eurac research

GWAS foundations

Cristian Pattaro

Eurac Research, Institute for Biomedicine, Bolzano, Italy

Johannesburg, 28 January 2026

cristian.pattaro@eurac.edu

@cpattaro.bsky.social

eurac research

Group of Biostatistics & Epidemiology

Fabiola Del Greco
senior researcher

Luisa Foco
senior researcher

Roberto Melotti
senior researcher

Martin Gögele
senior researcher

Rebecca Lundin
senior researcher

Sara Lago
senior researcher

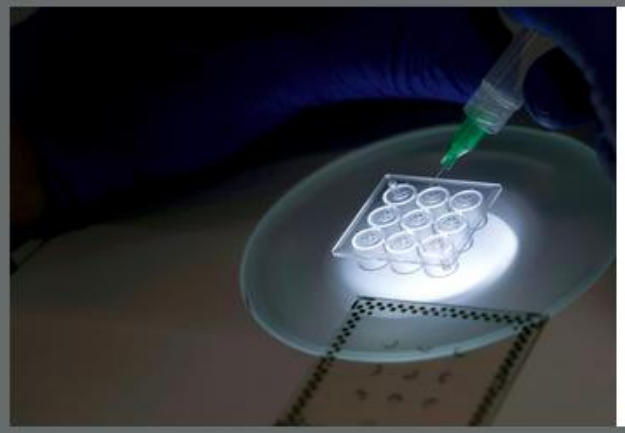
Nadia Alipour
Teheran **2026**
Seal of Excellence

Mousumi Banerjee
University of Michigan
Dept of Biostatistics, **2026**

Ping (Stella) Wang
PhD candidate
Northwestern University, Chicago
Visiting **2026**

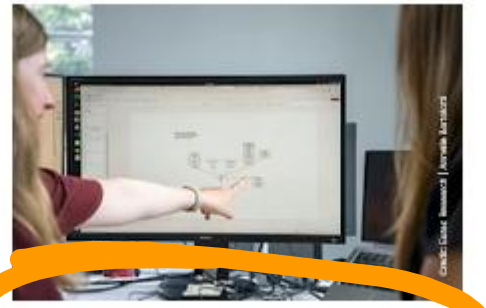
INSTITUTE FOR
Biomedicine

Health Data Science



Translational Biology

Biomedical Informatics



Biostatistics & Epidemiology

Computational Genomics

Biomedical Ethics and Policy

Genome-wide association studies

General concepts. Hypothesis-free vs hypothesis-driven approaches.

Complex traits and genetic models. CKD phenotypes.

