Deep learning to predict the impact of rare variation in drug metabolism genes

Russ B Altman, MD, PhD

Bioengineering, Genetics, Medicine, Biomedical Data Science

Stanford University

March 2021

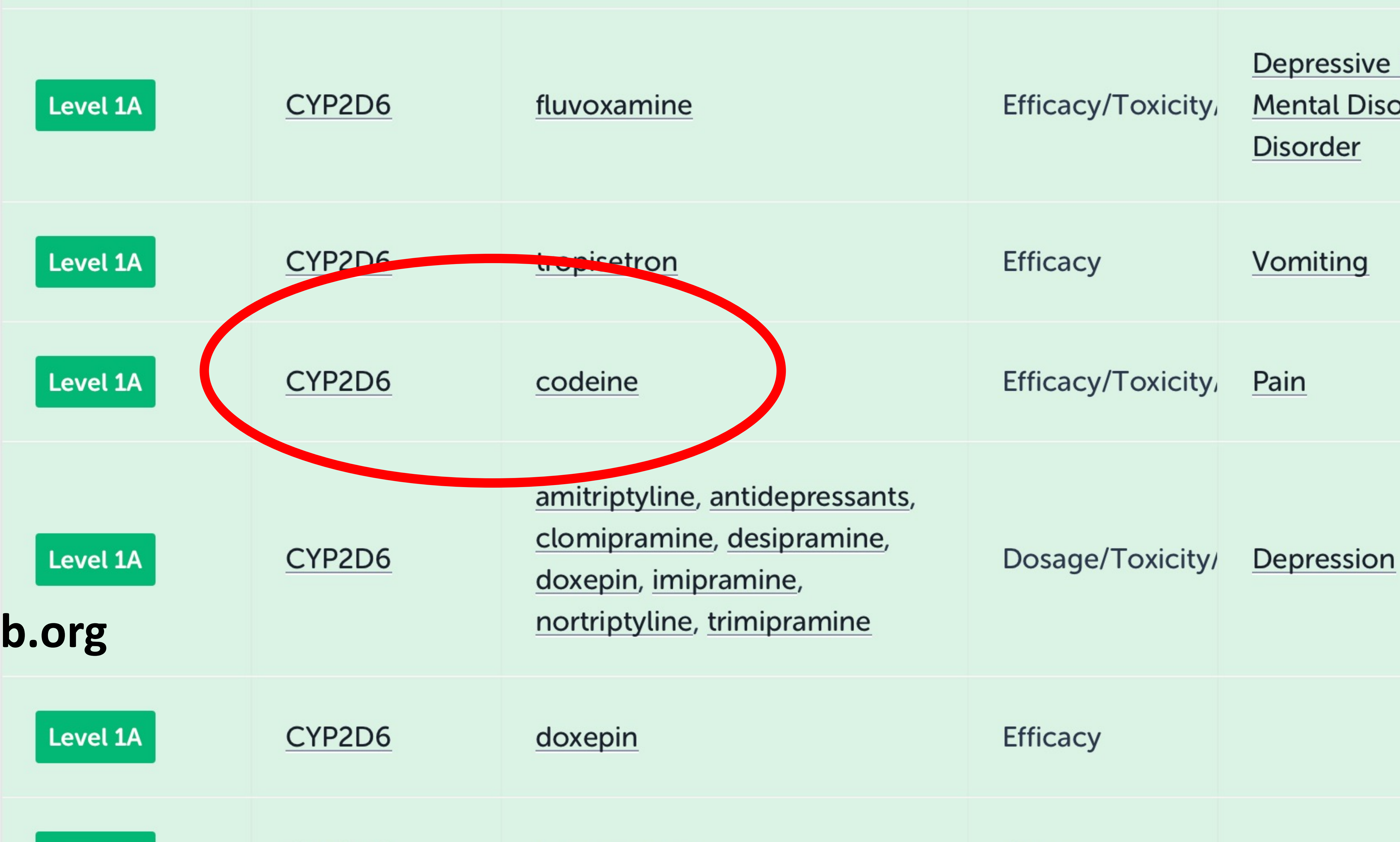
Greg McInness
Adam Lavertu

Rachel Dalton & Erica Woodahl,
U. Montana

PharmGKB & Helix Groups

Pharmacogenetics = variation in drug response due to genetic differences

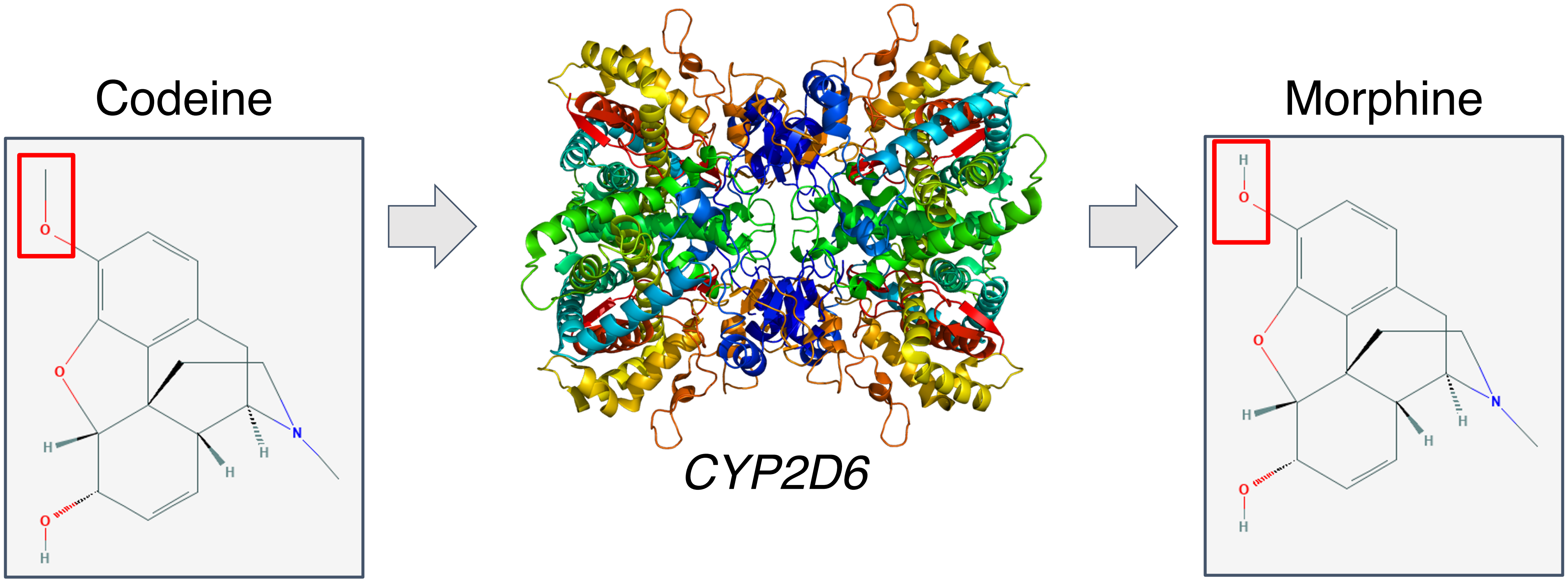
- Drug should work as expected
- Change the dose of the drug
- Increased chance of toxicity for drug
- Use another drug



Level 1A	CYP2D6	fluvoxamine	Efficacy/Toxicity,	Depressive Mental Diso Disorder
Level 1A	CYP2D6	tropisetron	Efficacy	Vomiting
Level 1A	CYP2D6	codeine	Efficacy/Toxicity,	Pain
Level 1A	CYP2D6	amitriptyline, antidepressants, clomipramine, desipramine, doxepin, imipramine, nortriptyline, trimipramine	Dosage/Toxicity/	Depression
Level 1A	CYP2D6	doxepin	Efficacy	

rmgkb.org

Codeine pharmacogenetics



As much as 23% of people in the US have a compromised ability to metabolize opioids

