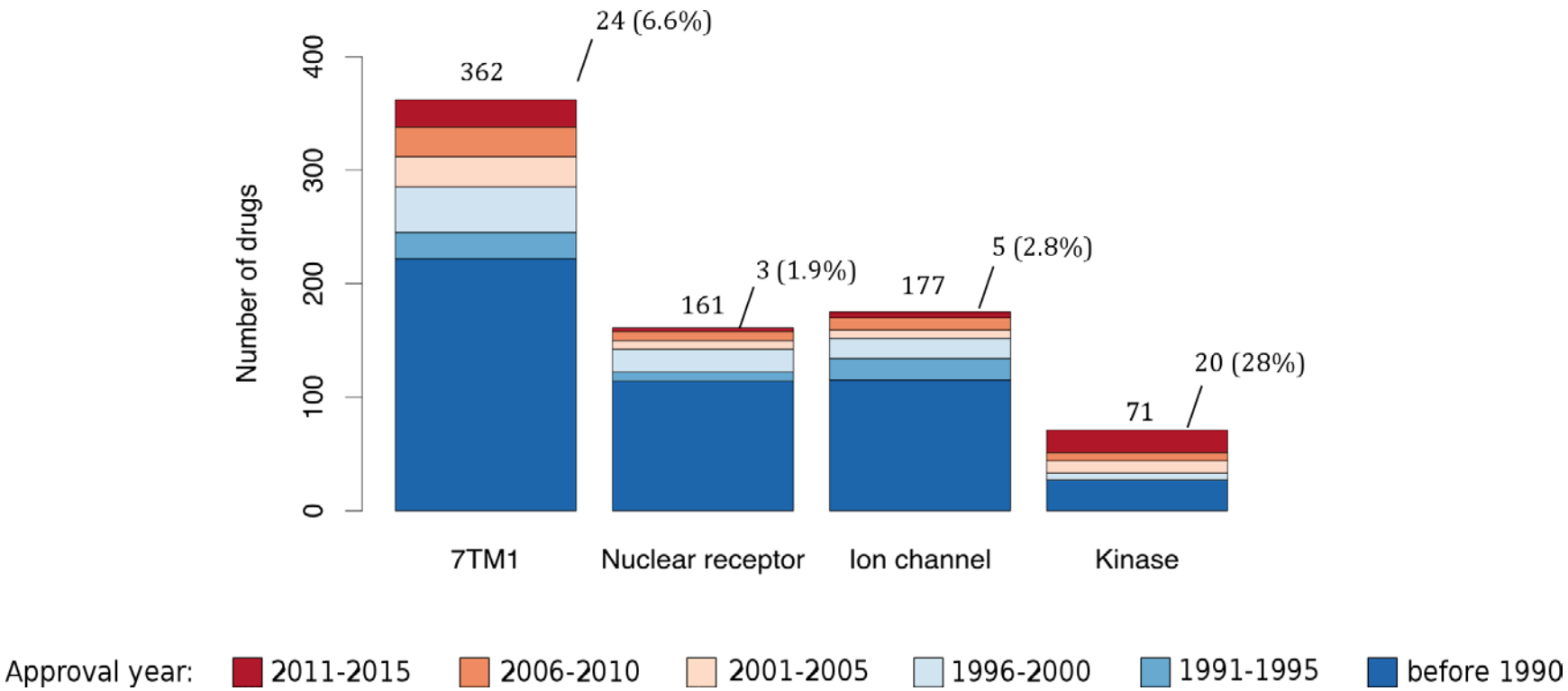
- We systematically compiled efficacy target information using drug label information and primary scientific literature. It is rather challenging to assign efficacy targets, especially to non-selective agents, particularly for anti-infective and anticancer drugs.
- Drugs targeting protein kinases have dramatically increased over the past 5 years, compared to e.g., the lack of innovation for nuclear receptor-targeted drugs over the same period.
- We analyzed Drugs and Target Classes according to their therapeutic area (ATC Codes). Most progress has been made in oncology, antivirals, immunosuppressants and diabetes.
- Small molecules targeting GPCRs are used in almost all therapeutic areas, while kinases are currently drug targets exclusively in the antineoplastic and immunomodulatory category.