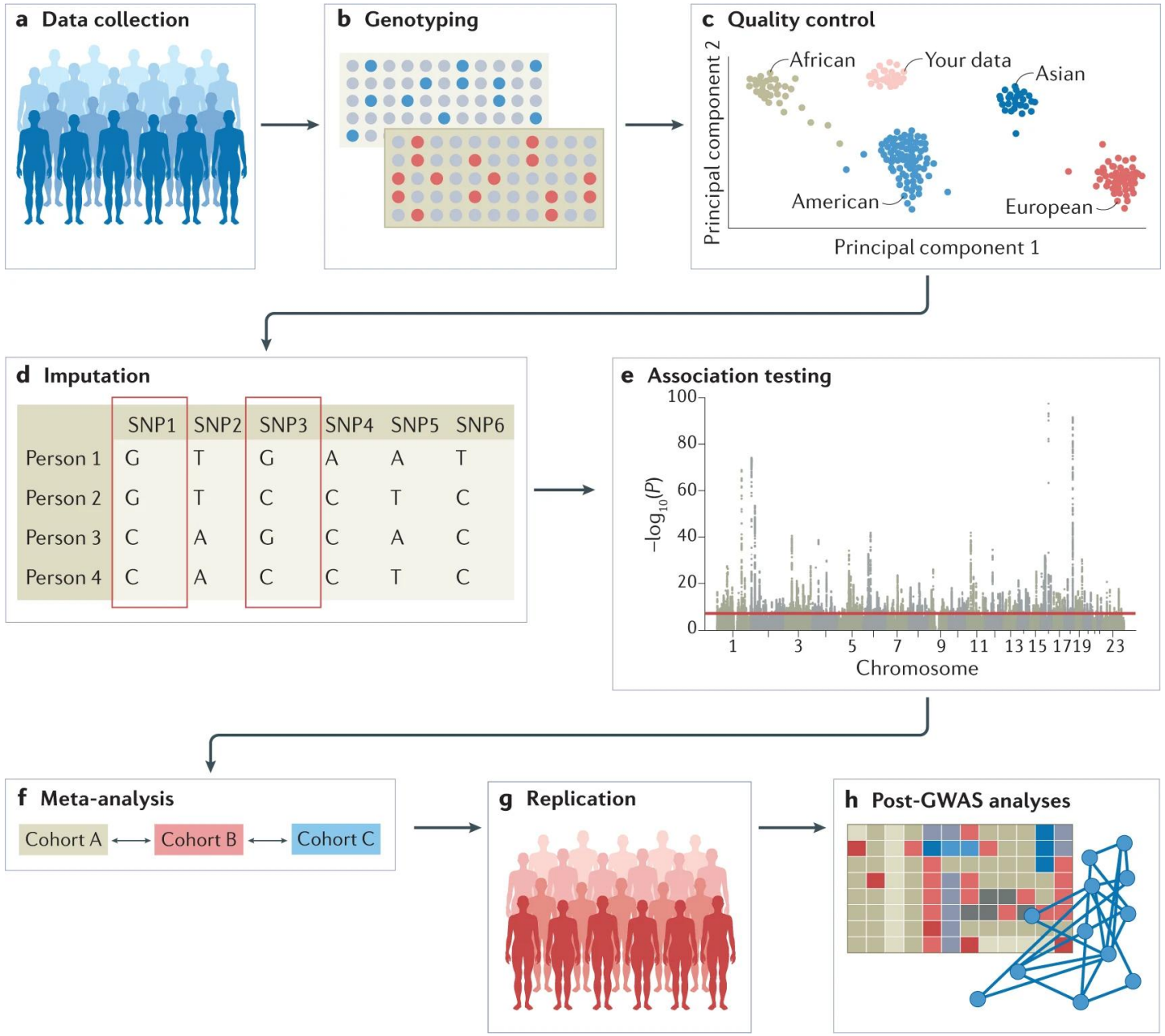
Primer | Published: 26 August 2021

Genome-wide association studies

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen & Danielle Posthuma

Nature Reviews Methods Primers **1**, Article number: 59 (2021) | Cite this article

Free at https://www.nature.com/articles/s43586-021-00056-9



Uffelmann et al, Nat Rev Methods Primers 2021; 1:59

Aim

To identify

genotypes – phenotypes associations

by testing for

differences in the allele frequency of genetic variants

between individuals

who are ancestrally similar but differ phenotypically.

Uffelmann et al, Nat Rev Methods Primers 2021; 1:59

Underlying philosophy

GWAS are hypothesis-free

any variant in the genome is given the same a priori probability

to results positively associated with the outcome

The replication crisis

Full text: https://www.gimjournal.org/article/S1098-3600(21)02956-7/fulltext

- Until 2000, genetic association studies were essentially candidate gene studies
- Authors performed extensive review of genetic association studies
- They identified >600 reported positive associations between common gene variants and disease
- (*these associations, if correct, would have had tremendous importance for the prevention, prediction, and treatment of most common diseases*)
- However, most reported associations didn't prove robust: of 166 putative associations studied at least 3 times, only 6 have been consistently replicated.