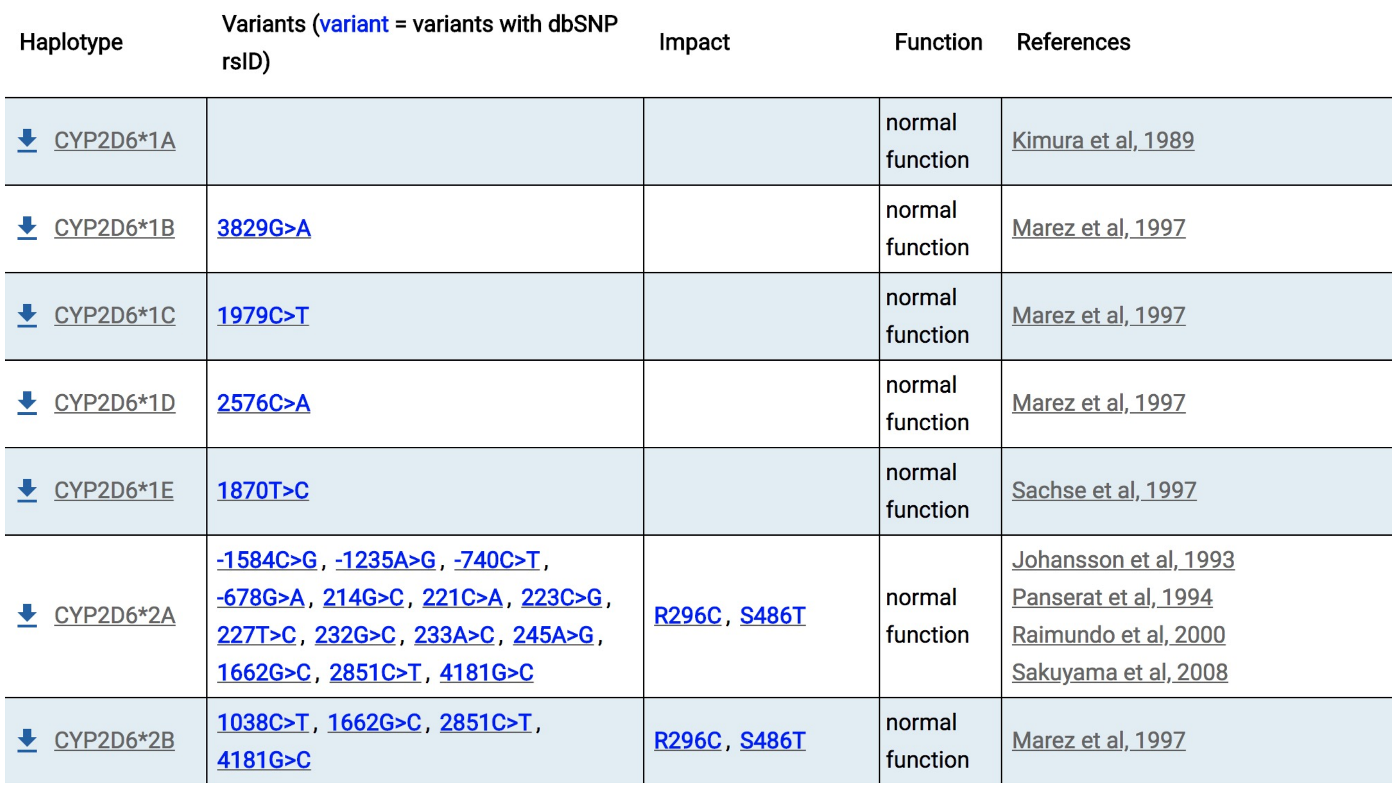
1-5% are poor metabolizers => CODEINE DOES NOT WORK
1-21% are ultra metabolizer => MORPHINE SPIKES IN BLOOD

Marcalus et al, 2019

“Star” Alleles = Haplotypes of pharmacogenes

*1 = Wildtype (Reference Sequence)
*2 = some combination of SNP alleles
*3 = another combination
*4 = etc…

From PharmVar DB



Haplotype	Variants (variant = variants with dbSNP rsID)	Impact	Function	References
CYP2D6*1A			normal function	Kimura et al, 1989
CYP2D6*1B	3829G>A		normal function	Marez et al, 1997
CYP2D6*1C	1979C>T		normal function	Marez et al, 1997
CYP2D6*1D	2576C>A		normal function	Marez et al, 1997
CYP2D6*1E	1870T>C		normal function	Sachse et al, 1997
CYP2D6*2A	-1584C>G, -1235A>G, -740C>T, -678G>A, 214G>C, 221C>A, 223C>G, 227T>C, 232G>C, 233A>C, 245A>G, 1662G>C, 2851C>T, 4181G>C	R296C, S486T	normal function	Johansson et al, 1993 Panserat et al, 1994 Raimundo et al, 2000 Sakuyama et al, 2008
CYP2D6*2B	1038C>T, 1662G>C, 2851C>T, 4181G>C	R296C, S486T	normal function	Marez et al, 1997

CYP2D6 has 161+ observed haplotypes

(many are common)

Some drugs metabolized by CYP2D6

Antidepressants	Beta Blockers	Anti-cancer	Antipsychotics	Other	
Amitriptyline	Alprenolol	Tamoxifen	Haloperidol	Mexiletine	Methamphetamine
Clomipramine	Carvedilol		Perphenazine	Minaprine	Bufuralol
Desipramine	Propafenone		Risperidone	Nebivolol	Chlorpheniramine
Imipramine	Bupranolol		Thioridazine	Nortriptyline	Chlorpromazine
Fluoxetine	Clonidine		Zuclopenthixol	Ondansetron	Clonidine
Paroxetine	Debrisoquine		Atomoxetine	Oxycodone	Codeine
Tamoxetine	Metoprolol		Alprenolol	Perhexiline	Debrisoquine
Trimipramine	Propranolol		Amphetamine	Phenacetin	Dexfenfluramine
Venlafaxine	Timolol		Aripiprazole	Phenformin	Dextromethorphan

see TRANSLATION page 321, February 2012

Clinical Pharmacogenetics Implementation Consortium Guidelines for Cytochrome P450 2D6 Genotype and Codeine Therapy: 2014 Update

KR Crews[1], A Gaedigk[2,3], HM Dunnenberger[1], JS Leeder[2,3], TE Klein[4], KE Caudle[1], CE Haidar[1], DD Shen[5,6], JT Callaghan[7,8], S Sadhasivam[9,10], CA Prows[11,12], ED Kharasch[13] and TC Skaar[7]

Codeine is bioactivated to morphine, a strong opioid agonist, by the hepatic cytochrome P450 2D6 (CYP2D6); hence, the efficacy and safety of codeine are governed by CYP2D6 activity. Polymorphisms are a major cause of CYP2D6 variability. We summarize evidence from the literature supporting this association and provide therapeutic recommendations for codeine based on *CYP2D6* genotype. This document is an update to the 2012 Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for *CYP2D6* genotype and codeine therapy.

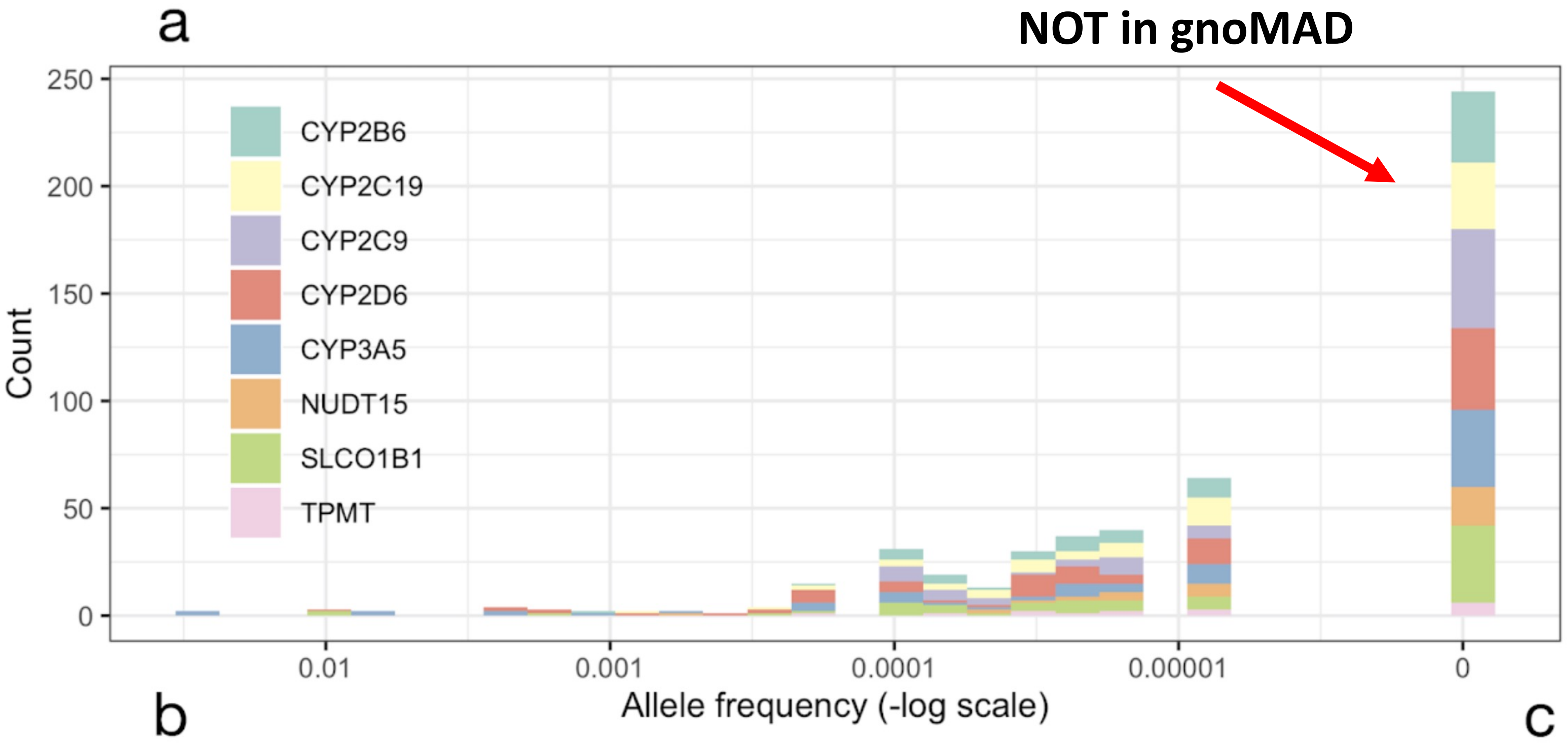
cypalleles.ki.se. Clinical phenotype data are available for common alleles (**Supplementary Tables S1–S5** online). However, many alleles have not been evaluated in clinical trials, and their clinical phenotypes are predicted based on the expected functional impact of their defining genetic variation or are extrapolated based on *in vitro* functional studies using different substrates.