R. Santos, O Ursu et al., *Nat. Rev Drug Discov,* 2016, accepted

6/03/16 revision

IDG

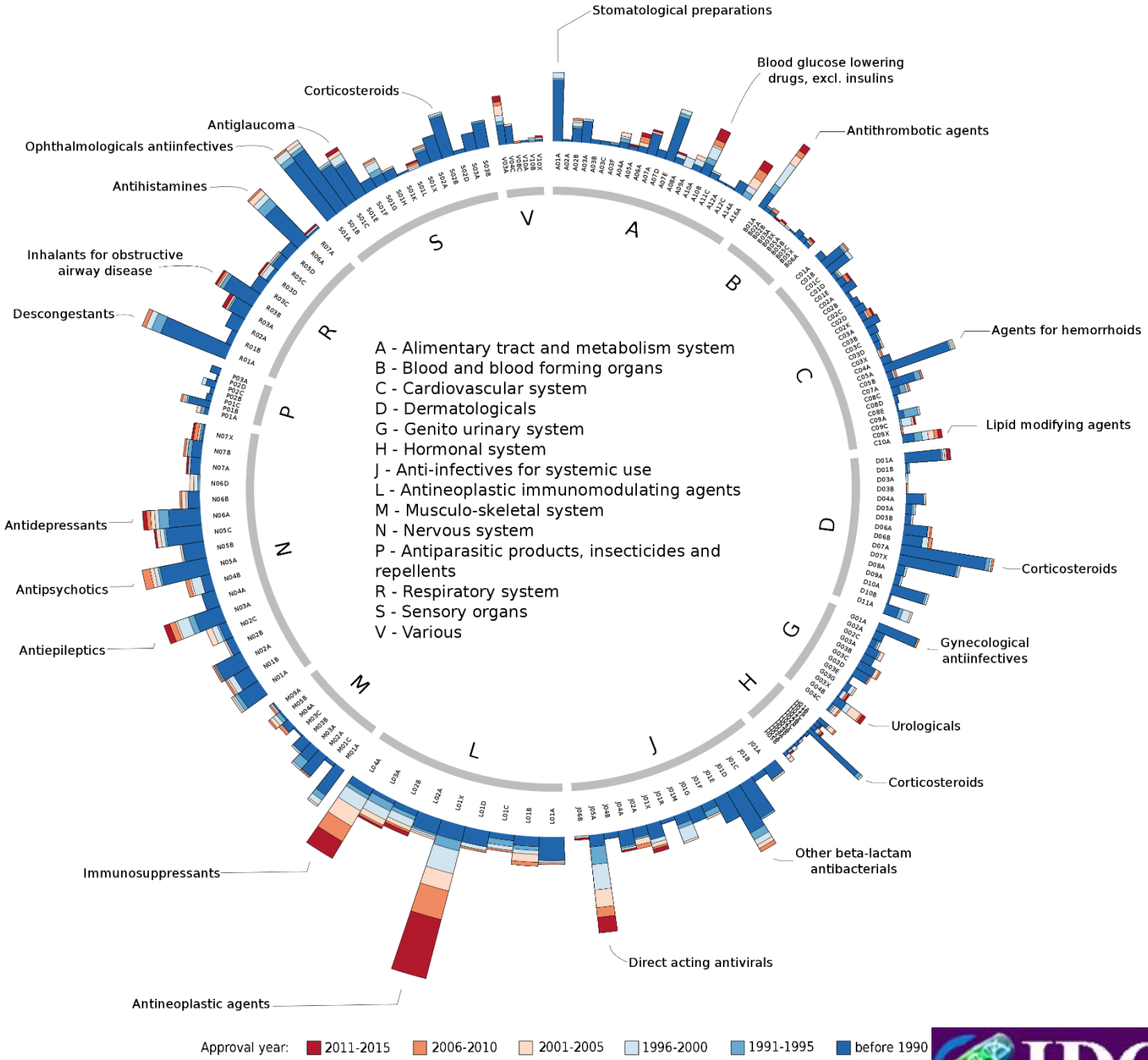
Innovation Patterns per Privileged Family Classes



R. Santos, O Ursu et al., *Nat. Rev Drug Discov,* 2016, accepted

6/03/16 revision

Innovation Patterns per Therapeutic Area



R. Santos, O Ursu et al., *Nat. Rev Drug Discov,* 2016, accepted

6/03/16 revision

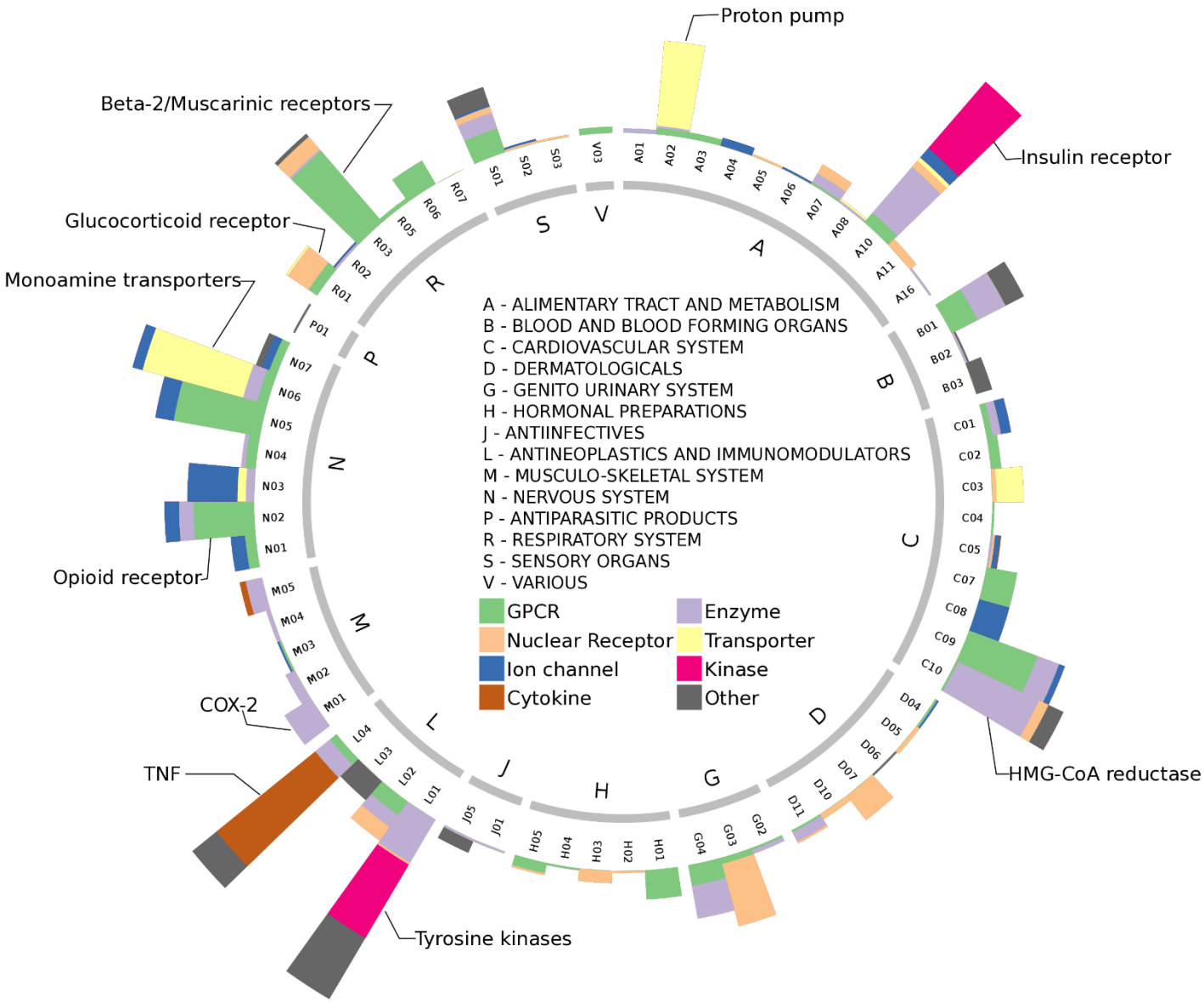
A Target-Centric Analysis of Global Drug Sales Data

