Integrating personalized gene expression profiles into predictive disease-associated gene pools

Emre Güney, PhD

Institute for Research in Biomedicine (IRB)
Barcelona, Spain

ISMB/ECCB'17
July 25th, 2017

npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

Integrating personalized gene expression profiles into predictive disease-associated gene pools

Jörg Menche[1,2,3], Emre Guney[1,4], Amitabh Sharma[1,4,5], Patrick J. Branigan[6], Matthew J. Loza[6], Frédéric Baribaud[6], Radu Dobrin[6] and Albert-László Barabási[1,2,4,5]

Gene expression data are routinely used to identify genes that *on average* exhibit different expression levels between a case and a control group. Yet, very few of such differentially expressed genes are detectably perturbed in individual patients. Here, we develop

ISMB/ECCB'17
July 25th, 2017

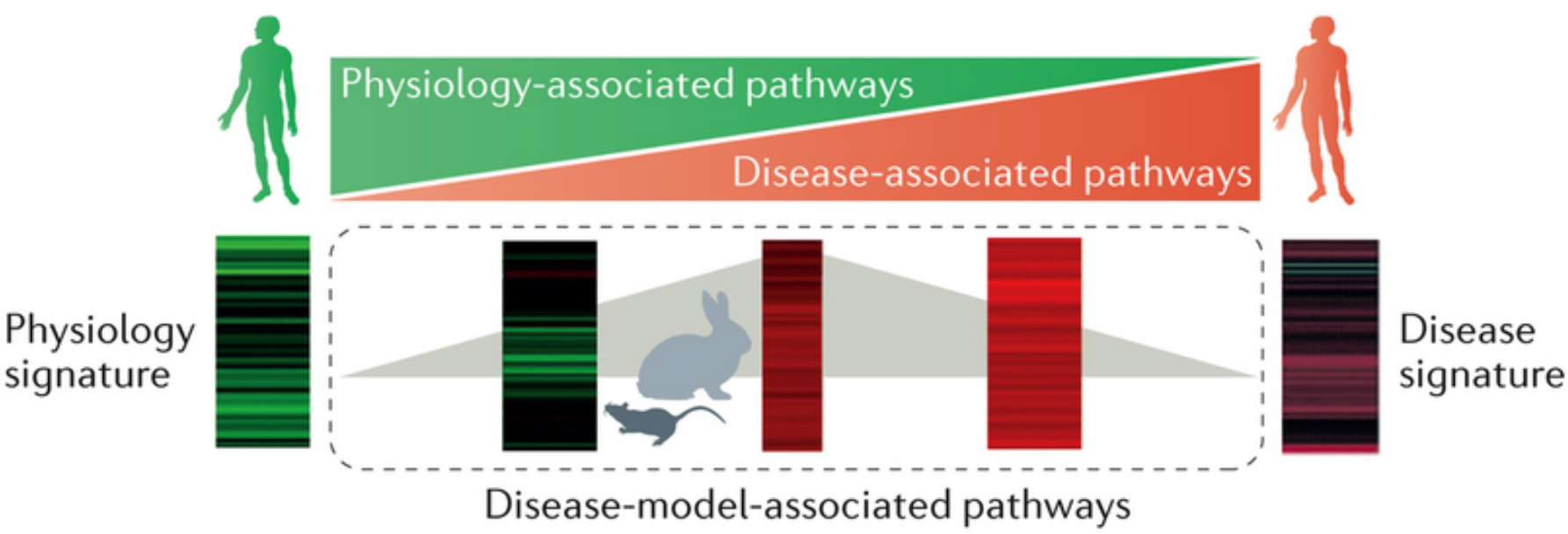


Image from *Moffat et al., 2017, Nat Rev Drug Discov*

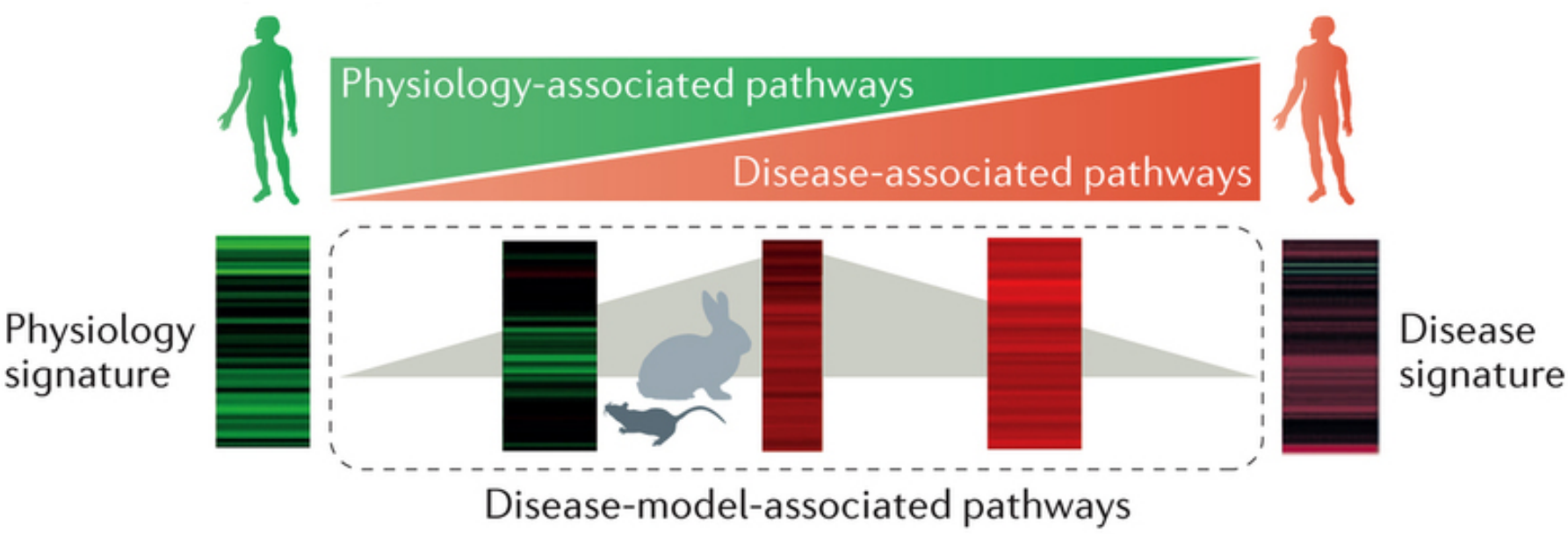


Image from *Moffat et al., 2017, Nat Rev Drug Discov*

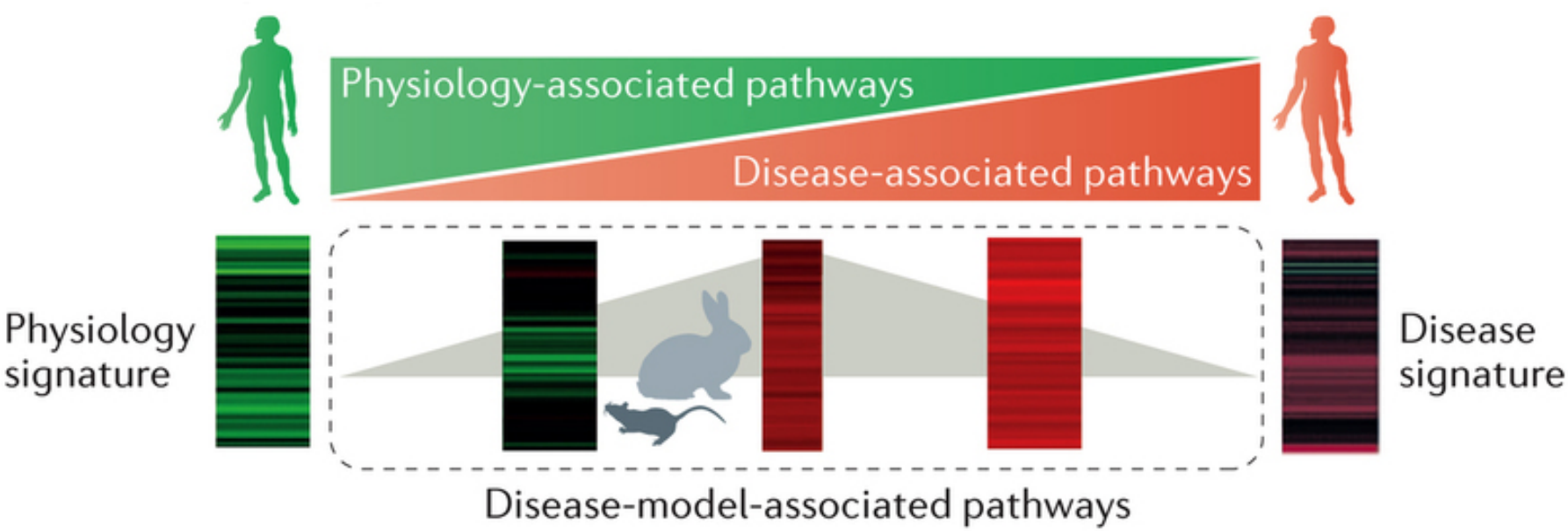
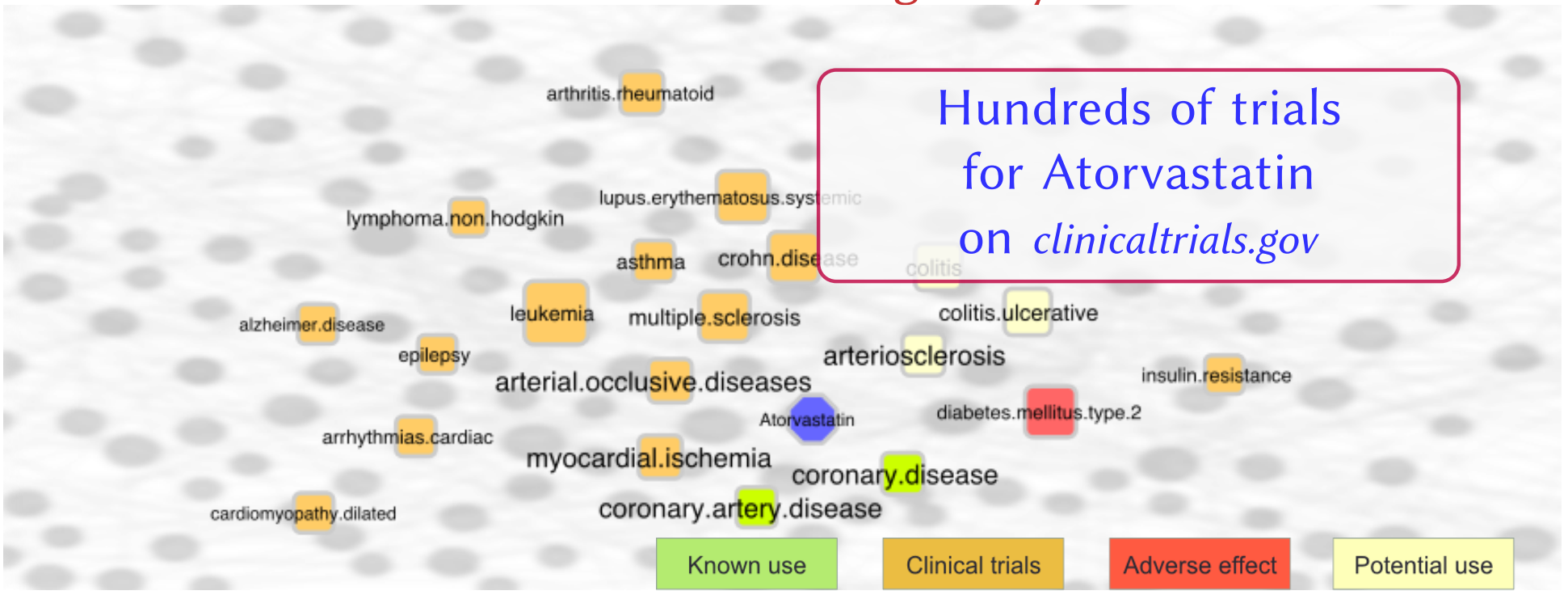