Genetic test interpretation

Most clinical laboratories report *CYP2D6* genotype using the star (*) allele nomenclature and may provide interpretation of

Rare variants in UK Biobank Exomes

- Evaluated: variation in 8 key pharmacogenes (metabolizing enzymes, transporters) including CYP2D6
- 478 predicted-deleterious variants across all 8
- 244 of these not in gnomAD (resource for population variation)
- 6.1% of individuals carry one novel deleterious variant
- Each individual has an average of **12 drugs** for which unusual response might expected
- Novel variants enriched in non-European populations



NOT in gnoMAD

We need methods to assess the impact of novel or rare variations!

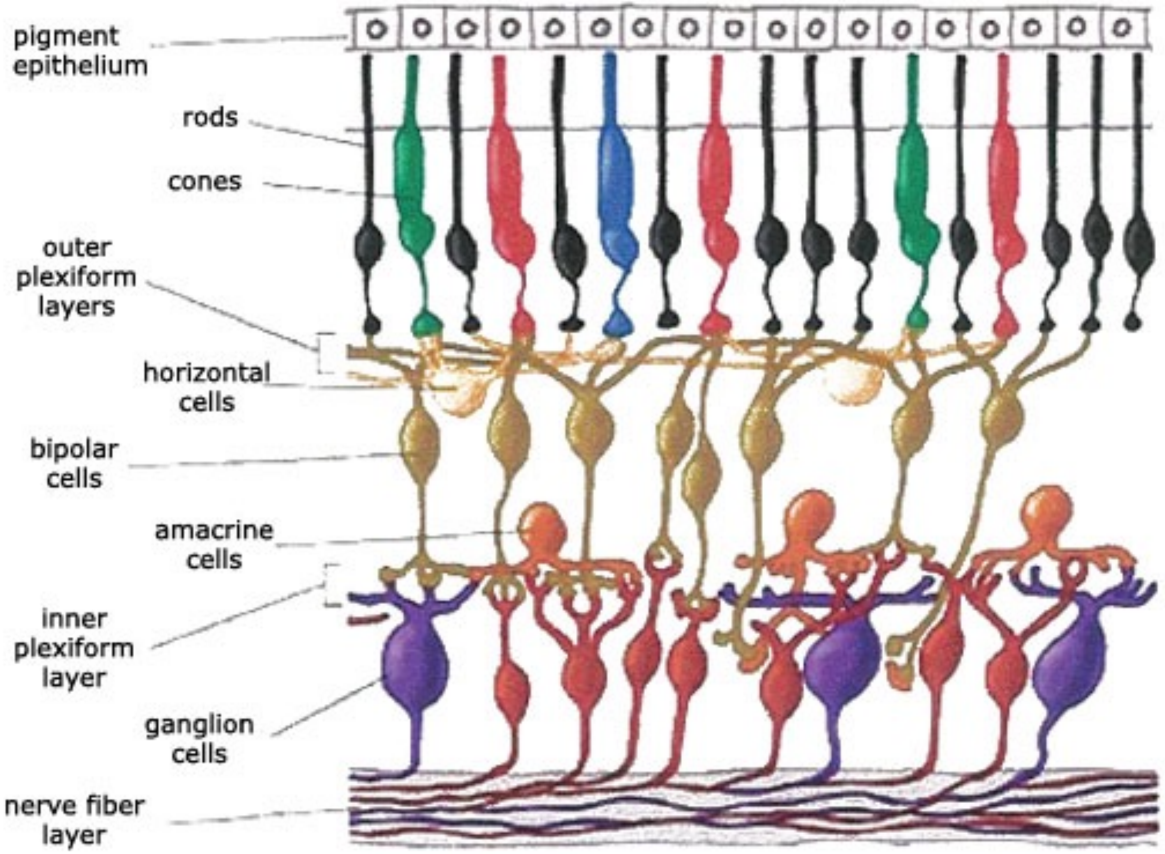
How can we predict the function of the novel haplotypes observed in population surveys?

==

How can we bring clinical pharmacogenetics to patients with rare variants?

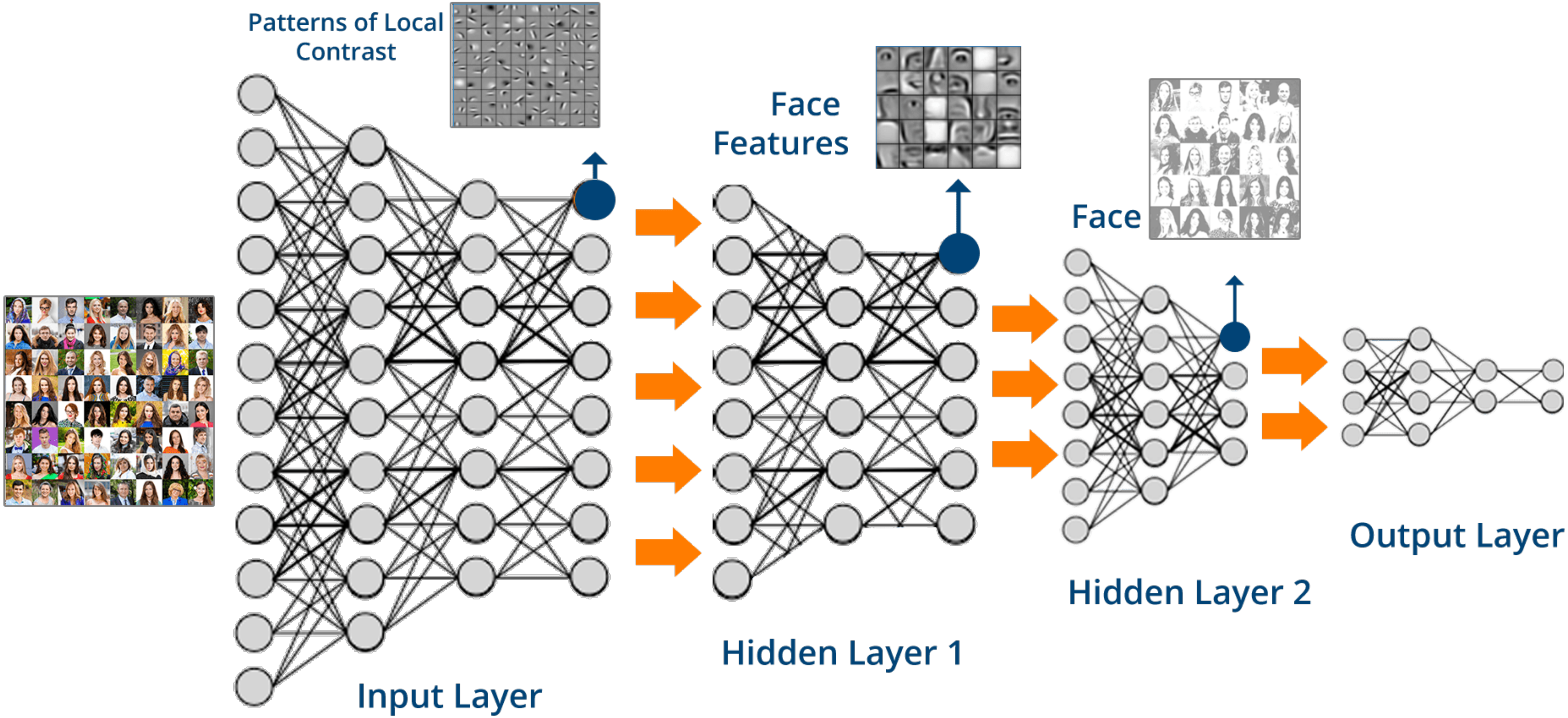
Deep Learning

- Deep Learning is based on an analogy to neural processing = neural networks
- cf. processing of light in the retina.