- Aggregated sales from 75 countries, including Europe, North America and Japan, over a five year period (2011-2015), collected by IMS Health, were interrogated from a drug target (**Tclin**) perspective.
- Data were normalized by mapping revenue for pharmaceutical products to Active Pharmaceutical Ingredients using DrugCentral, corrected by number of APIs per product and by the number of efficacy (**Tclin**) targets per API.
- We analyzed all targets according to ATC therapeutic area Codes for the corresponding drugs.
- Sales by Level 2 ATC code levels and by target class were normalized to percent values in a circular histogram.
- These ATC chapters show that the top earning mode-of-action drug categories are “antineoplastics and immunomodulators”, followed by the “nervous system” chapter.

T Oprea et al., *Nat. Rev Drug Discov,* in preparation

8/22/16 revision

Financial Activity per Therapeutic Area

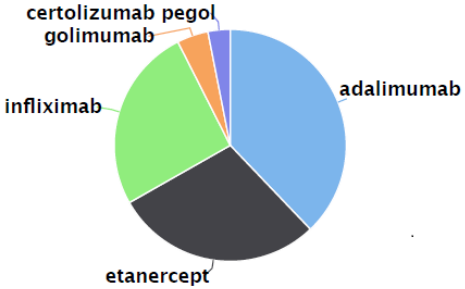


T Oprea et al., *Nat. Rev Drug Discov,* in preparation

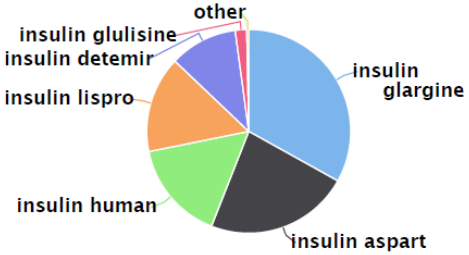
8/22/16 revision

IMS Health: Revenue Per Drug Target

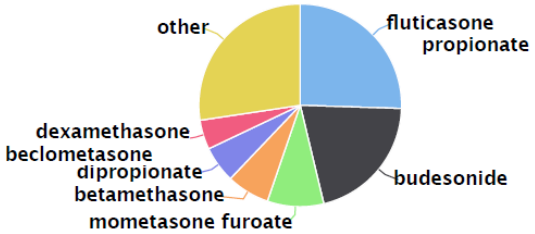
TNF - $163.39 bil



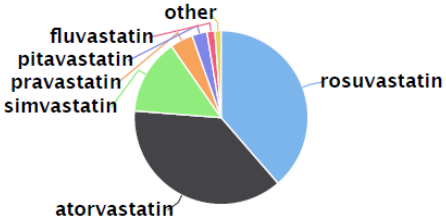
INSR - $143.55 bil



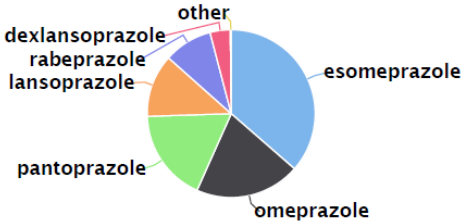
NR3C1 - $142.75 bil



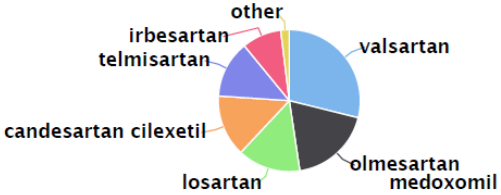
HMGCR - $122.55 bil



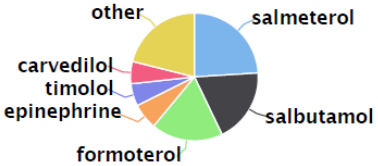
H+/K+ ATPase - $118.16 bil



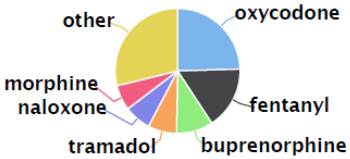
AGTR1 - $99.98 bil



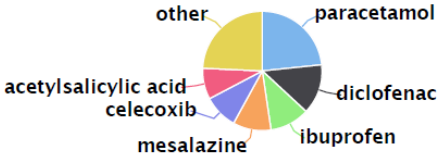
ADRB2 - $90.02 bil



OPRM1 - $87.97 bil



PTGS2 - $84.04 bil



<u>Most lucrative targets between 2011 and 2015:</u> the TNFalpha receptor; the insulin R; the glucocorticoid R; HMG-CoA-reductase, the gastric proton pump; the angiotensin R1; adrenergic β_2-R; μ-opioid-R; and cyclooxygenase-2. *Based on global drug sales data (75 countries)*