Image from *Moffat et al., 2017, Nat Rev Drug Discov*

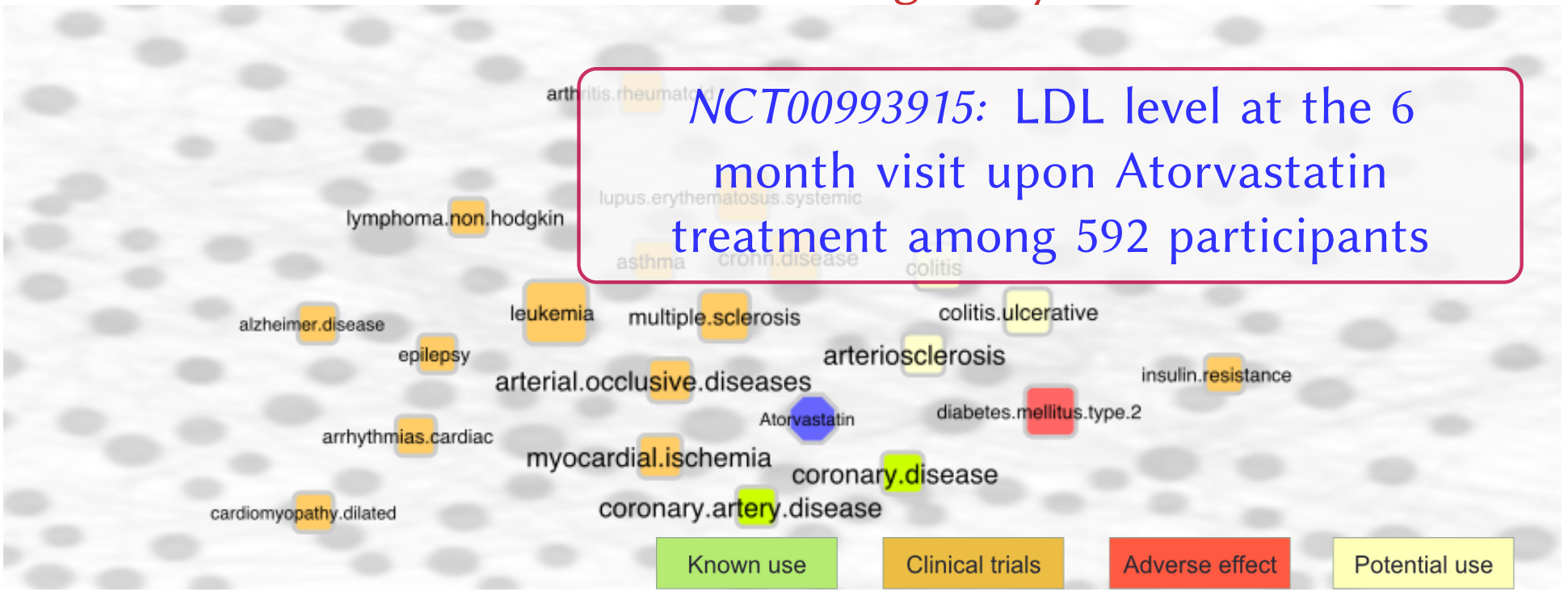
Patient-level heterogeneity



Hundreds of trials for Atorvastatin on *clinicaltrials.gov*

Atorvastatin: one of the most successful cholesterol-lowering drug

Patient-level heterogeneity



Atorvastatin: one of the most successful cholesterol-lowering drug

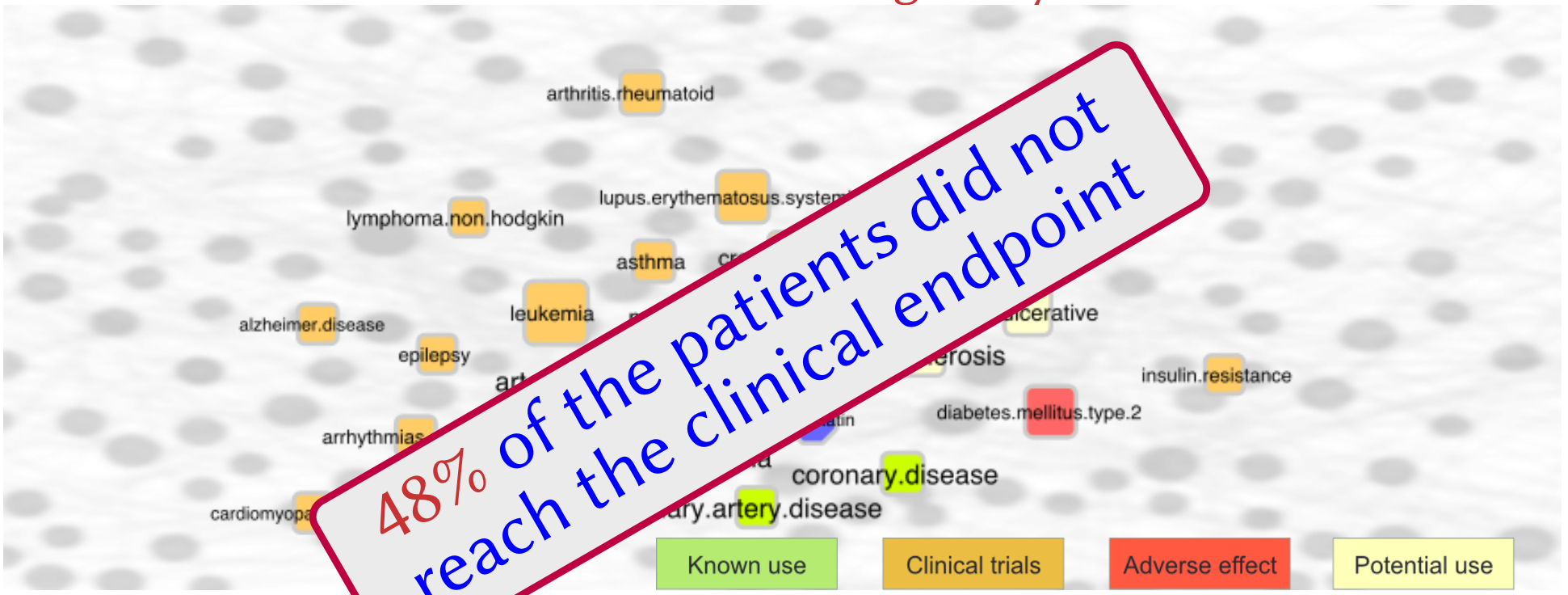
Clinical endpoint: LDL $\leq$ 100 mg/dL after treatment