The replication crisis

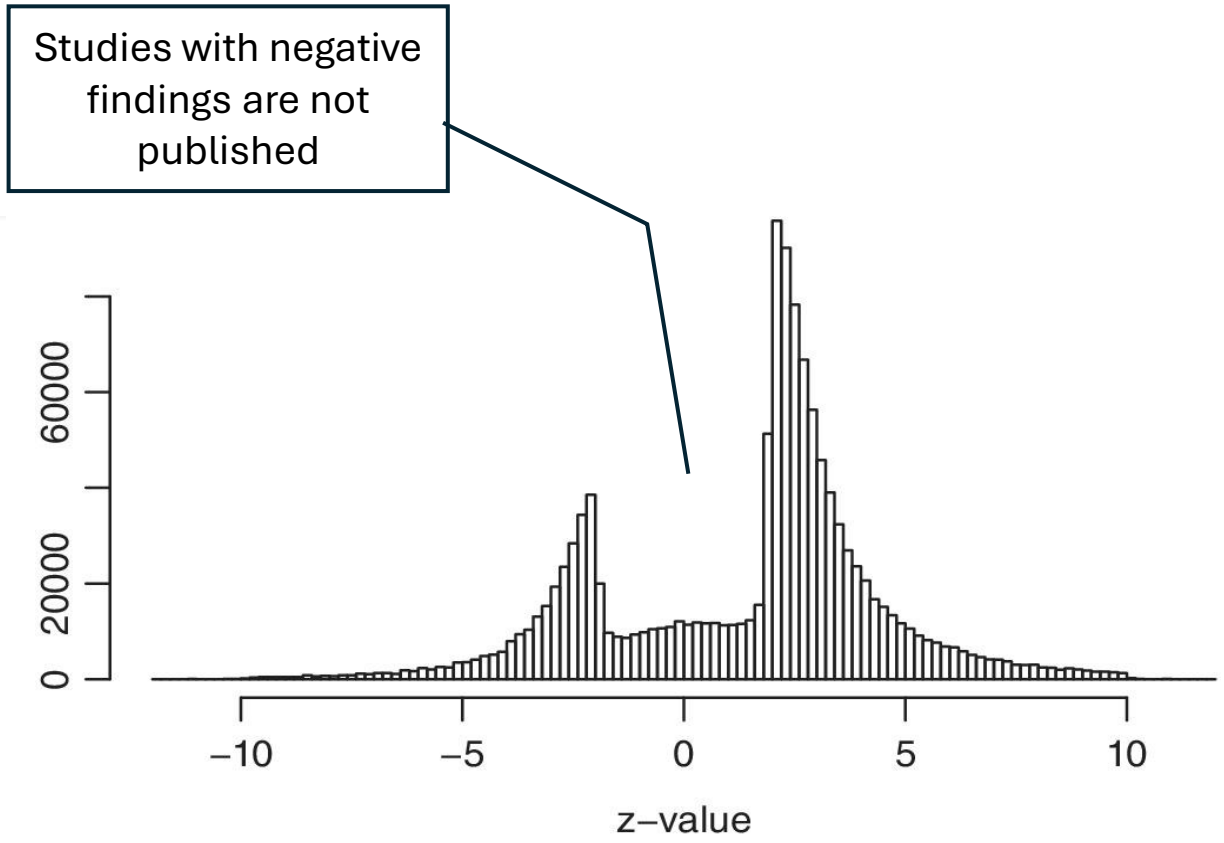
Why Most Discovered True Associations Are Inflated

Ioannidis, John P. A.

Author Information

Epidemiology: September 2008 - Volume 19 - Issue 5 - p 640-648

→ The Winner's course
specially severe when power is low)



statistica neerlandica

ORIGINAL ARTICLE | Open Access

The significance filter, the winner's curse and the need to shrink

Erik W. van Zwet | Eric A. Cator

First published: 22 March 2021 | https://doi.org/10.1111/stan.12241 | Citations: 1

Statistica Neerlandica, Volume: 75, Issue: 4, Pages: 437-452,
First published: 22 March 2021, DOI: (10.1111/stan.12241)

Consider a **sample of N study participants**

In each participant

- we measured a **quantitative trait or diagnosed a disease of interest**
- we **genotyped a few million SNPs**

- is any SNP associated with the trait of interest?
- statistical significance penalty: 0.05 / 1,000,000 = 5e-08

Omnigenic model

Cell
A Cell Press journal

PERSPECTIVE · Volume 169, Issue 7, P1177-1186, June 15, 2017 · *Open Archive*

Download Full Issue

An Expanded View of Complex Traits: From Polygenic to Omnigenic

Evan A. Boyle [1] · Yang I. Li [1] · Jonathan K. Pritchard [1,2,3]