T Oprea et al., *Nat. Rev Drug Discov,* in preparation

8/22/16 revision

Take Home Message 2

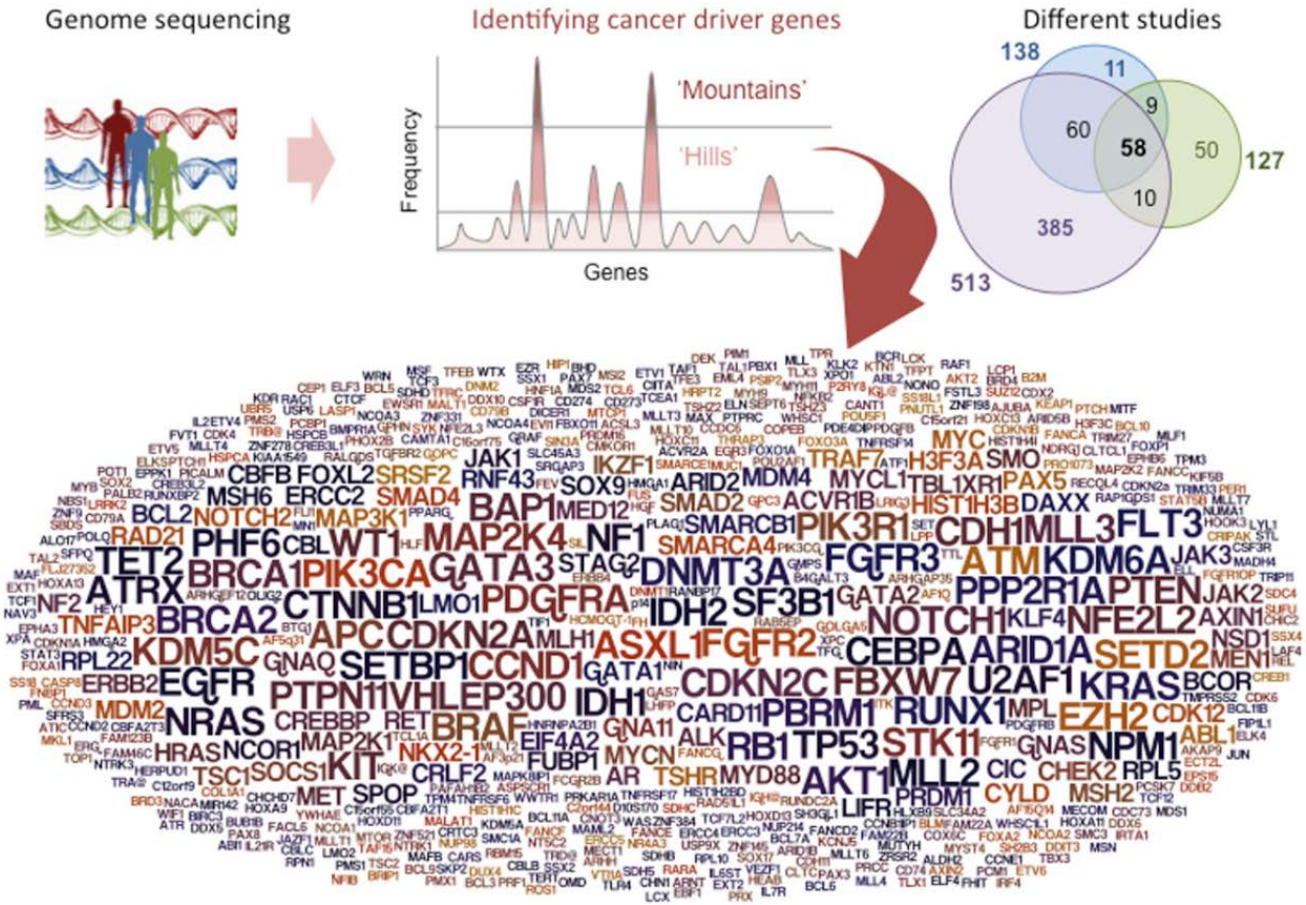
We have therapeutic agents for ~15% of human diseases *)

…the top 5 best-earning targets are not GPCRs

There are many new therapeutic opportunities

*) disease-ontology.org catalogs ~9,000 disease concepts. This lacks ~6,000 rare diseases. Thus we estimate ~15,000 disease concepts, of which ~2500 have therapeutic agents

Cancer Driver Genes: How Many?