[*Pre-treatment*] The average LDL level 179 mg/dL

[*Post-treatment*] 52% of the participants had LDL $\leq$ 100 mg/dL

clinicaltrials.gov/ct2/show/results/NCT00993915

Patient-level heterogeneity



Atorvastatin: one of the most successful cholesterol-lowering drug

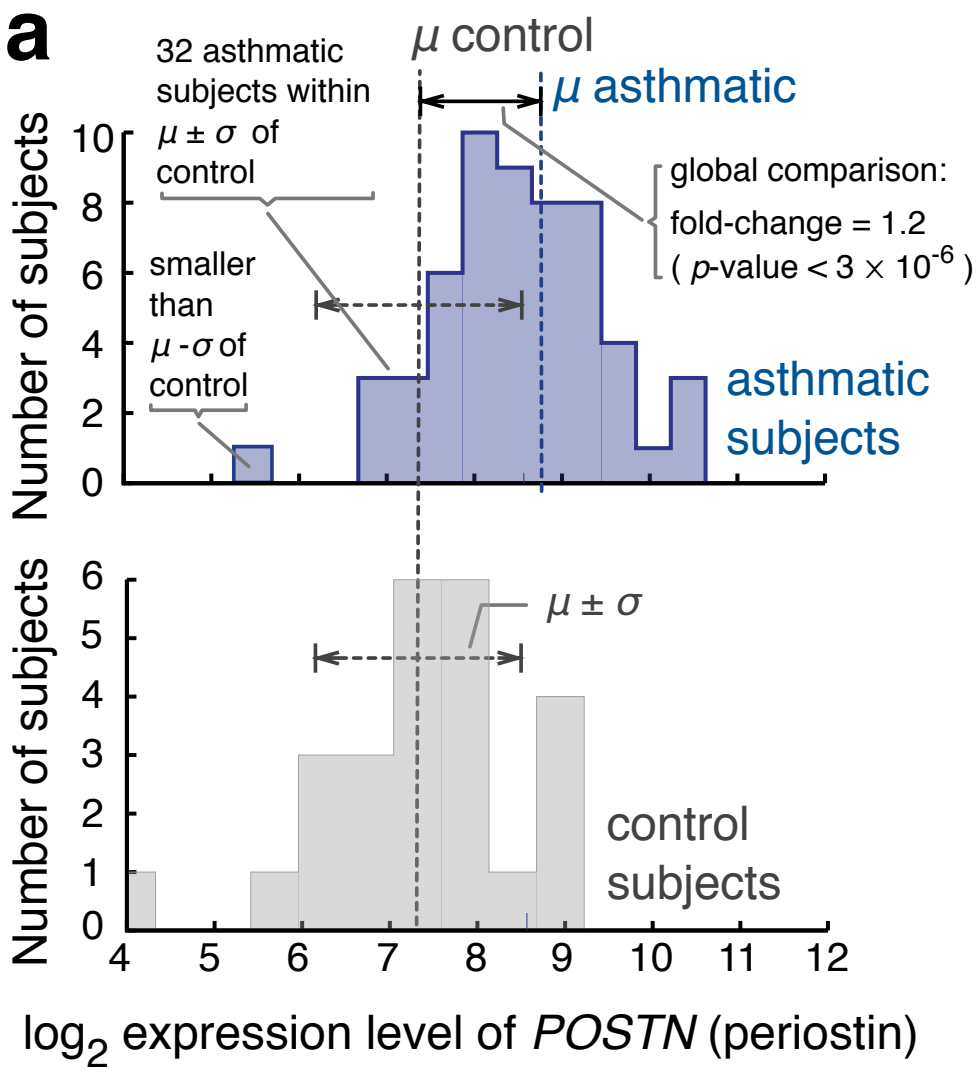
Clinical endpoint: LDL $\leq$ 100 mg/dL after treatment

[*Pre-treatment*] The average LDL level 179 mg/dL

[*Post-treatment*] 52% of the participants had LDL $\leq$ 100 mg/dL

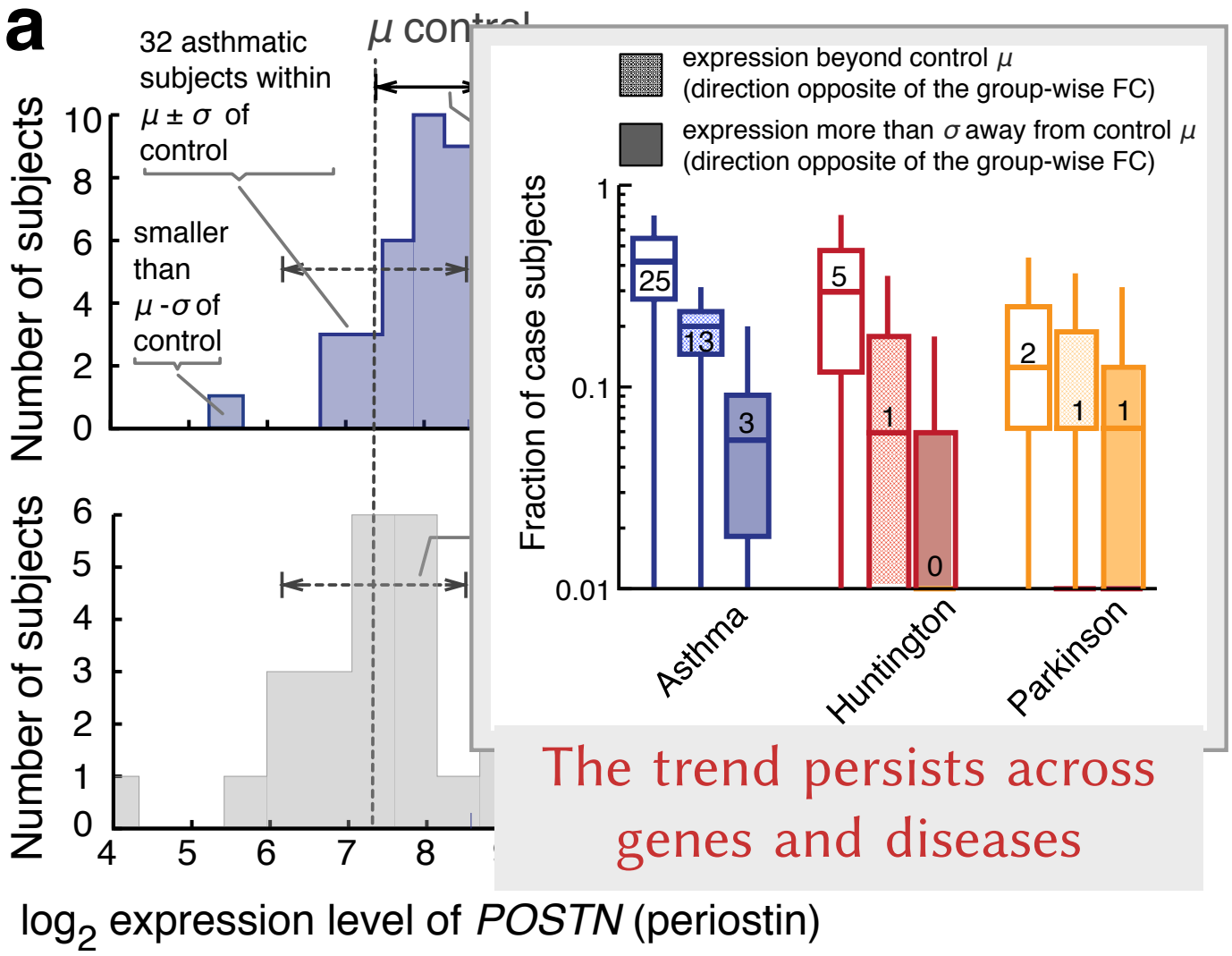
clinicaltrials.gov/ct2/show/results/NCT00993915

Gene expression of POSTN in asthmatic and control invidividuals



Periostin (*POSTN*) functions in the adhesion and migration of epithelial cells, whose up-regulation is a biomarker for Asthma.

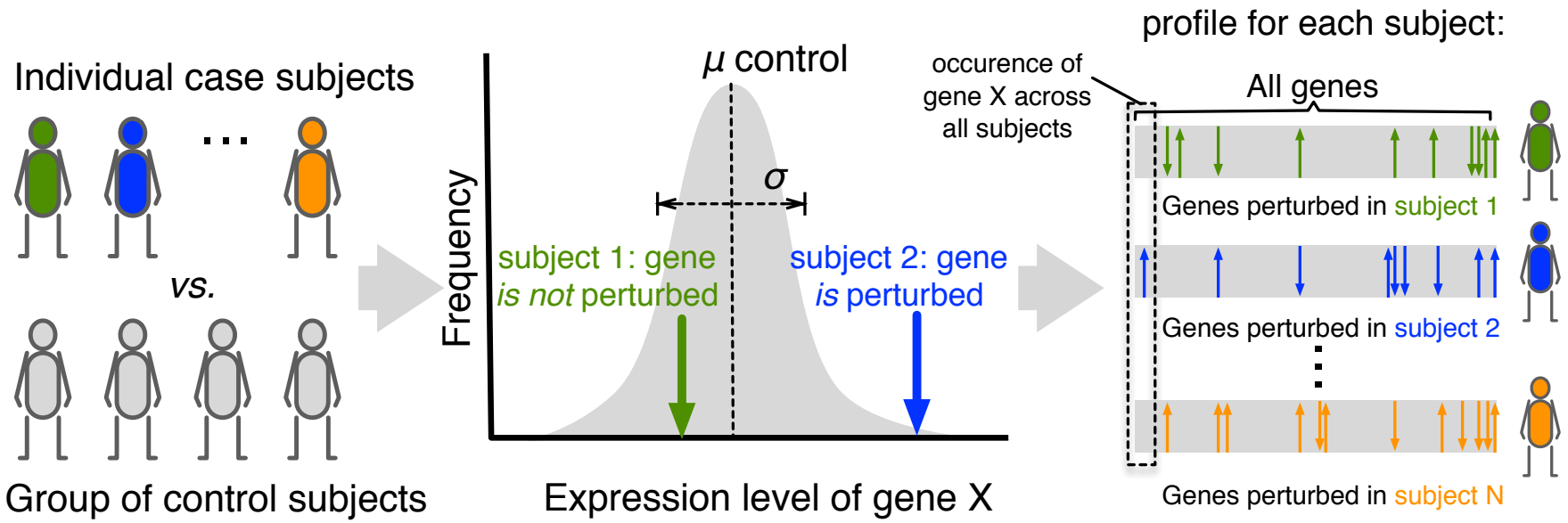
Gene expression of POSTN in asthmatic and control invidividuals



The trend persists across genes and diseases

Periostin (*POSTN*) functions in the adhesion and migration of epithelial cells, whose up-regulation is a biomarker for Asthma.