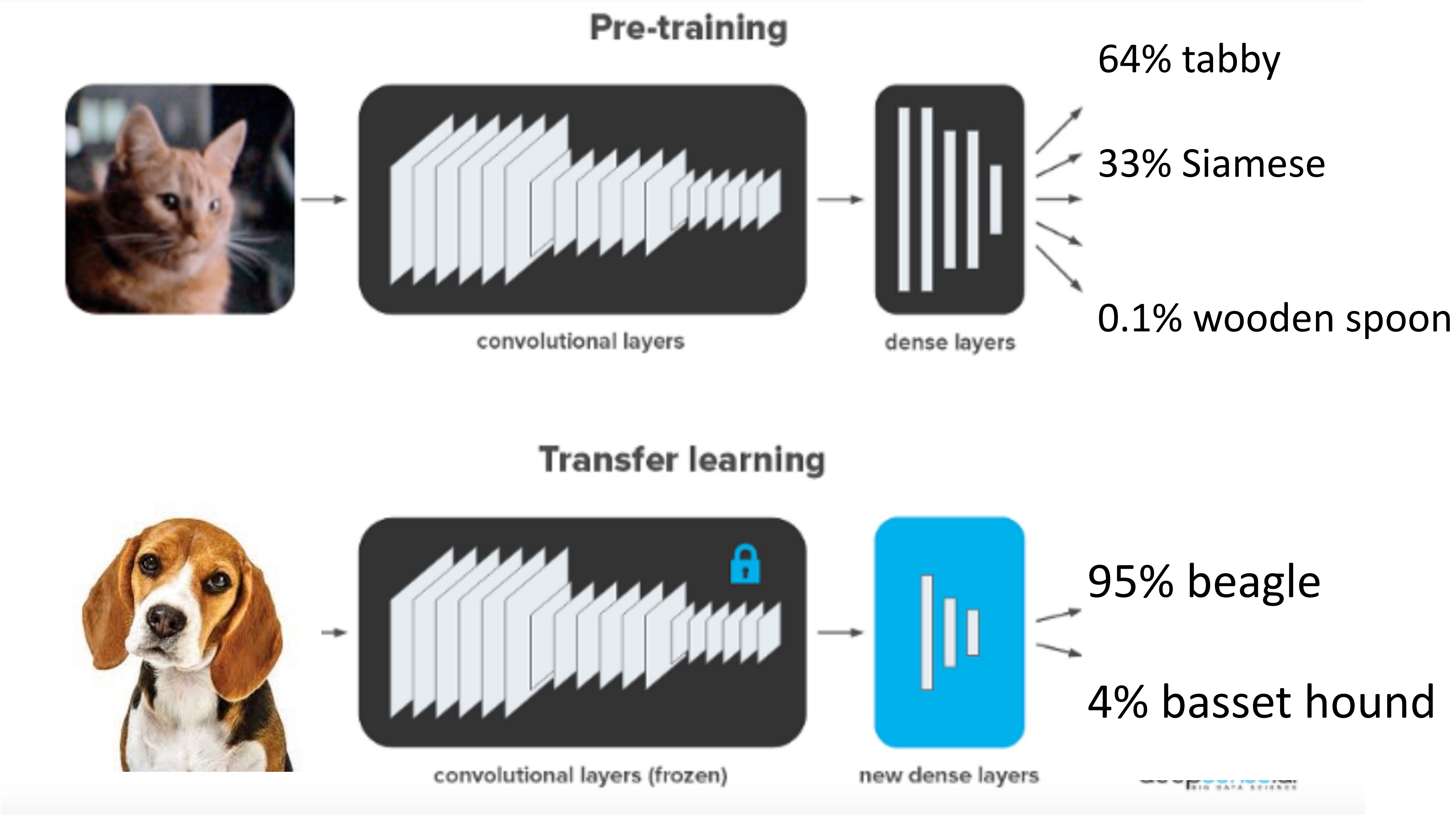


http://www.arn.org/docs/glicksman/eyw_041101.htm

Deep Learning



https://www.michaelchimenti.com/2017/11/deep-neural-nets-software-2-0/



64% tabby

33% Siamese

0.1% wooden spoon

95% beagle

4% basset hound

https://www.freecodecamp.org/news/keras-vs-pytorch-avp-transfer-learning-c8b852c31f02/

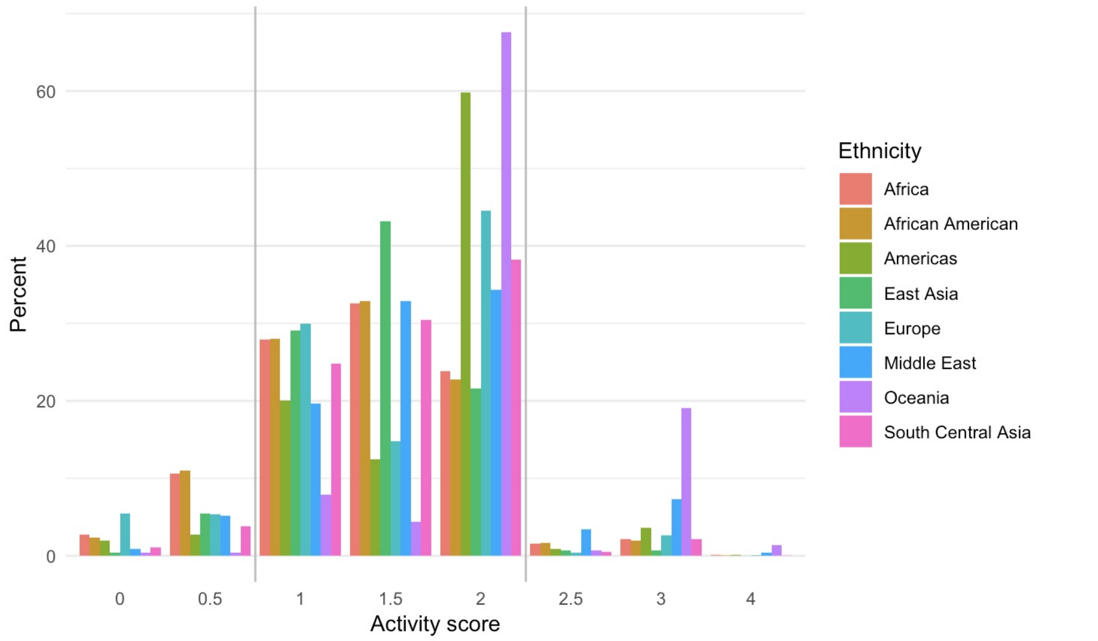
CYP2D6 “Activity Score”

Method for predicting metabolic phenotype from genotype (* allele)

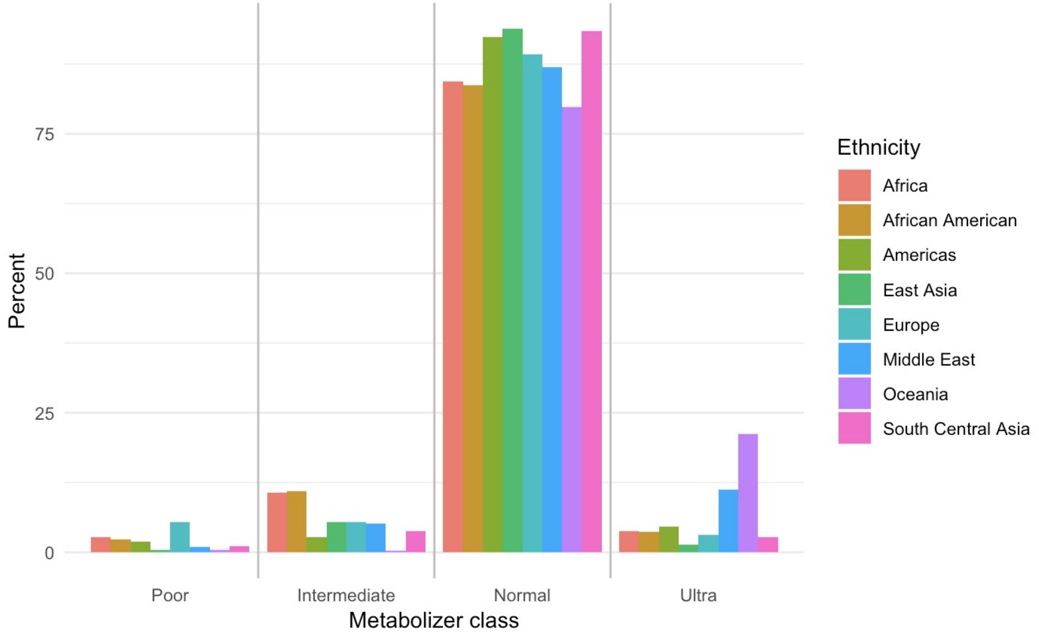
Assigns a score to each haplotype based on *known* functional variants = sum of the haplotype scores

Adapted from Gaedigk et al, 2007

Value assigned	Alleles
0	**3, *4, *4xN, *5, *6, *7, *16, *36, *40, *42, *56B*
0.5	**9, *10, *17, *29, *41, *45, *46*
1	**1, *2, *35, *43, *45xN*
2	**1xN, *2xN, *35xN*



Adapted from Gaedigk et al, 2017



IDEA for CYP2D6 Transfer Learning

- Generate 50,000 sequences on a natural gnoMAD background with known CYP2D6 variations embedded/spiked into these sequences
- Estimate the Activity Score of these sequences
- Train a model to learn how to assign Activity Scores
- (This should force CNN to learn key sequence features)
- Use SPARSE **experimental** (360 samples) & **database** data (~60 * alleles with known function) to refine final layers
- Predict function of haplotypes & assess

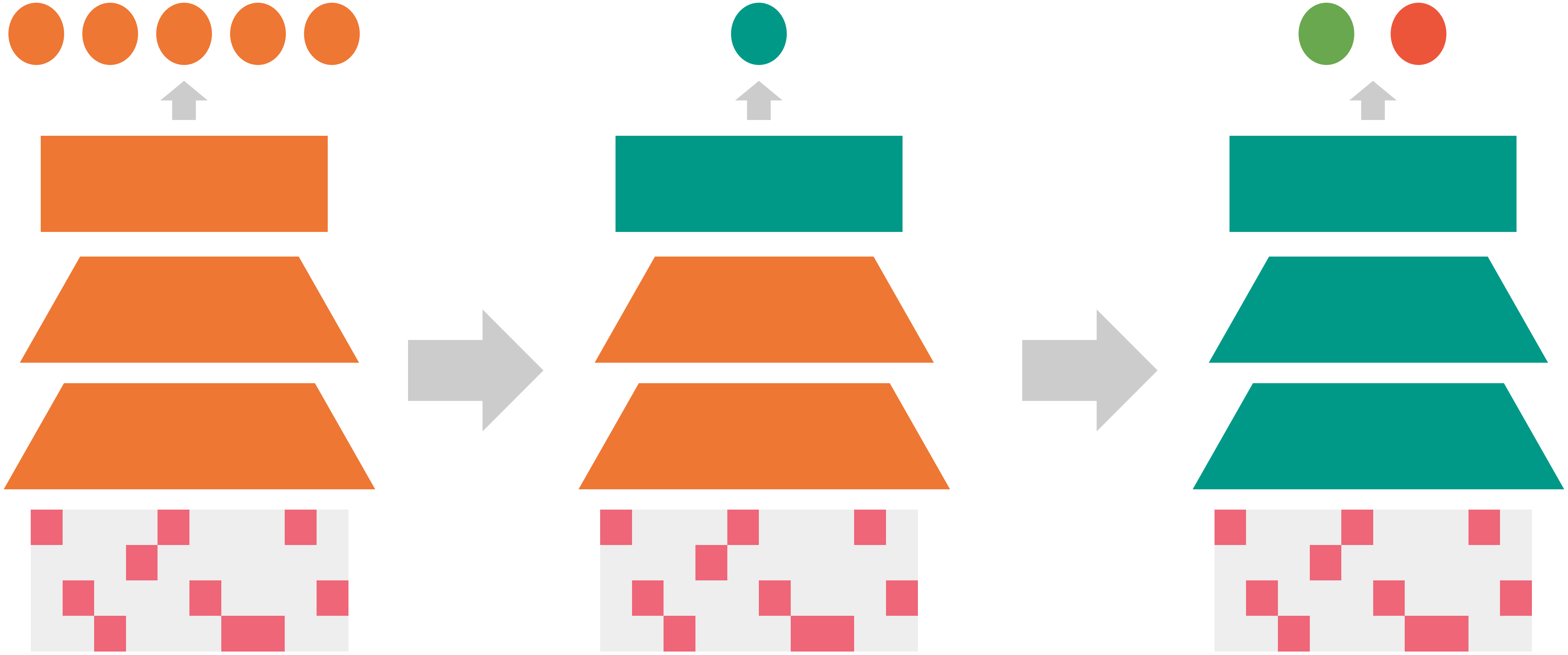
PLoS Comput Biol. 2020 Nov 2;16(11):e1008399