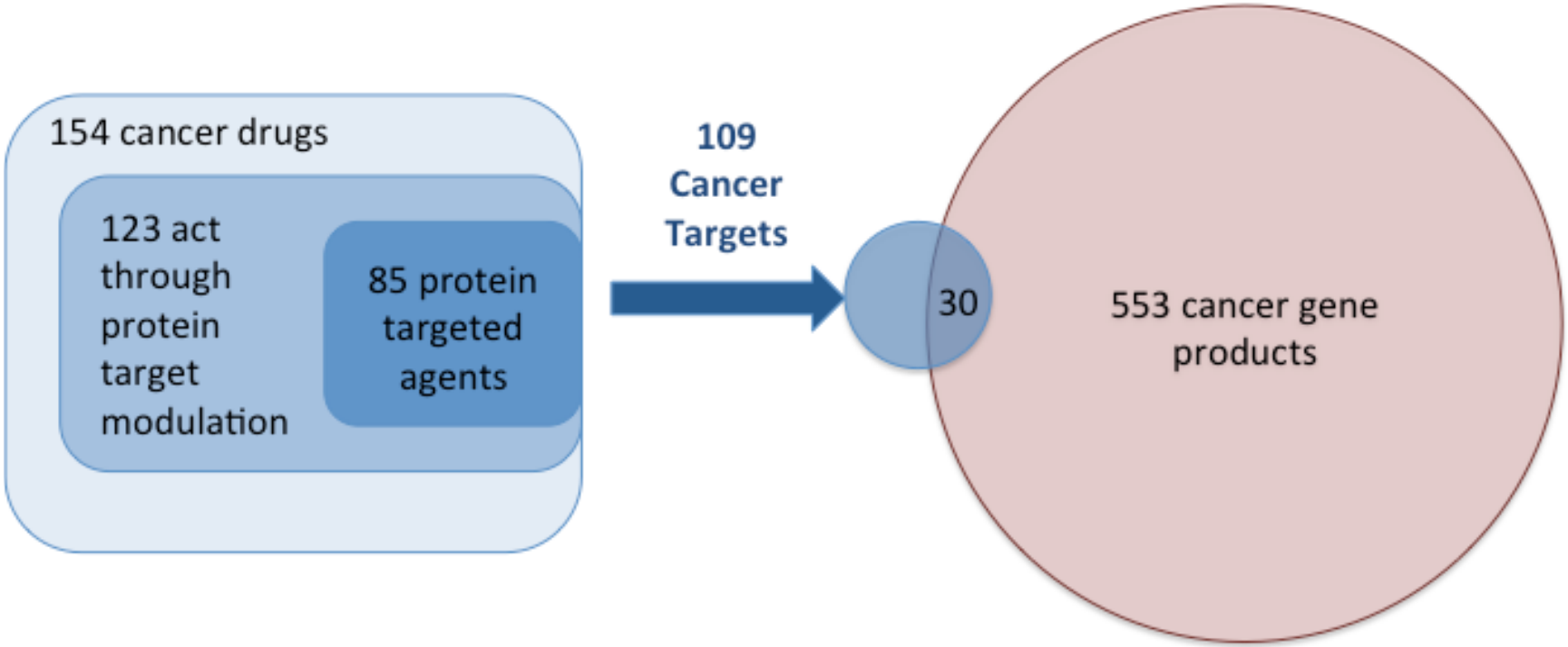


- TCGA's pan-cancer analysis: 127 significantly mutated genes across 12 tumor types (out of 3281 genomes), which is similar to the ~140 genes identified from 3,284 cancer genomes.
- The COSMIC Cancer Gene Census contains 595 genes (513 in the 2013 figure, above)
- **Only 58 genes that are common among the three** (67 genes between the 2 pan-cancer studies)

Workman, P. & Al-Lazikani, B. *Nat. Rev. Drug Discov.* **12,** 889—890 (2013)

6/03/16 revision

Overlap of Cancer Drug Targets with Cancer Drivers



Workman, P. & Al-Lazikani, B. Drugging cancer genomes. *Nat. Rev. Drug Discov.* **12,** 889—890 (2013)

R. Santos, O Ursu et al., *Nat. Rev Drug Discov,* 2016, accepted

6/03/16 revision

We Track Expression Data

GTEx Portal

THE HUMAN PROTEIN ATLAS

HUMAN PROTEOME MAP

We already process these resources in TCRD

These resources would have to be processed for UNMCCC

canSAR

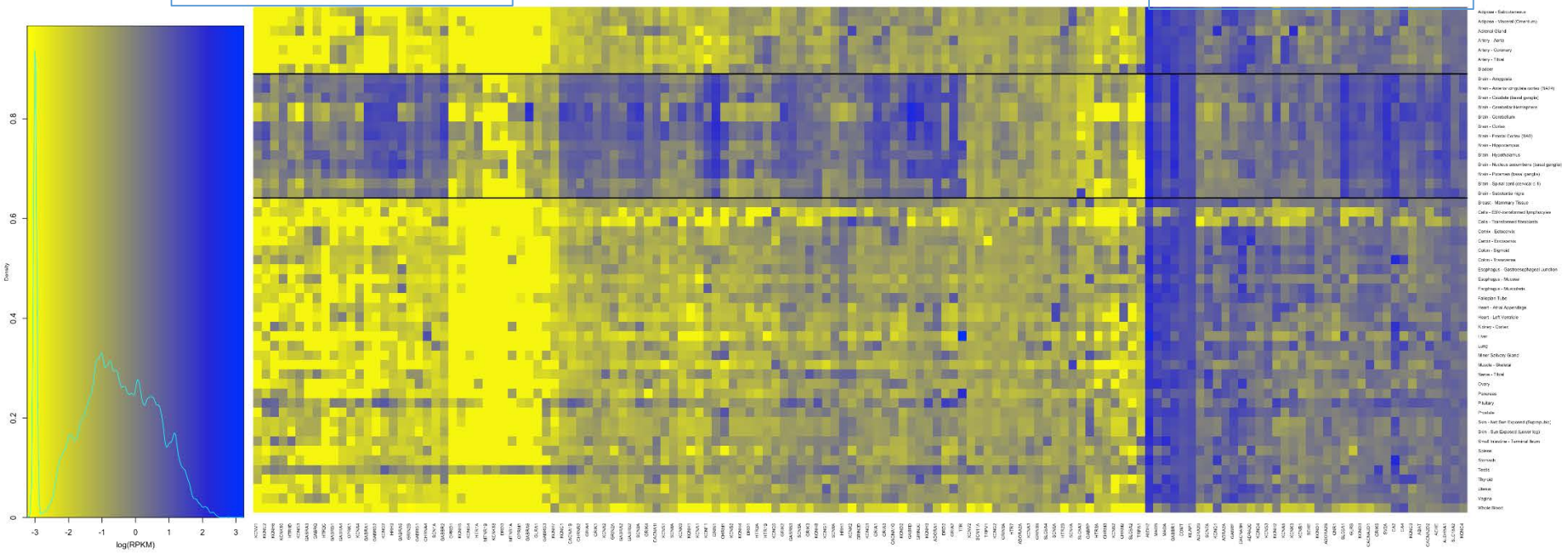
COSMIC
Catalogue of somatic mutations in cancer

nature | TCGA pan-cancer analysis

GTEx Expression for CNS drug targets

~ 25% higher specificity for brain tissues, HRH3, DRD3, HTR1A, MTNR1B

~ 27% not specific for brain tissues, MAOA, MAOB, COMT



It's possible that some drugs localize preferentially in the brain.
But it's also possible that some expression data are inconsistent.