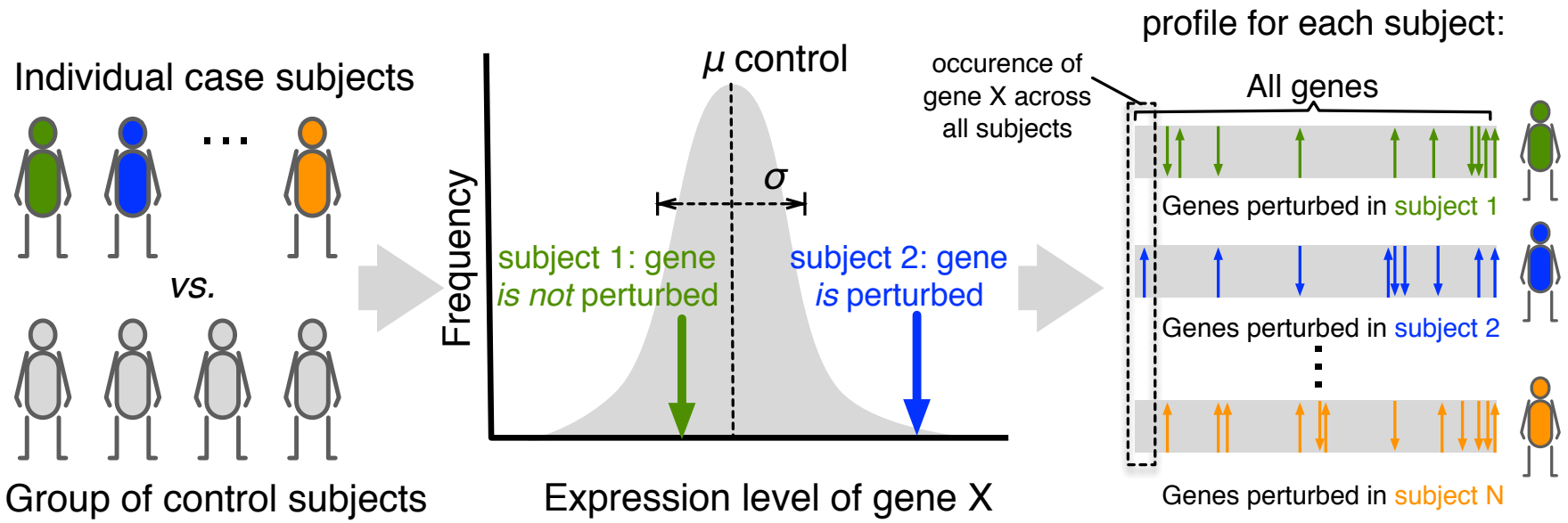
PEEP: PErsonalized Expression Profile



	Samples			
Case	x	x	...	x
Control	c	c	c	c

$$z(gene \text{ in } x) = \frac{expression_x(gene) - \mu_c(gene)}{\sigma_c(gene)}$$

PEEP: PErsonalized Expression Profile



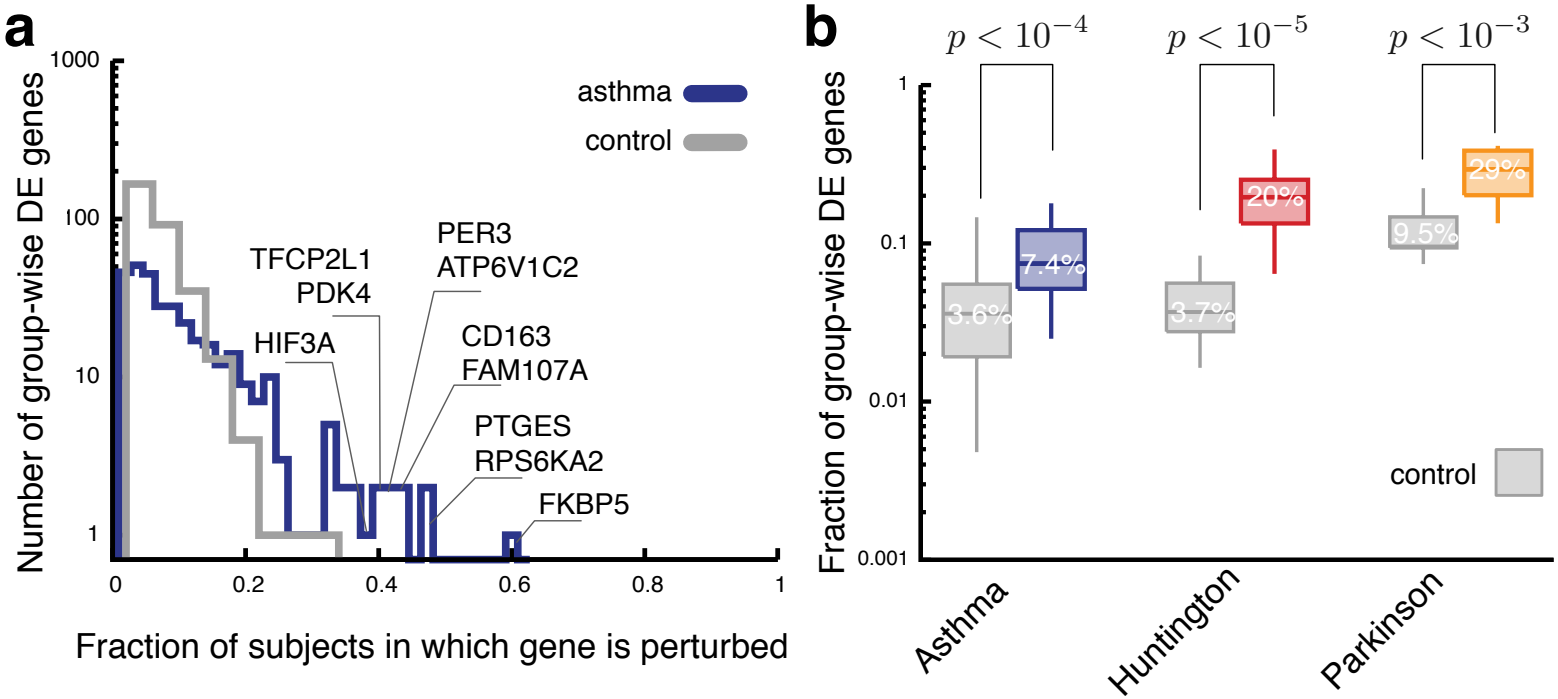
	Samples			
Case	x	x	...	x
Control	c	c	c	c

$$z(gene \text{ in } \mathrm{x}) = \frac{expression_{x}(gene) - \mu_{c}(gene)}{\sigma_{c}(gene)}$$

$$PEEP(x) : \forall gene \; |z(gene \text{ in } \mathrm{x})| > z_{threshold}$$

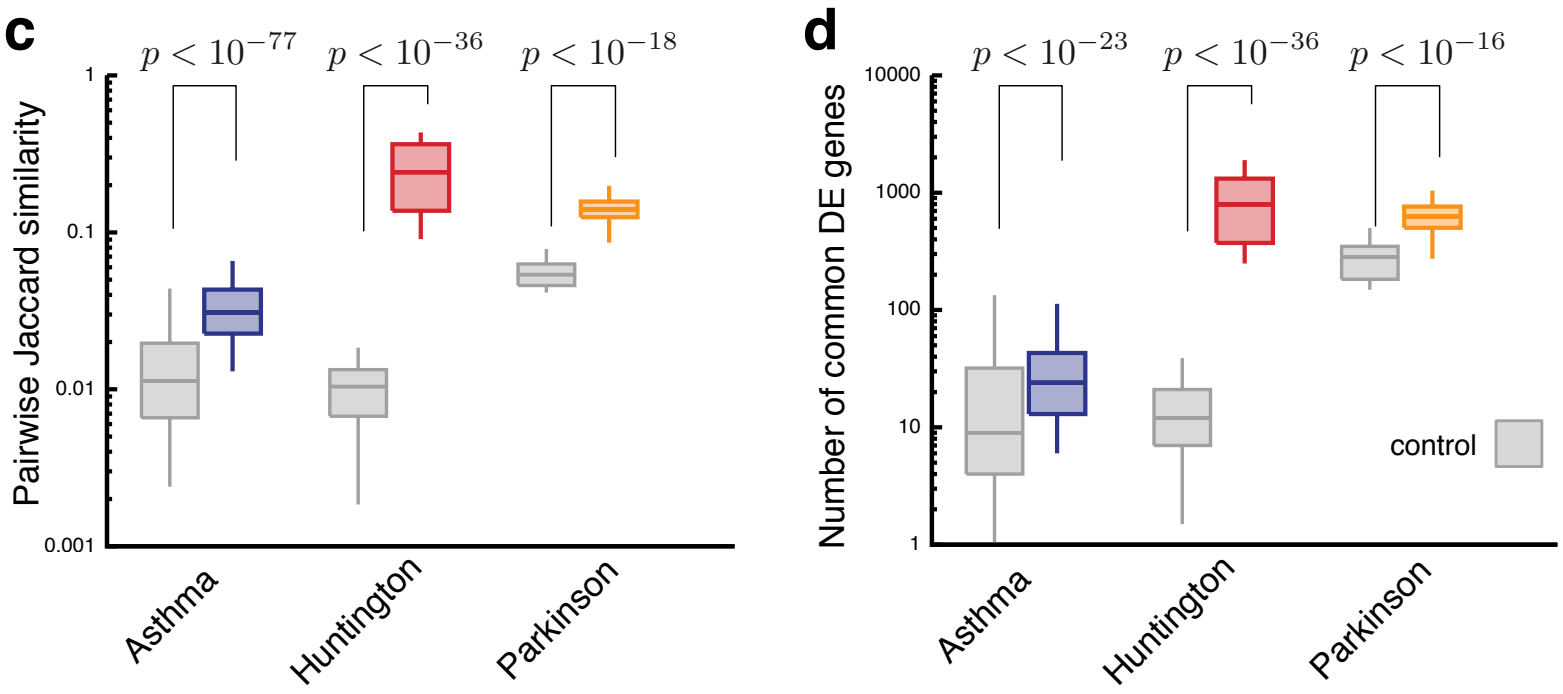
[Genes that are significantly perturbed in each individual]