Transfer learning used to train network

Activity score classification

Measured activity regression

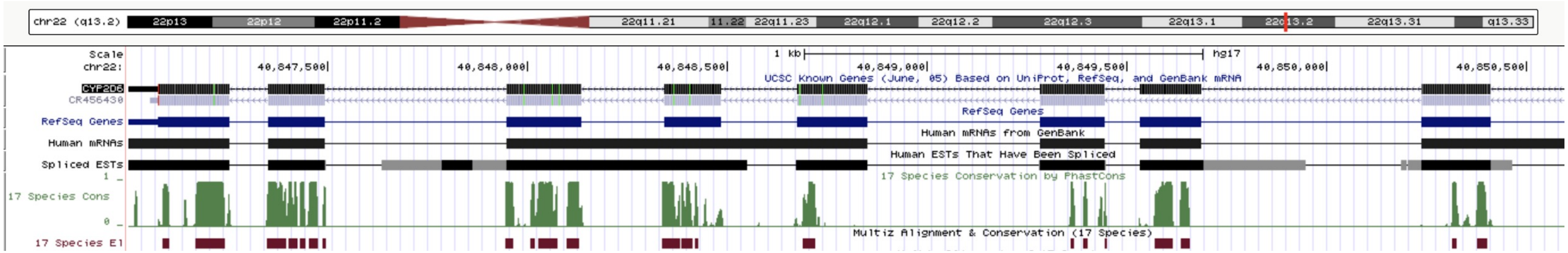
Star allele Classifier



Pretrain on simulated data

Tune on real data via semi-supervised learning

Fine tune on star alleles sequences



CYP2D6: 14,407 base pairs in 9 exons

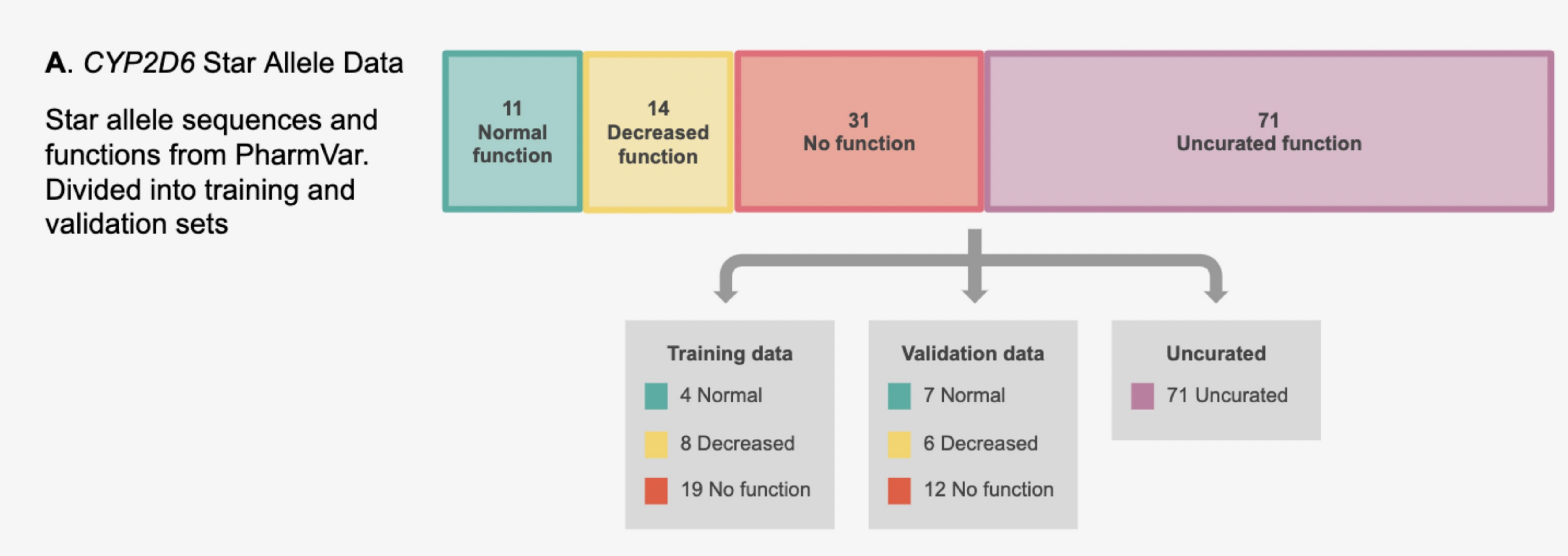
Experimental data (Erica Woodahl & Rachel Dalton)

(360 liver samples, sequenced CYP2D6, 2 activity measurements/sample)

161 variant sites

60 intronic, 56 exonic, 45 upstream/downstream

Gold standard data available from databases



A. *CYP2D6* Star Allele Data

Star allele sequences and functions from PharmVar. Divided into training and validation sets

11
Normal function

14
Decreased function

31
No function

71
Uncurated function

Training data

- 🟩 4 Normal
- 🟨 8 Decreased
- 🟥 19 No function

Validation data

- 🟩 7 Normal
- 🟨 6 Decreased
- 🟥 12 No function

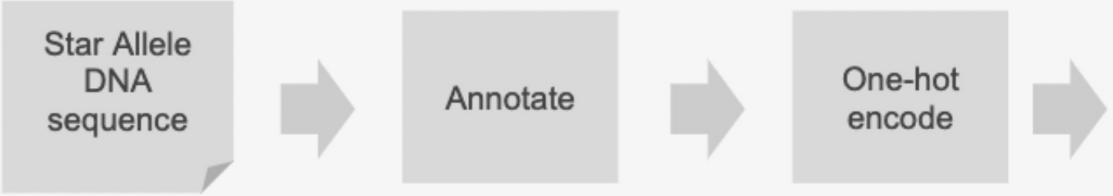
Uncurated

- 🟪 71 Uncurated

Representation of (phased) sequence data

B. Data formatting

Input: *CYP2D6* star allele sequence
Output: One-hot encoded sequence and annotation data



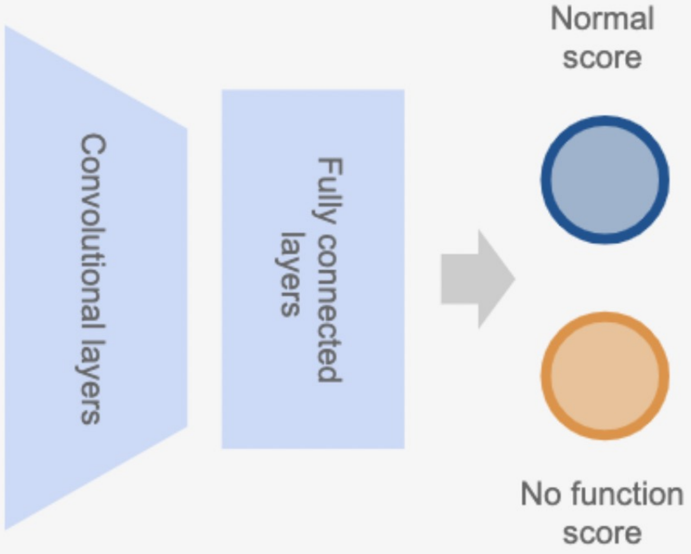
A	🟥				🟥				🟥	
C				🟥						
T		🟥				🟥				🟥
G			🟥				🟥	🟥		
Deleterious		🟥								
eQTL						🟥				🟥

C.Functional prediction

Input: One-hot encoded sequence and annotation data
Output: Functional probabilities: normal function score, no function score.