GTEx Portal

THE HUMAN PROTEIN ATLAS

HUMAN PROT

Challenge

nature | TCGA pan-cancer analysis

COSMIC
Catalogue of somatic mutations in cancer

- Large-scale expression data are rarely in agreement (even with peer-reviewed literature). This is our biggest challenge.
- COSMIC vs. TCGA vs. others – agreement is *partial*
- There is no mathematical way to establish what is the "truth". Thus, we have no programmatic way to assign higher levels of confidence to one source over another.
  - Math & stats can show trends, and where data are consistent…
  - Analytics & modeling can help us look for inconsistencies, but only based on existing evidence

FAERS processing (Aug, 2016)

- Removed duplicated reports (last update kept)
- Added missing APIs mappings – additional information based on product names was added to openFDA mappings
- We removed all reports with no product – APIs mappings

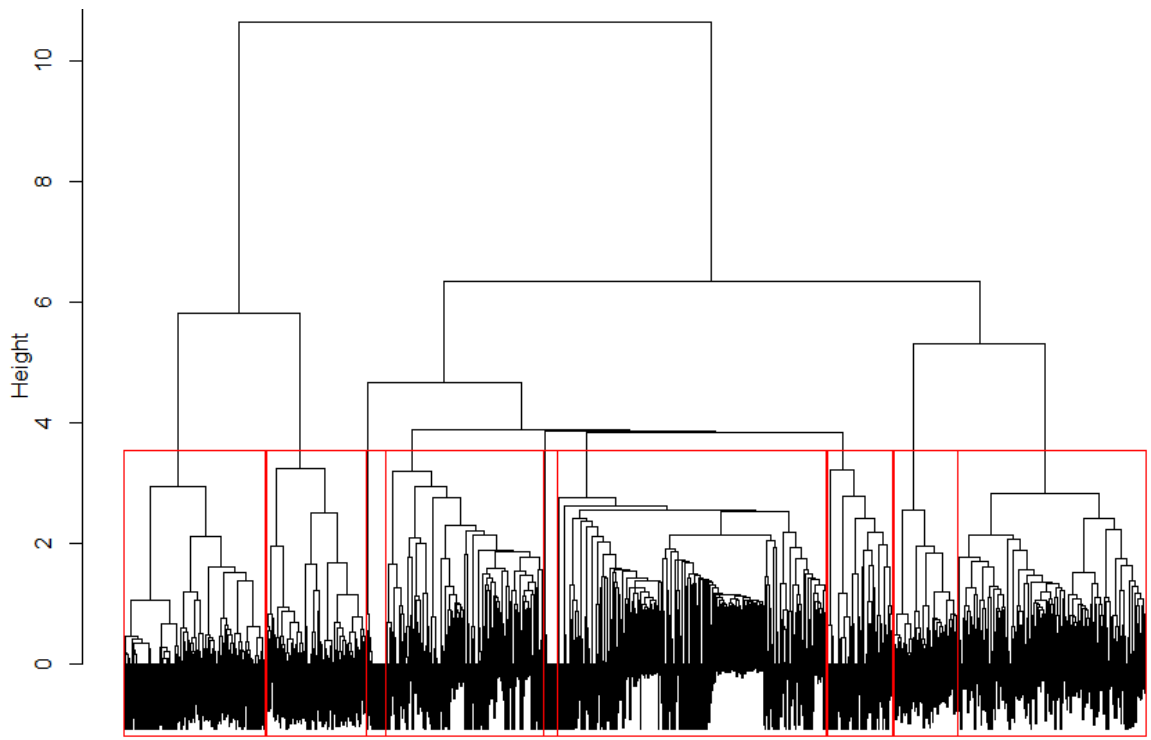
Reports	Drugs (unique APIs)	MedDRA terms	Reports with PRR* >= 2
6,534,096	3,193	19,238	944,471

PRR – proportional reporting ratio

- FAERS Total: 86,014,009 API – AE pairs
- Filtering for Drug *suspected* to cause AE:

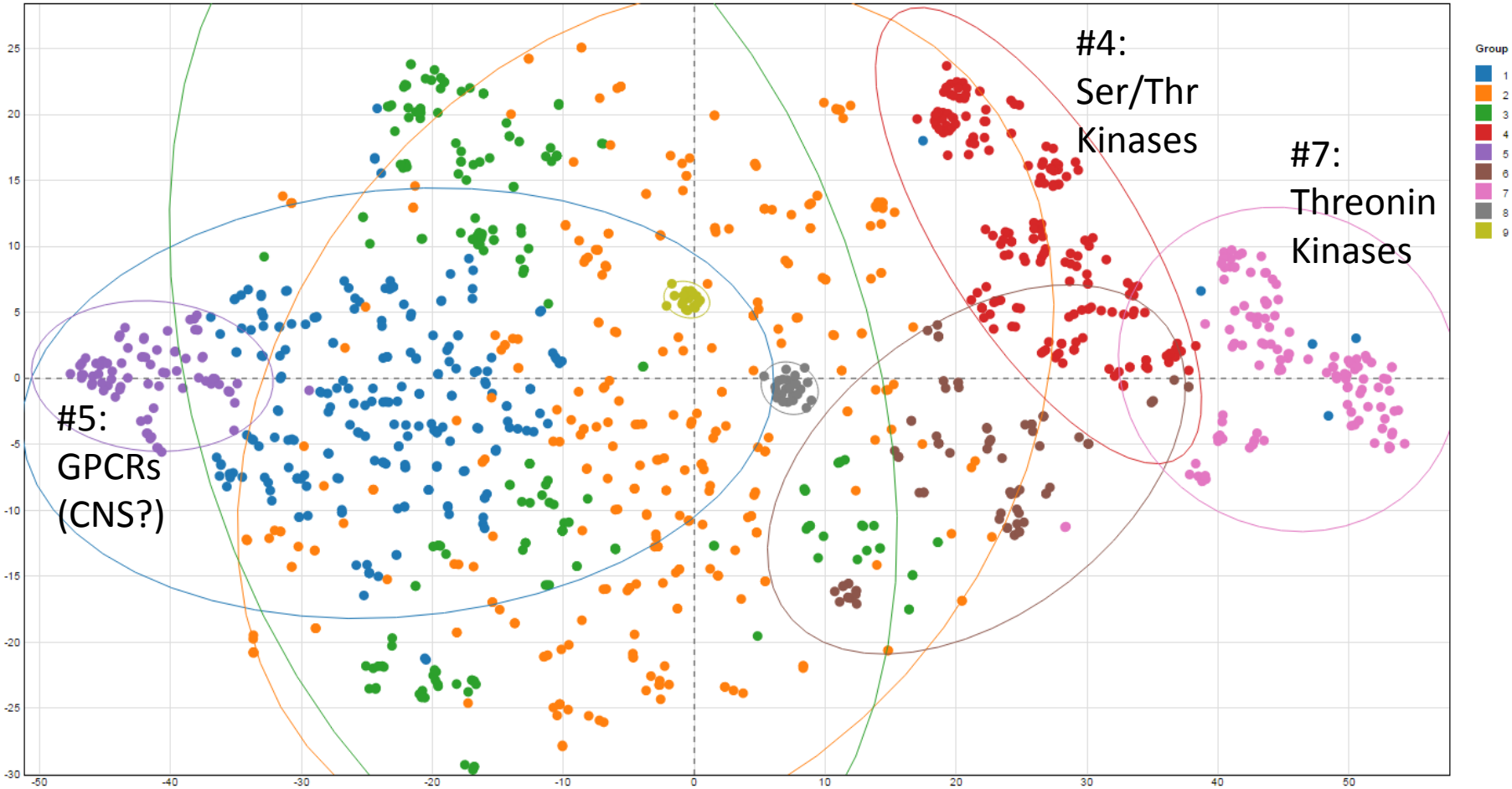
36,283,400 API – AE pairs

Drug-AE-Target Relationships



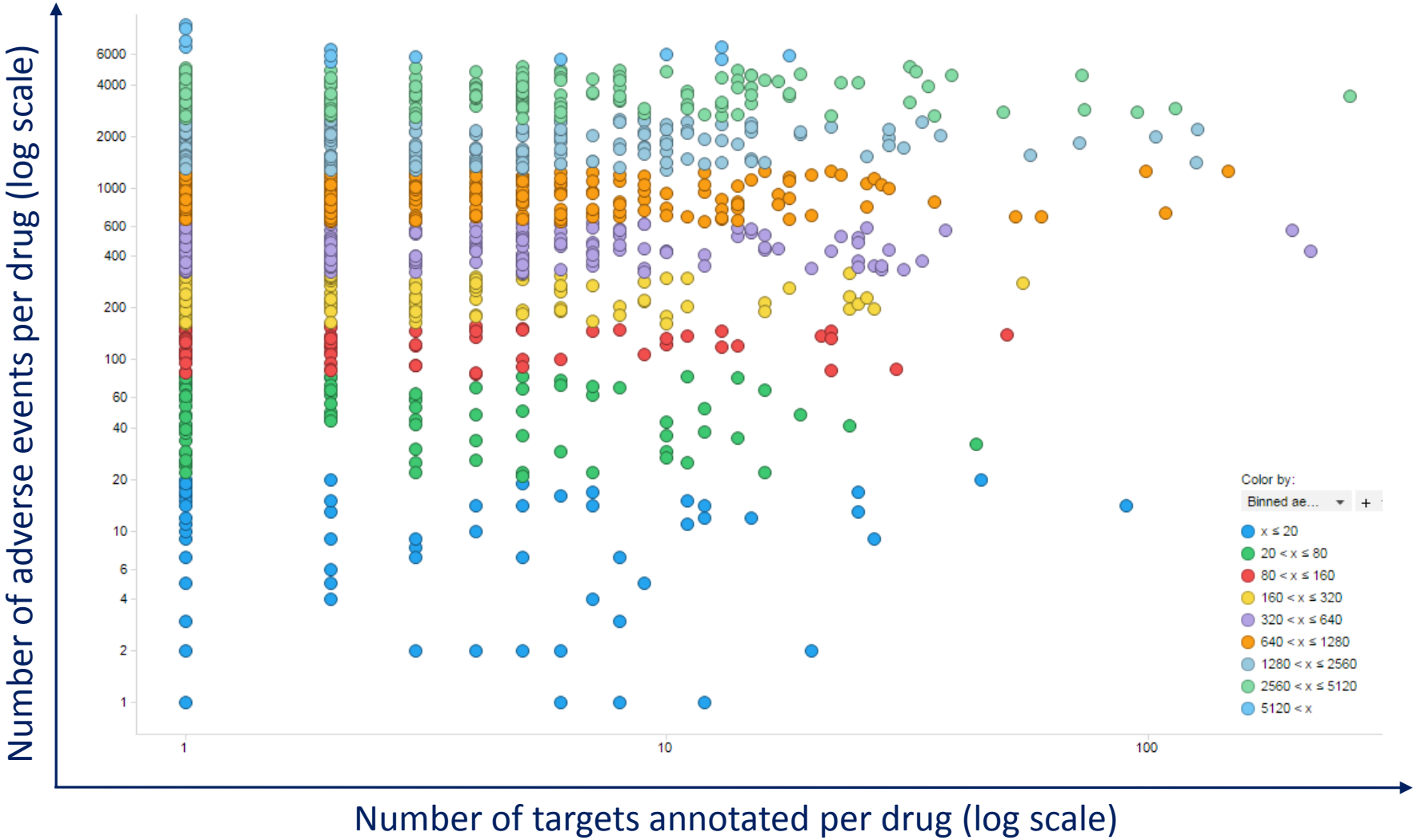
- Hierarchical (Ward) clustering was applied to the dis-similarity matrix computed from 17,848 AEs recorded for the 3,193 APIs, which in turn bind to 1,247 targets [these are mapped into **Tclin** & **Tchem**]
- The 17848x1247 dis-similarity matrix was projected onto 2D using Stochastic Neighbor Embedding

Targets Clustered in AE Space



- Nine clusters representing Target relationships derived from the 17,848 AE-Drug pairs and the 3,193 Drug – 1,247 Targets matrix

AE vs Target Annotations



- How many AEs per drug vs. known targets per drug?
- Short answer: There is no relationship