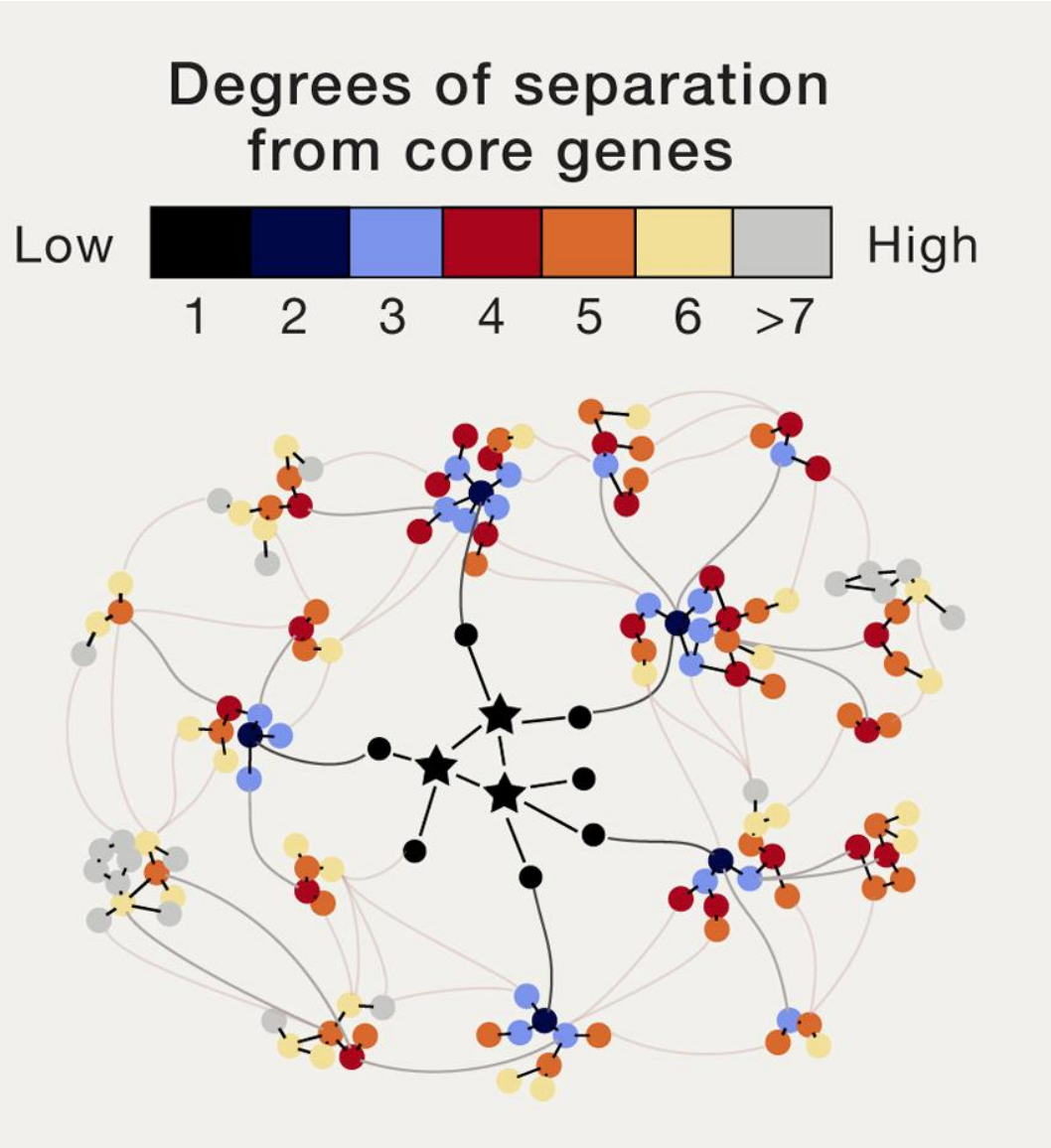
Omnigenic model

Most genes affect disease risk through highly connected cellular networks.