Annotations +

G				🟥
C		🟥		
A	🟥			
T			🟥	
G				🟥
C		🟥		
A	🟥			
A	🟥			
G				🟥
C		🟥		

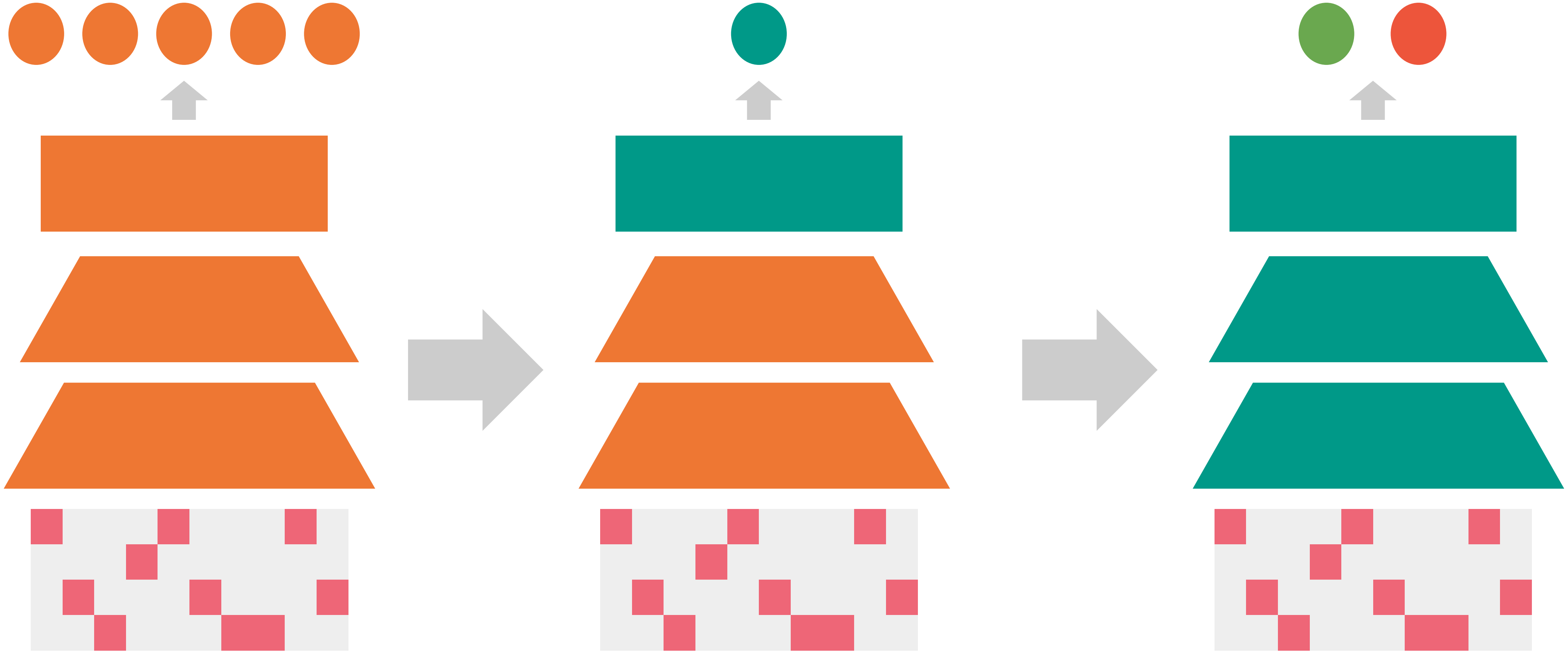


Binary annotations for variants

- In coding region?
- Allele freq < 0.05?
- Deleterious per vote of CADD, DANN, FATHMM, LOFTEE?
- Indel?
- In methylation mark?
- DNA hypersensitivity site?
- TF binding site?
- Known eQTL site?
- Known active site amino acid?

PLoS Comput Biol. 2020 Nov 2;16(11):e1008399

Transfer learning used to train network



Activity score classification

Measured activity regression

Star allele Classifier

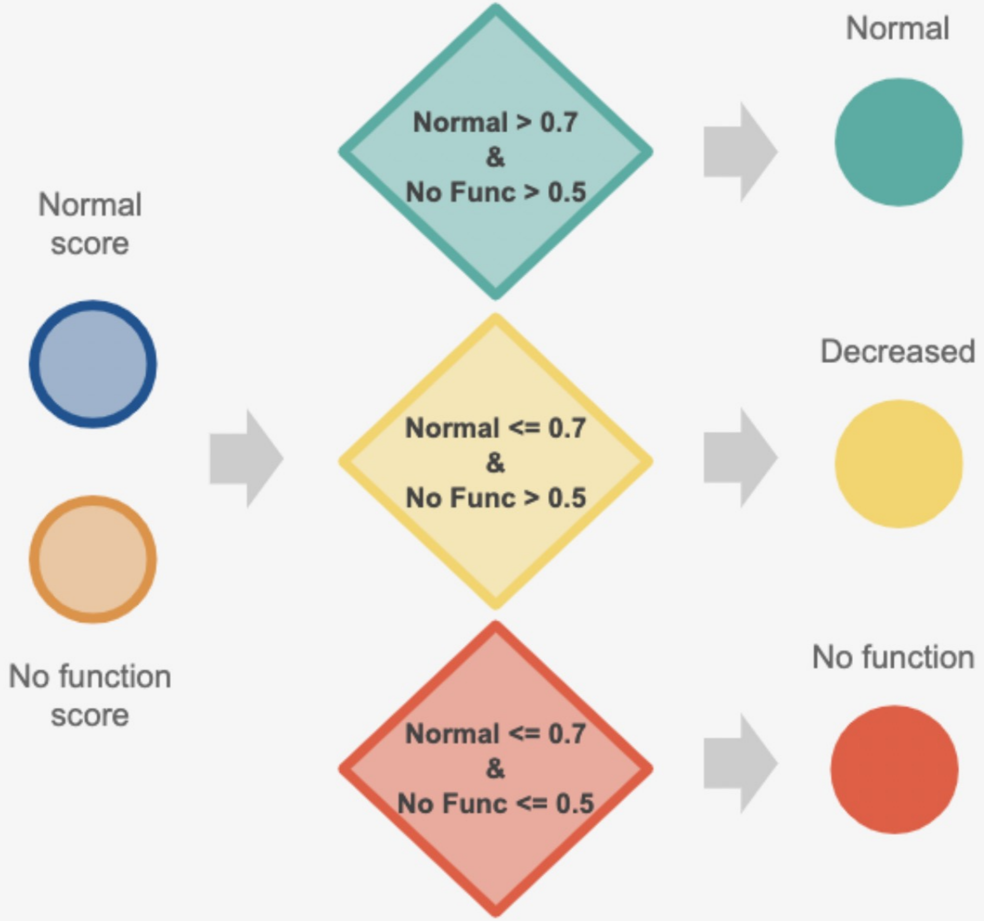
Pretrain on simulated data

Tune on real data via semi-supervised learning

Fine tune on star alleles sequences

D.Conversion of ordinal scores to functional classes

Input: Functional probabilities
Output: CYP2D6 functional prediction



Normal score

No function score

Normal > 0.7 & No Func > 0.5

Normal <= 0.7 & No Func > 0.5

Normal <= 0.7 & No Func <= 0.5

Normal

Decreased

No function

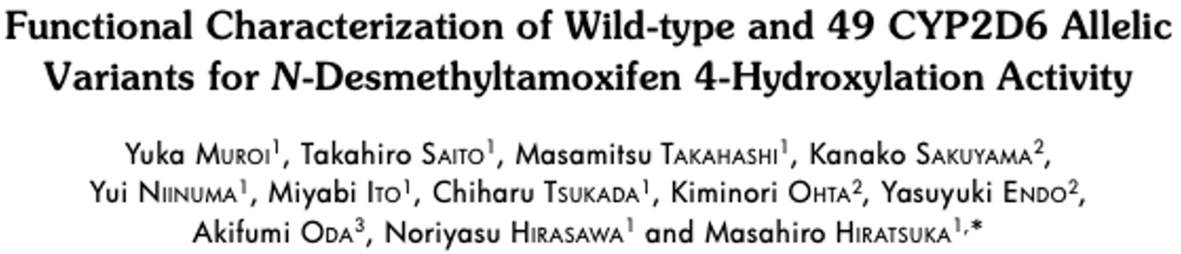
PLoS Comput Biol. 2020 Nov 2;16(11):e1008399

Comparison of predicted function with *in vitro* data

Validate predictions using *in vitro* data from large study.