Quantifying the heterogeneity using PEEPs



Only a small fraction of group-wise DE genes appears in PEEP of each patient

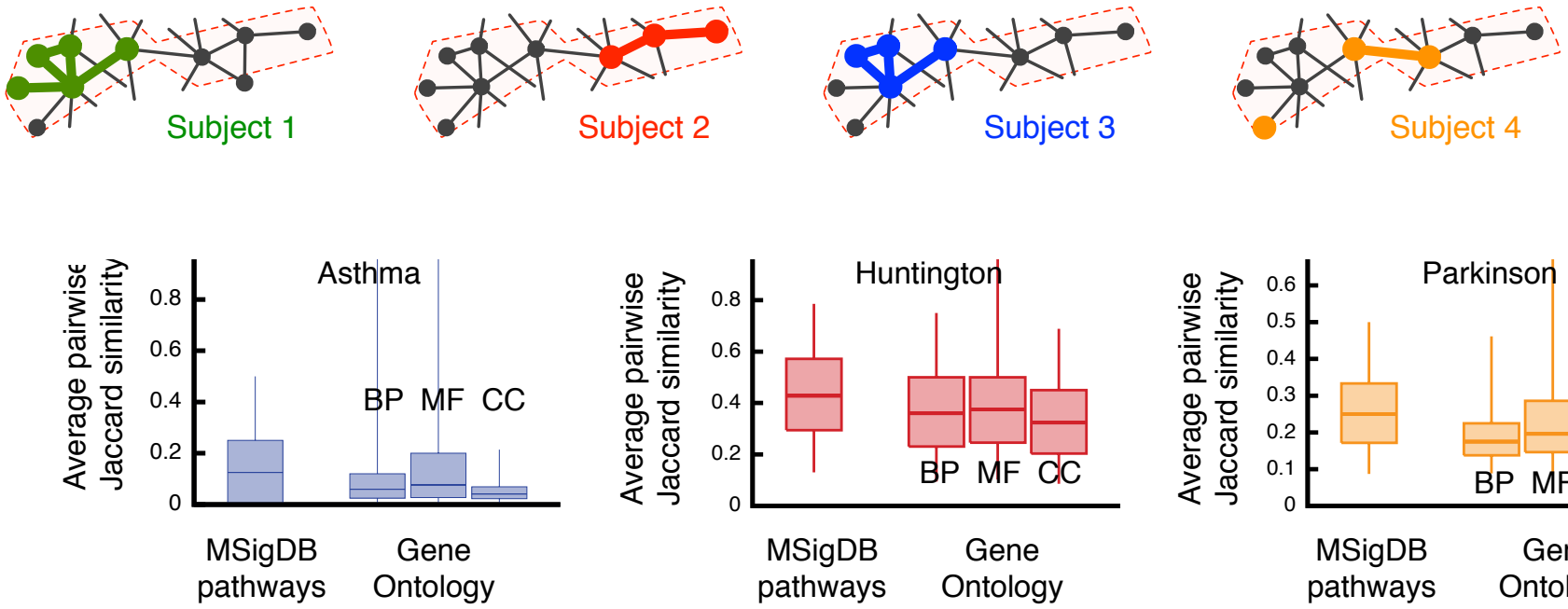
Quantifying the heterogeneity using PEEPs



The overlap between PEEPs of two individuals with the same disease

- is low ($< 30\%$), suggesting high heterogeneity at the transcription level
- is higher than the overlap between the PEEPs of healthy subjects

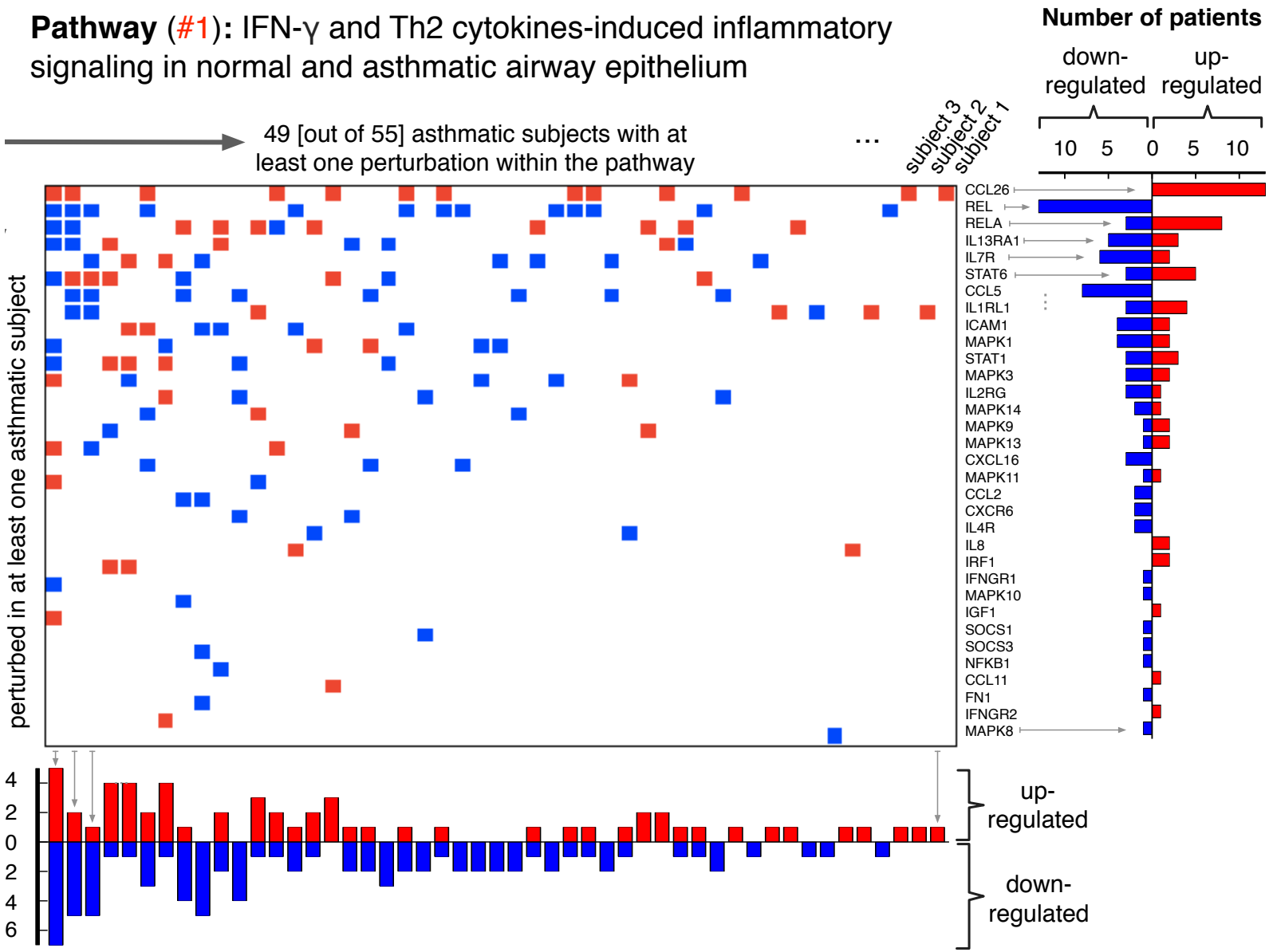
From personalized gene-level signatures to pathway-level signatures



Enrichment of disease-specific pathways in PEEPs (assessed by Fisher's test followed by Bonferroni correction) reveals that

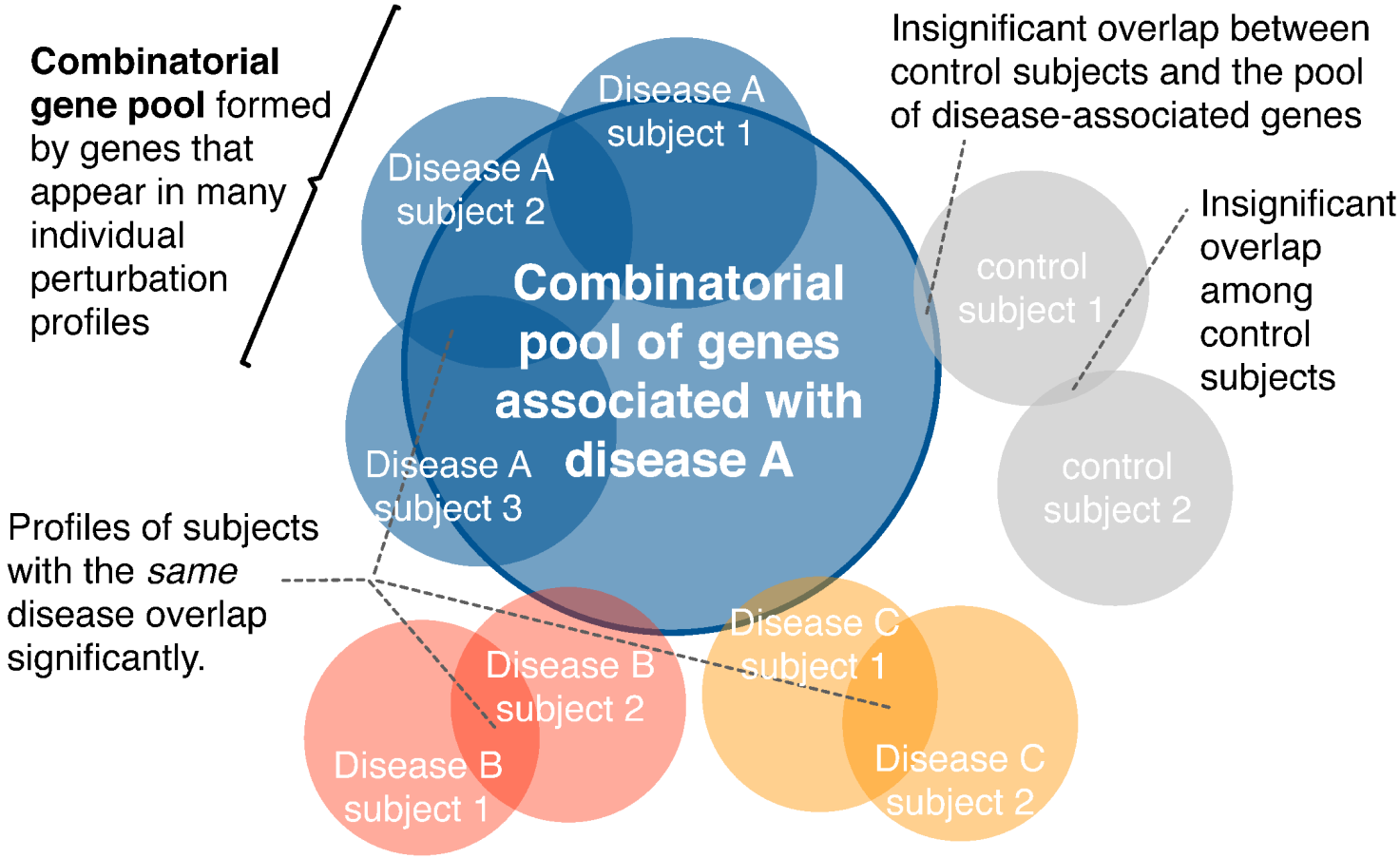
- almost all the individuals show significant perturbations in disease-specific pathways
- the specific perturbations differ greatly across subjects