Oleg Ursu, C Bologa and T Oprea, unpublished

8/21/16 revision

Take Home Message 3

There is a great deal of uncertainty with respect to experimental data

FAERS data may provide
an independent angle for target prioritization
and shortcuts to *druggable* targets

IDG KMC Team

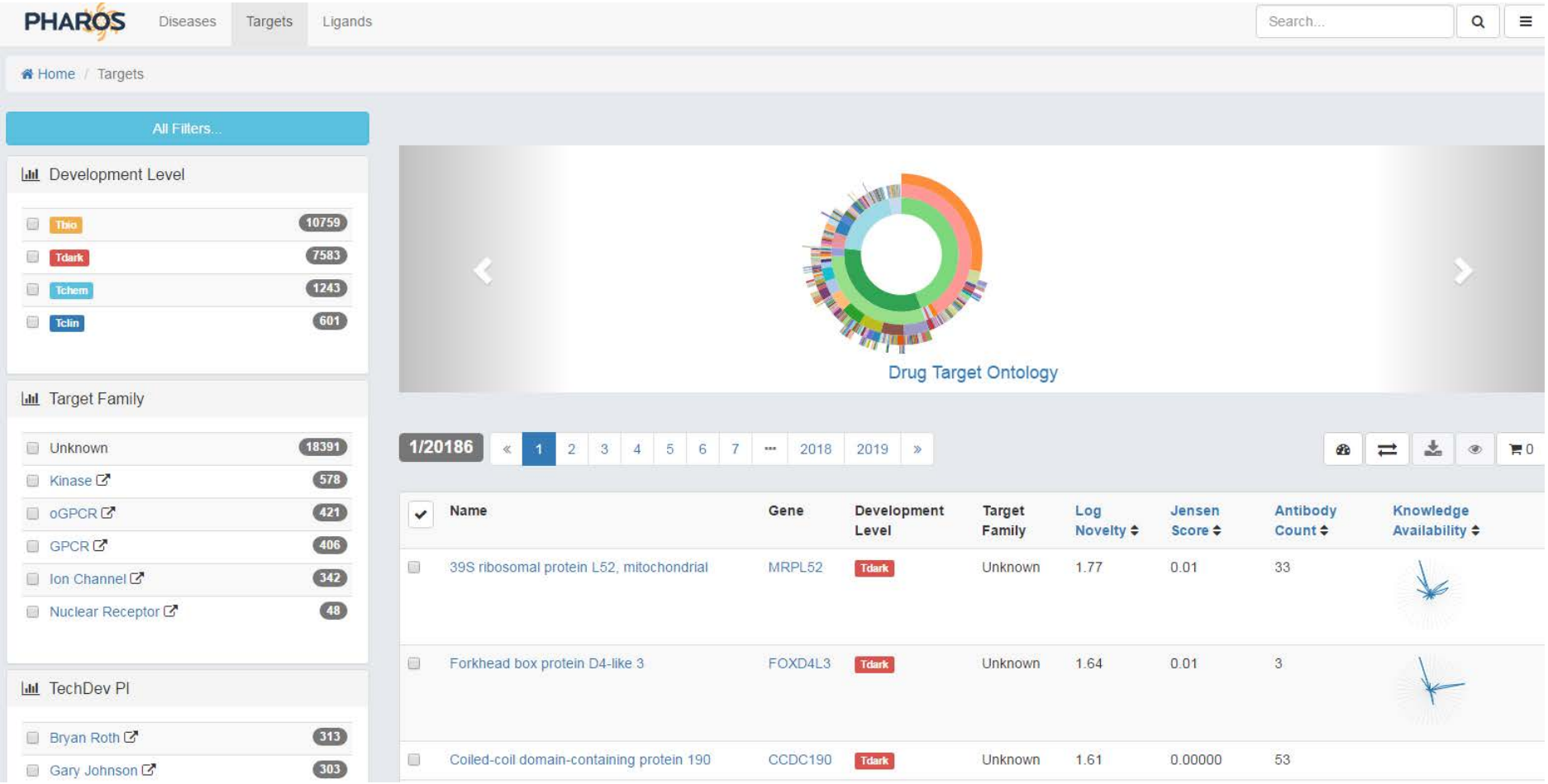
- University of New Mexico: Cristian Bologa, Jayme Holmes, Steve L. Mathias, Tudor Oprea, Larry Sklar, Oleg Ursu, Anna Waller, Jeremy J Yang, Gergely Zahoranszky-Kohalmi[1)]
- Novo Nordisk Foundation Center for Protein Research: Lars Juhl Jensen, Søren Brunak
- Icahn School of Medicine at Mount Sinai: Avi Ma'ayan, Joel Dudley, Andrew Rouillard[2)]
- EMBL-EBI – European Bioinformatics Institute (ChEMBL team): John Overington[3)], Anne Hersey, Anna Gaulton, Anneli Karlson[3)], George Papadatos[2)]
- NIH-NCATS: Rajarshi Guha, Ajit Jadhav, Dac-Trung Nguyen, Anton Simeonov, Noel Southall
- University of Miami: Stephan Schürer, Dusica Vidovic
- with help from IMS Health: Allen Campbell, Christian Reich

[1)] NIH-NCATS; [2)] GSK; [3)] Stratified Medical

IDG Illuminating the Druggable Genome Consortium

The IDG Consortium is an NIH network of Knowledge Management Centers that collect & integrate data from across various resources to aid in prioritizing *illumination of understudied protein targets*, and connecting these with Technology Development Centers that bring forth new technologies and tool sets to shed light on to these targets.

Pharos: The IDG KMC Portal



Watch the 2-minute YouTube video here:
https://pharos.nih.gov/idg/index#

9/14/16 revision

Illuminating the Druggable Genome

~25 Million Papers
6.6 million Patents
>100 Million EHRs (RUF)

20,200 Proteins

Seeking New Knowledge

~15,000 Diseases

~4,400 Drugs

8/21/16 revision

IDG