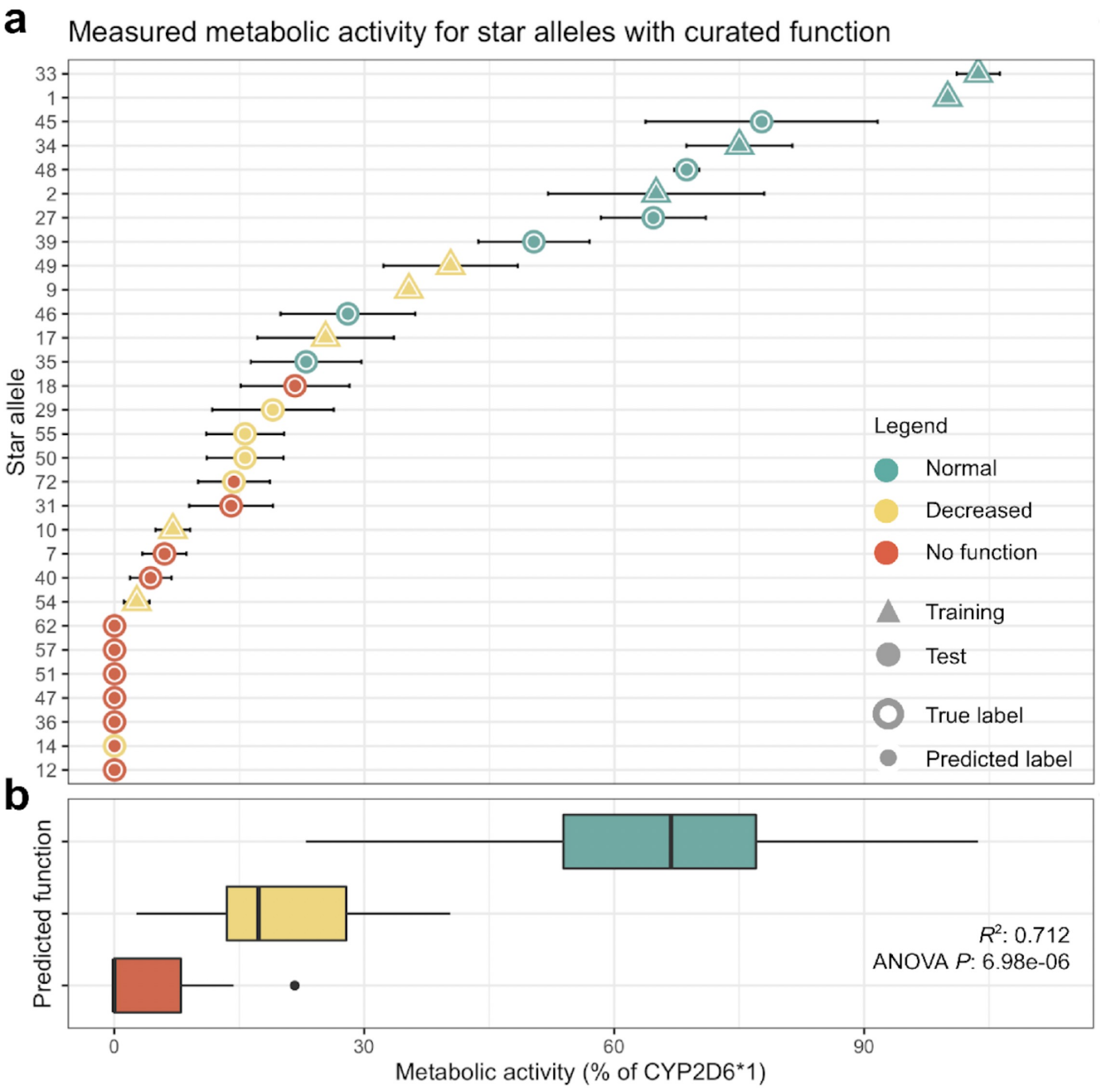
71% variance explained by functional labels

Star allele function measured

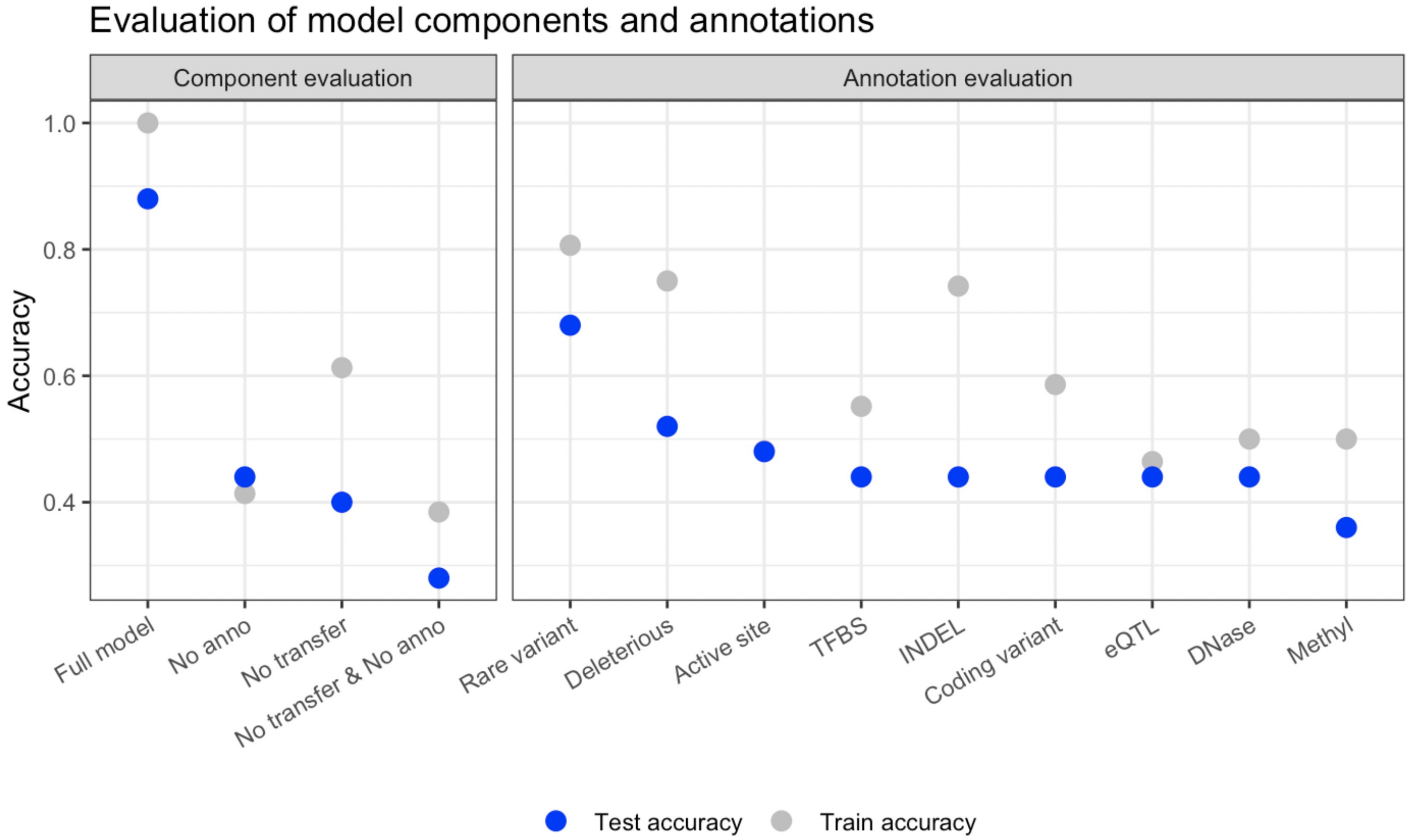


Functional Characterization of Wild-type and 49 CYP2D6 Allelic Variants for *N*-Desmethyltamoxifen 4-Hydroxylation Activity

Yuka Muroi[1], Takahiro Saito[1], Masamitsu Takahashi[1], Kanako Sakuyama[2], Yui Niinuma[1], Miyabi Ito[1], Chiharu Tsukada[1], Kiminori Ohta[2], Yasuyuki Endo[2], Akifumi Oda[3], Noriyasu Hirasawa[1] and Masahiro Hiratsuka[1,*]