Personalized pathway-level signatures



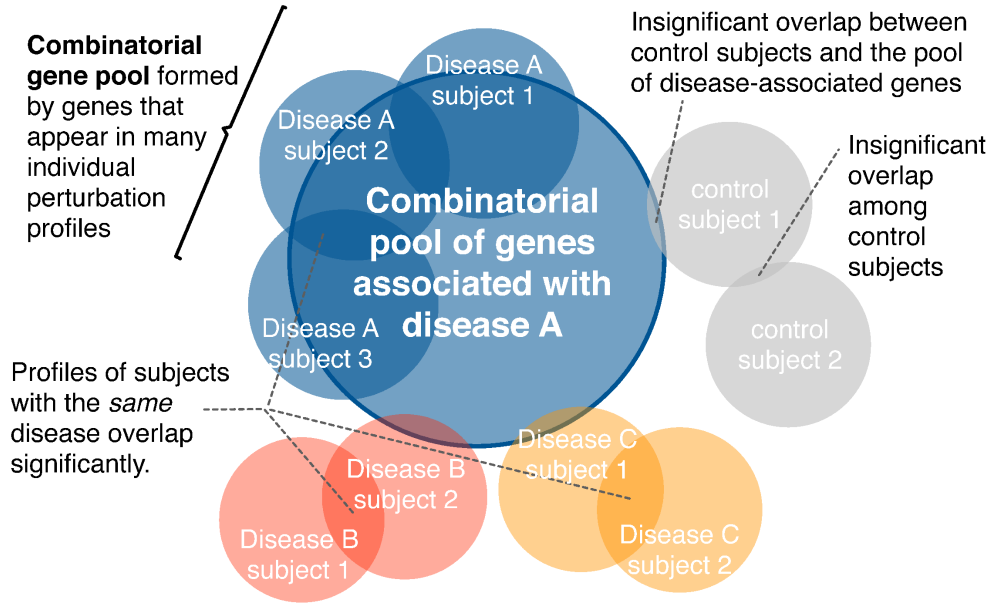
Pathways are perturbed in different ways in different patients

Disease module hypothesis



- Different perturbations within certain molecular pathways lead to the same disease
- Genes that are perturbed in a significant fraction of the case subjects define the disease module

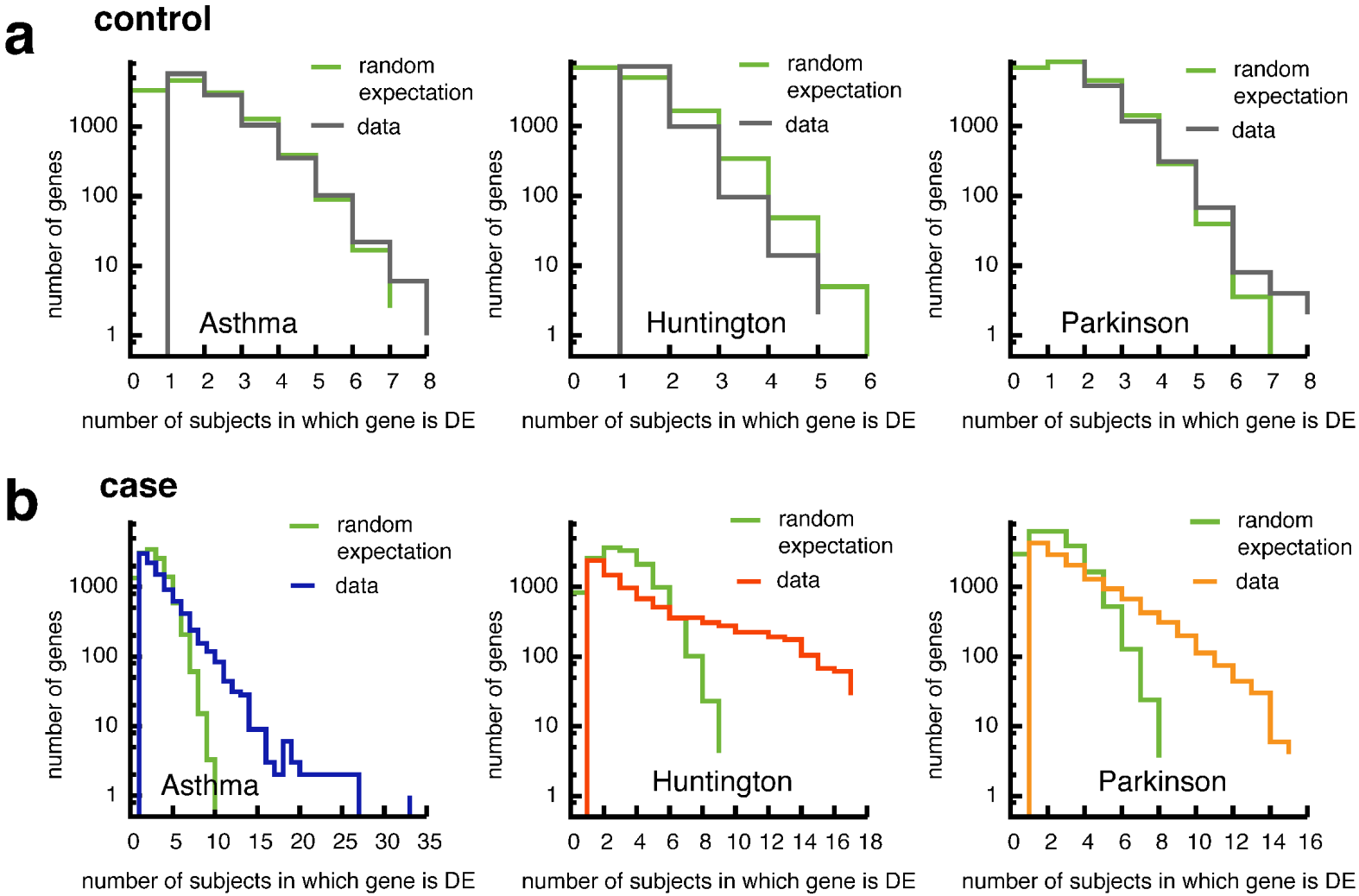
Can PEEPs predict disease status?



Need to make sure that

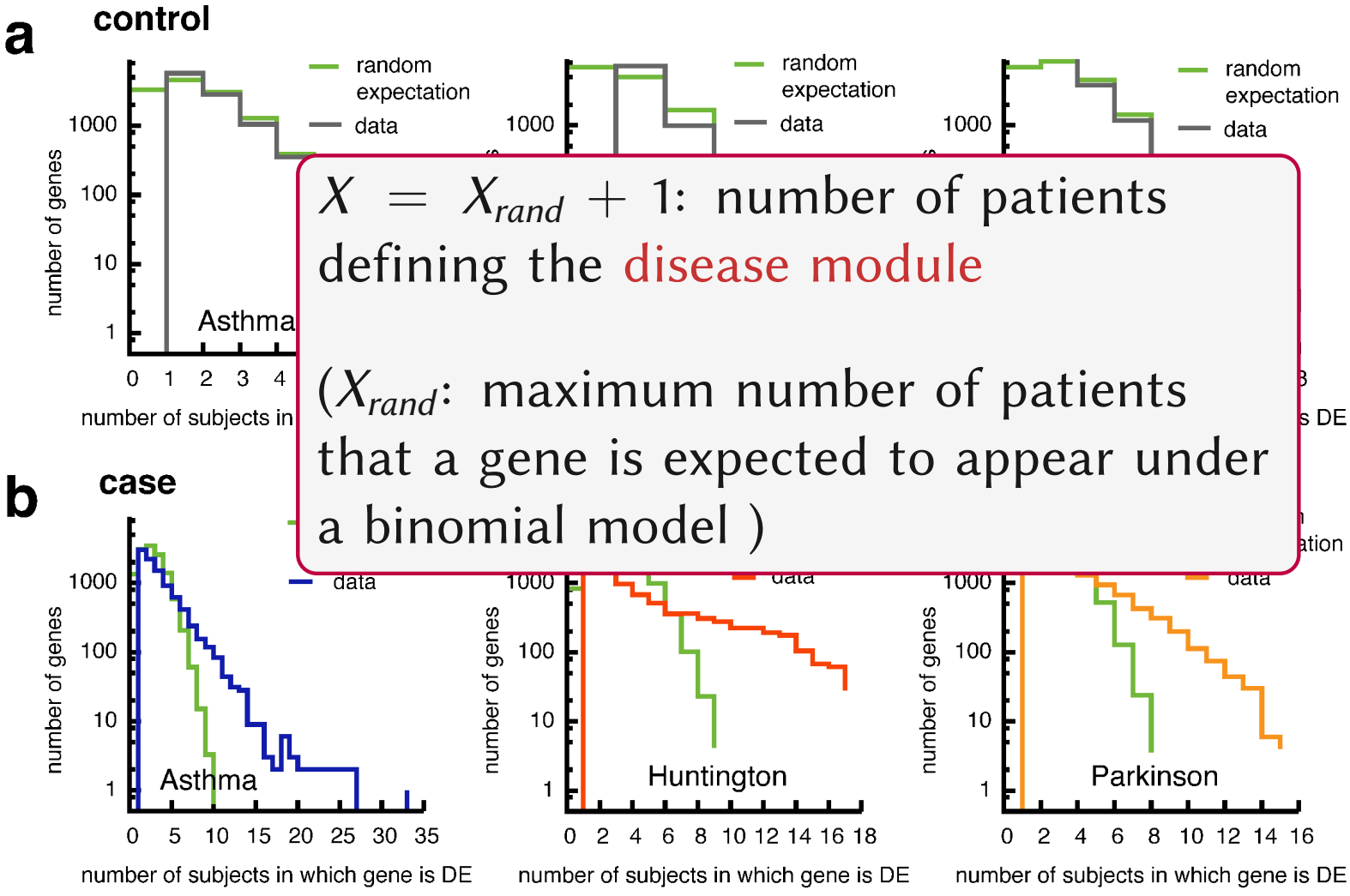
- PEEPs are not an artifact of technical or biological noise in transcriptomics data
- Disease module captures disease heterogeneity observed across patients

Can PEEPs predict disease status?