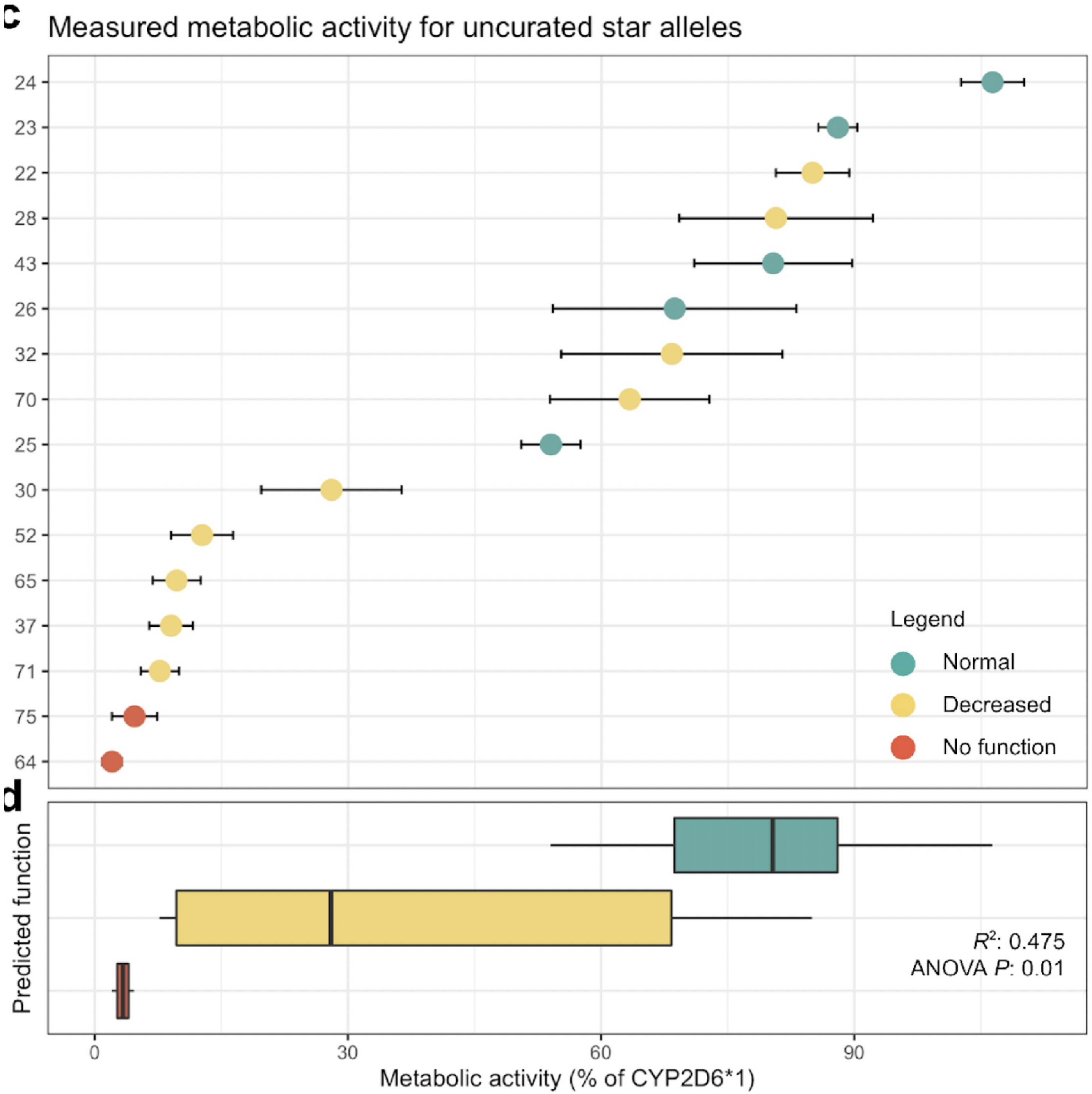


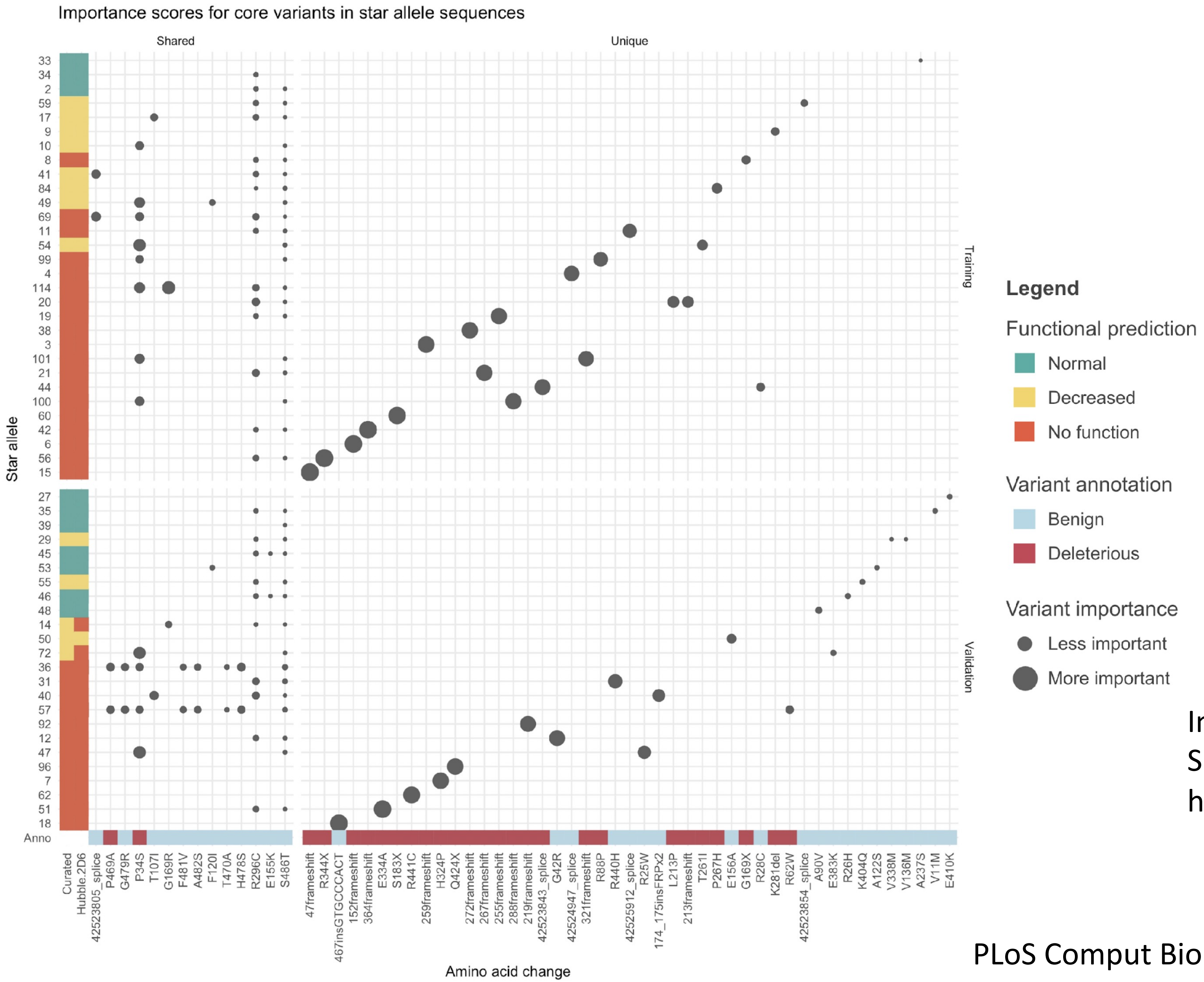
PLoS Comput Biol. 2020 Nov 2;16(11):e1008399


Evaluation of model components and annotations
Component evaluation
Annotation evaluation
Accuracy
1.0
0.8
0.6
0.4
Full model
No anno
No transfer
No transfer & No anno
Rare variant
Deleterious
Active site
TFBS
INDEL
Coding variant
eQTL
DNase
Methyl
Test accuracy
Train accuracy

Performance on novel uncurated star alleles

- Patients with these haplotypes would currently be told: "no information available”





Importance using DeepLift
Shrikumar, Greenside & Kundaje
https://arxiv.org/abs/1704.02685

PLoS Comput Biol. 2020 Nov 2;16(11):e1008399

Conclusions

- Pharmacogenomics is entering clinical care and is useful chiefly in the context of common variants

- UK Biobank analysis indicates large numbers of people with variations in pharmacogenes that are not currently characterized, thus limiting impact.

- Deep learning methods (in this case with transfer learning) hold promise for predicting clinically useful pharmacogenomic phenotypes for novel (chiefly rare) variations in important genes.

Rachel Dalton & Erica Woodahl

Thanks!

russ.altman@stanford.edu
www.pharmgkb.org

NIH National Human Genome Research Institute

FDA U.S. FOOD & DRUG ADMINISTRATION

NIH National Institute of General Medical Sciences

NIH National Center for Advancing Translational Sciences

NIH NLM

CHAN ZUCKERBERG BIOHUB

Table 2. Drug-gene side effect relationship results. Associations are presented in three groups: drug-gene pairs with CPIC guidelines, pairs with no guidelines but evidence in PharmGKB, and novel associations. Phenotype is the gene phenotype (IM: Intermediate Metabolizer, PM: Poor Metabolizer, RM: Rapid Metabolizer, UM: Ultrarapid Metabolizer, IF: Increased Function, PF: Poor Function). Odds ratio is the odds ratio relative to normal metabolizer or normal function alleles. * indicates significance with Bonferroni adjusted p-value threshold of 1.0 x 10^{-5}. Only results with a standard error less than 0.2 are included.

Group	Drug	Gene	Level of Evidence	Phenotype	ICD-10	Code definition	Odds ratio	p-value
CPIC Guidance	citalopram	CYP2C19	1A	IM	B02	Herpes zoster	0.53	8.76E-05
	simvastatin	SLCO1B1	1A	IF	M65	Synovitis and tenosynovitis	1.82	1.42E-04
	amitriptyline	CYP2C19	1A	RM	R53	Malaise and fatigue	1.55	1.74E-04
	amitriptyline	CYP2C19	1A	UM	J30	Vasomotor and allergic rhinitis	1.94	2.75E-04
	codeine	CYP2D6	1A	PM	A52	Late syphilis	1.78	3.30E-04
	ibuprofen	CYP2C9	1A	PM	E13	Other specified diabetes mellitus	2.00	4.90E-04
	clopidogrel	CYP2C19	1A	RM	B08	Viral infections characterized by skin and mucous membrane lesions	0.59	5.17E-04
	tamoxifen	CYP2D6	1A	IM	C50	Malignant neoplasm of breast	0.62	6.98E-04
	simvastatin	SLCO1B1	1A	PF	M79	Unspecified soft tissue disorders	1.49	7.46E-04
	simvastatin	SLCO1B1	1A	DF	M65	Synovitis and tenosynovitis	1.79	7.75E-04
No Guidance	citalopram	CYP2D6	3	IM	J45	Asthma	1.44	9.13E-05
	citalopram	CYP2D6	3	IM	I50	Heart failure	1.56	1.12E-04
	simvastatin	CYP2C9	3	PM	J01	Acute sinusitis	1.74	1.56E-04
	citalopram	CYP2D6	3	IM	J64	Unspecified pneumoconiosis	1.56	5.74E-04
	propranolol	CYP2D6	4	IM	O86	Other puerperal infections	1.85	6.38E-04
Novel associations	diazepam	CYP2C9	NA	PM	M19	Osteoarthritis	2.33	4.52E-06*
	zopiclone	CYP2C9	NA	IM	H91	Unspecified hearing loss	2.20	1.73E-05
	loratadine	CYP2D6	NA	IM	M16	Osteoarthritis of hip	1.98	1.20E-04
	tramadol	CYP2B6	NA	PM	H61	Disorders of external ear	1.95	1.86E-04
	quinine	SLCO1B1	NA	IF	N39	Disorders of urinary system	1.95	1.87E-04