The number of shared genes between case subjects significantly exceeds the random expectation

Can PEEPs predict disease status?

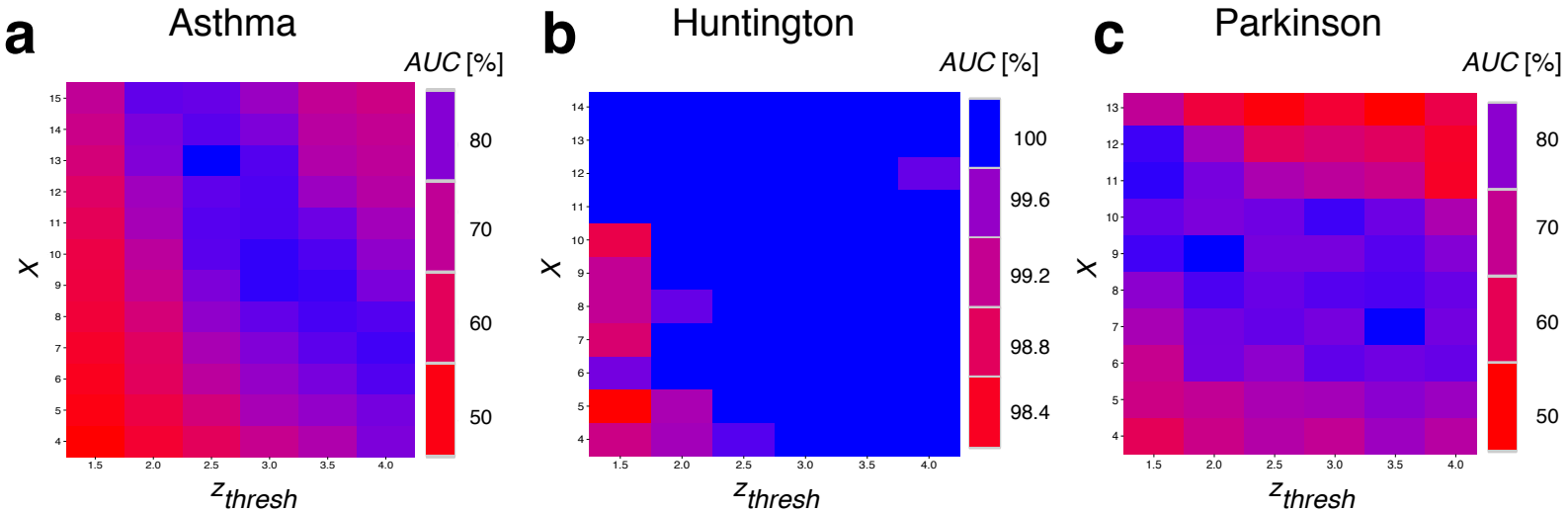


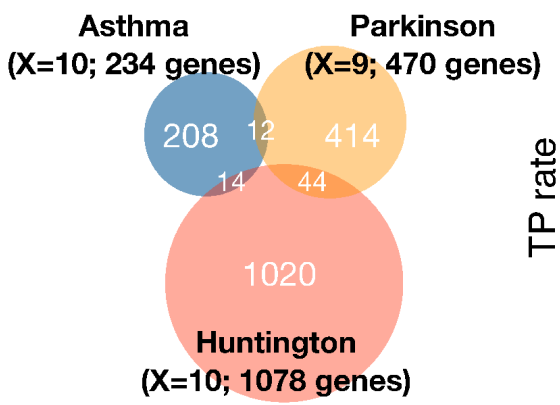
$X = X_{rand} + 1$: number of patients defining the disease module

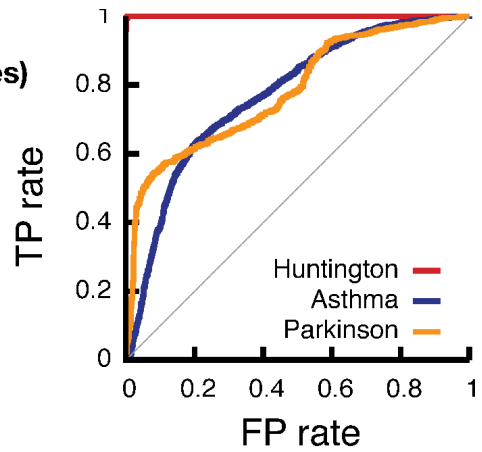
(X_{rand}: maximum number of patients that a gene is expected to appear under a binomial model)

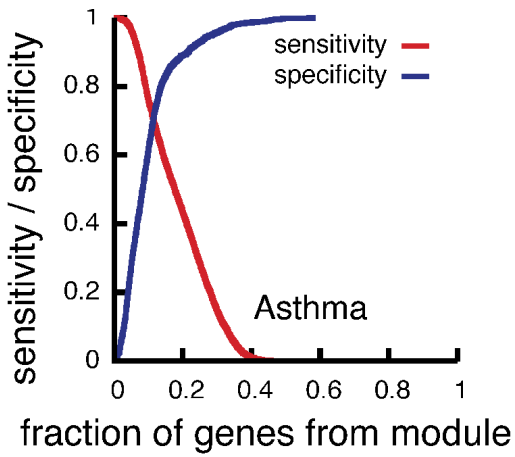
The number of shared genes between case subjects significantly exceeds the random expectation

Predicting diseases from PEEPs









Conclusions

- High heterogeneity across patients at the transcription level
- PErsonalized Expression Profiles (PEEPs)
  - quantify heterogeneity
  - define a pool of perturbed genes across patients
  - predict disease status

R package for generating PEEPs is available at
github.com/emreg00/pepper

Acknowledgements

VIENNA SCIENCE
AND TECHNOLOGY FUND

Agència
de Gestió
d'Ajuts
Universitaris
i de Recerca

P50-HG004233, U01-HG001715, UO1-HG007690 from NHGRI

PO1-HL083069, R37-HL061795, RC2-HL101543, U01-HL108630 from NHLBI

VRG15-005 from WWTF

Beatriu de Pinós Fellowship from AGAUR

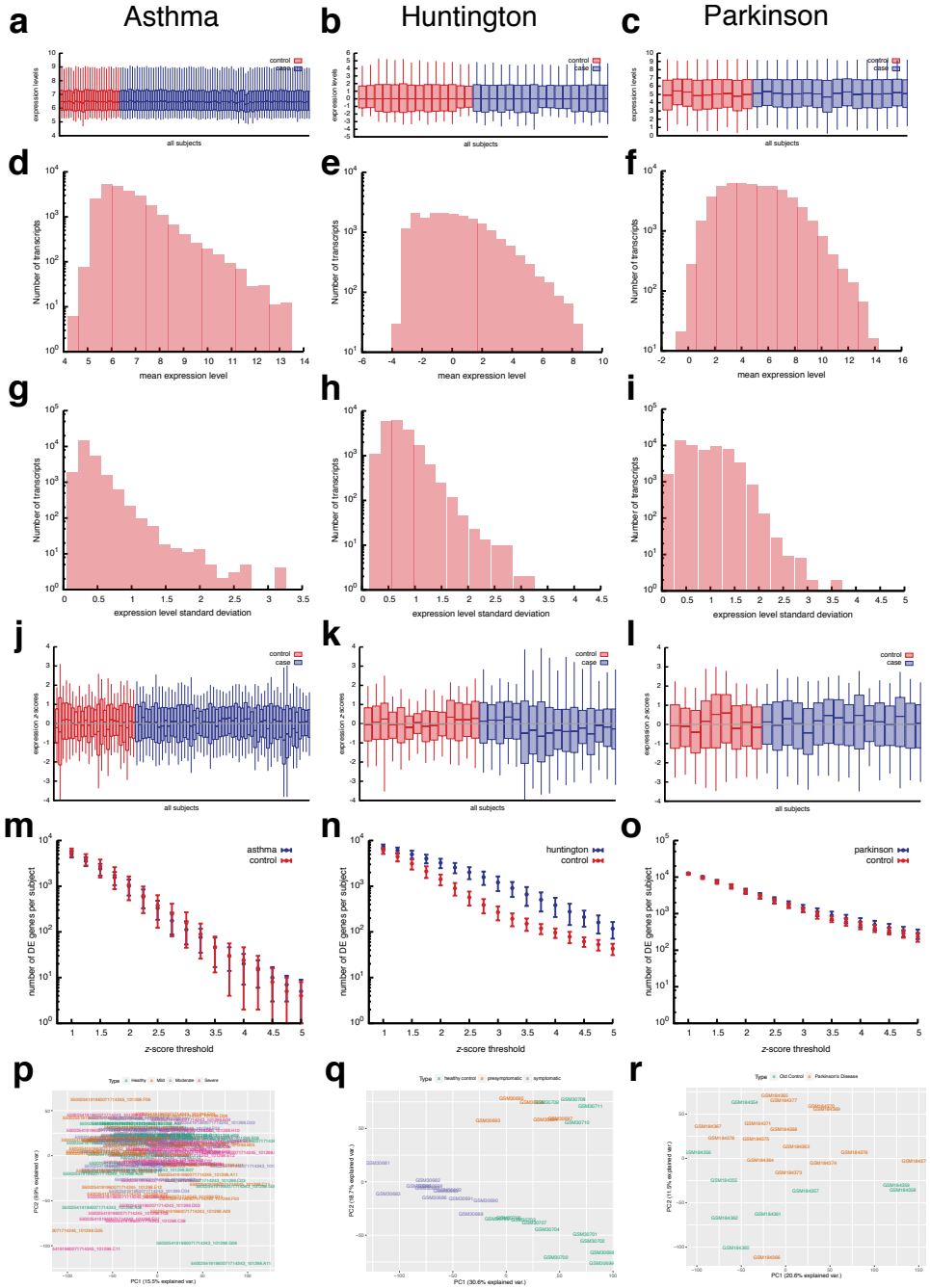
Integrating personalized gene expression profiles into predictive disease-associated gene pools

Jörg Menche[1,2,3], Emre Guney[1,4], Amitabh Sharma[1,4,5], Patrick J. Branigan[6], Matthew J. Loza[6], Frédéric Baribaud[6], Radu Dobrin[6] and Albert-László Barabási[1,2,4,5]

Marc Santolini & Alan Karma

R package for generating PEEPs is available at
github.com/emreg00/pepper

Preprocessing and normalization of transcriptomics data



Defining disease modules

X, the number of patients defining the disease module is given by

$$X = X_{rand} + 1$$

where X_{rand} given by

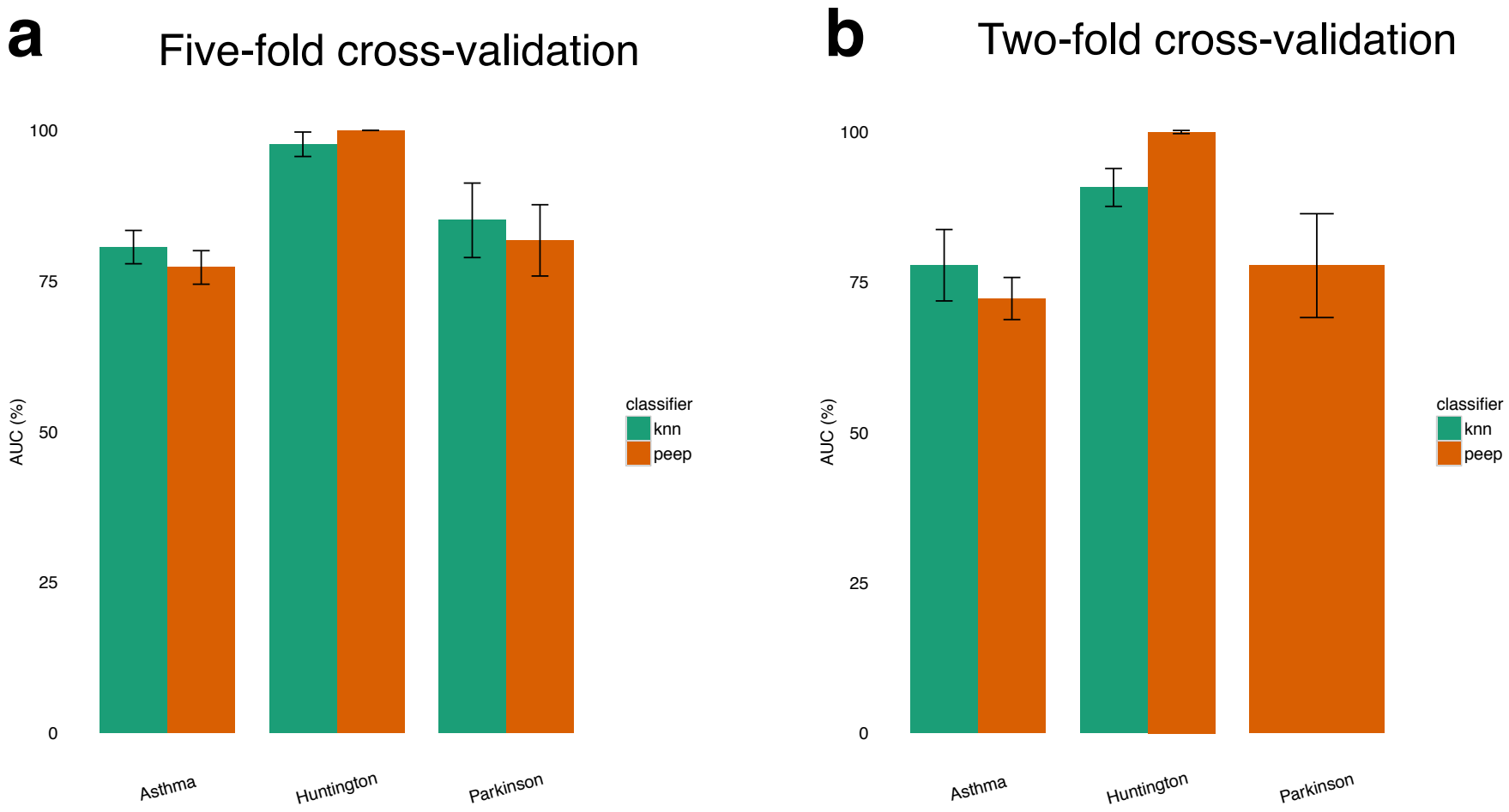
$$argmax_{X_{rand}}\left(\sum_{k=X_{rand}}^{n} G * f(k; n, p) < 1\right)$$

and

$$f(k; n, p) = Pr(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

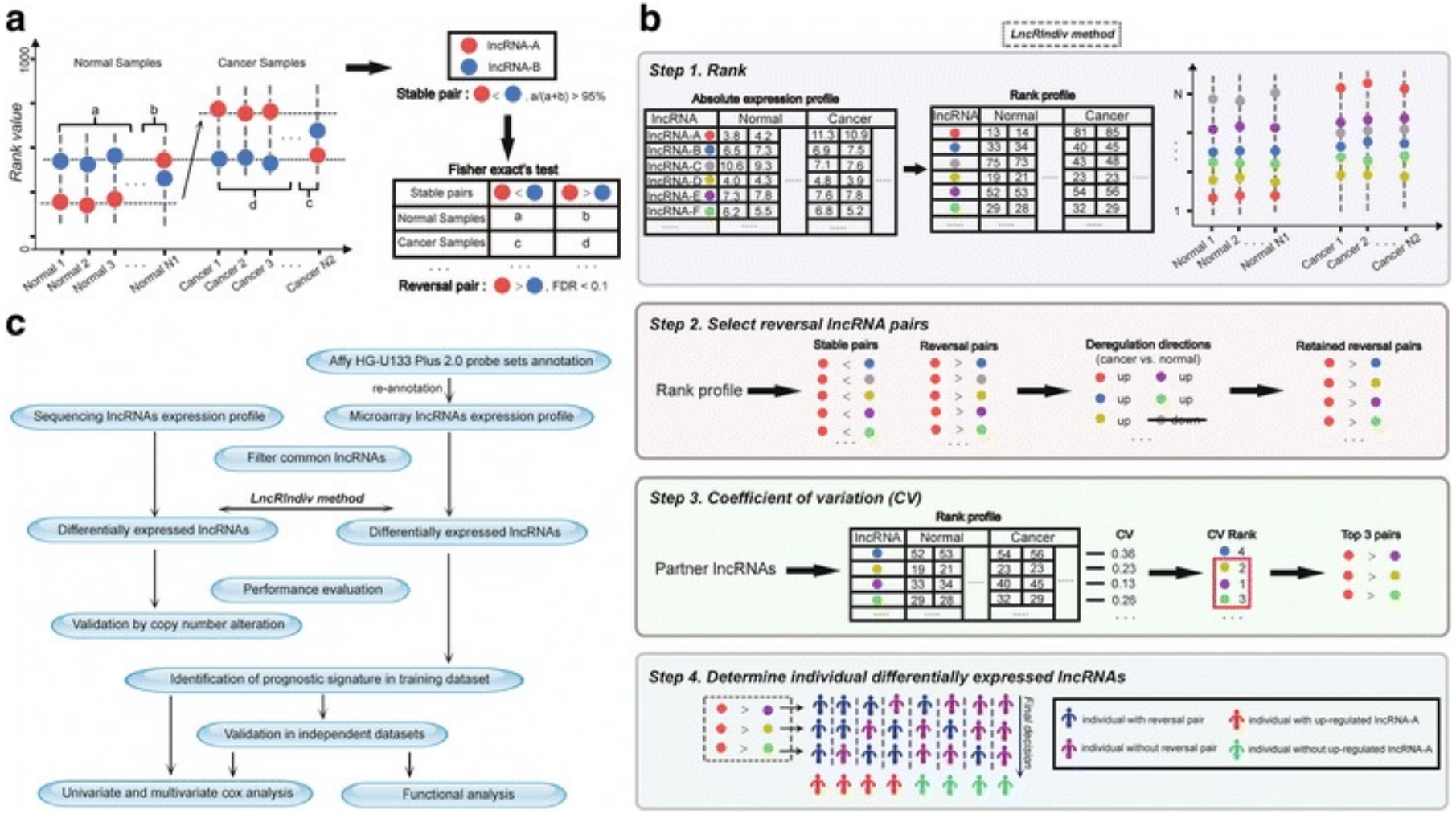
corresponding to the probability for a gene to be perturbed in exactly k out of n subjects given that each subject has g perturbed genes on average (out of all possible G genes in the expression profiling platform) and thus, $p = g/G$

PEEP vs K-nearest-neighbor based predictions



- Results with $K = 15$ (maximizes accuracy among 3, 5, 10, 15, 20)
- AUCs are calculated using a repeated cross validation approach (100 repetitions)

Rank-based approaches for calculating individual perturbation profiles



RankComp *Wang et al., 2015, Bioinformatics*

LncRIndiv *Peng et al., 2017, Mol Cancer*