For any given disease phenotype:

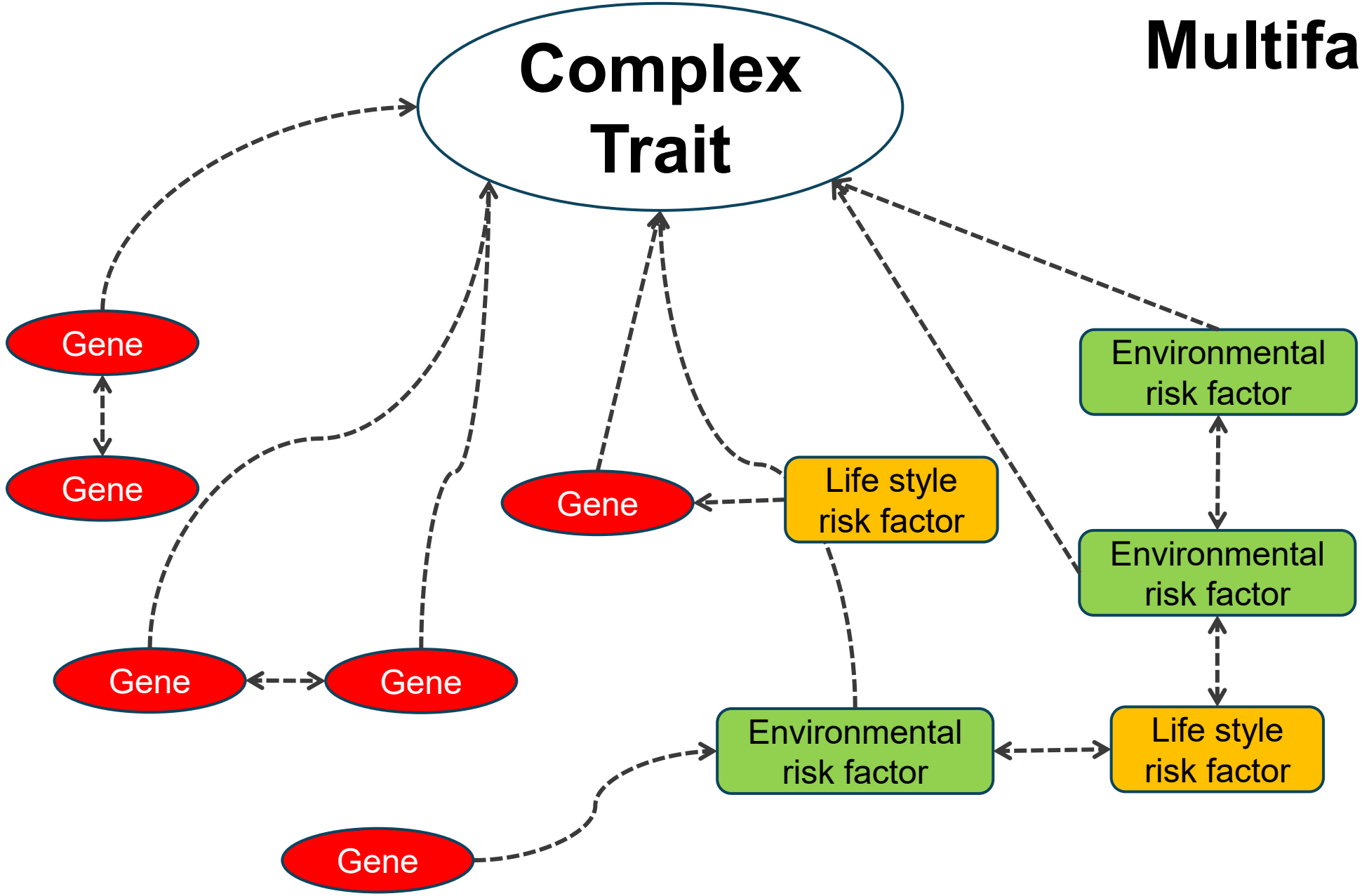
1. a limited number of genes have direct effects on disease risk
2. most expressed genes are only a few steps from the nearest core gene and thus may have non-zero effects on disease.

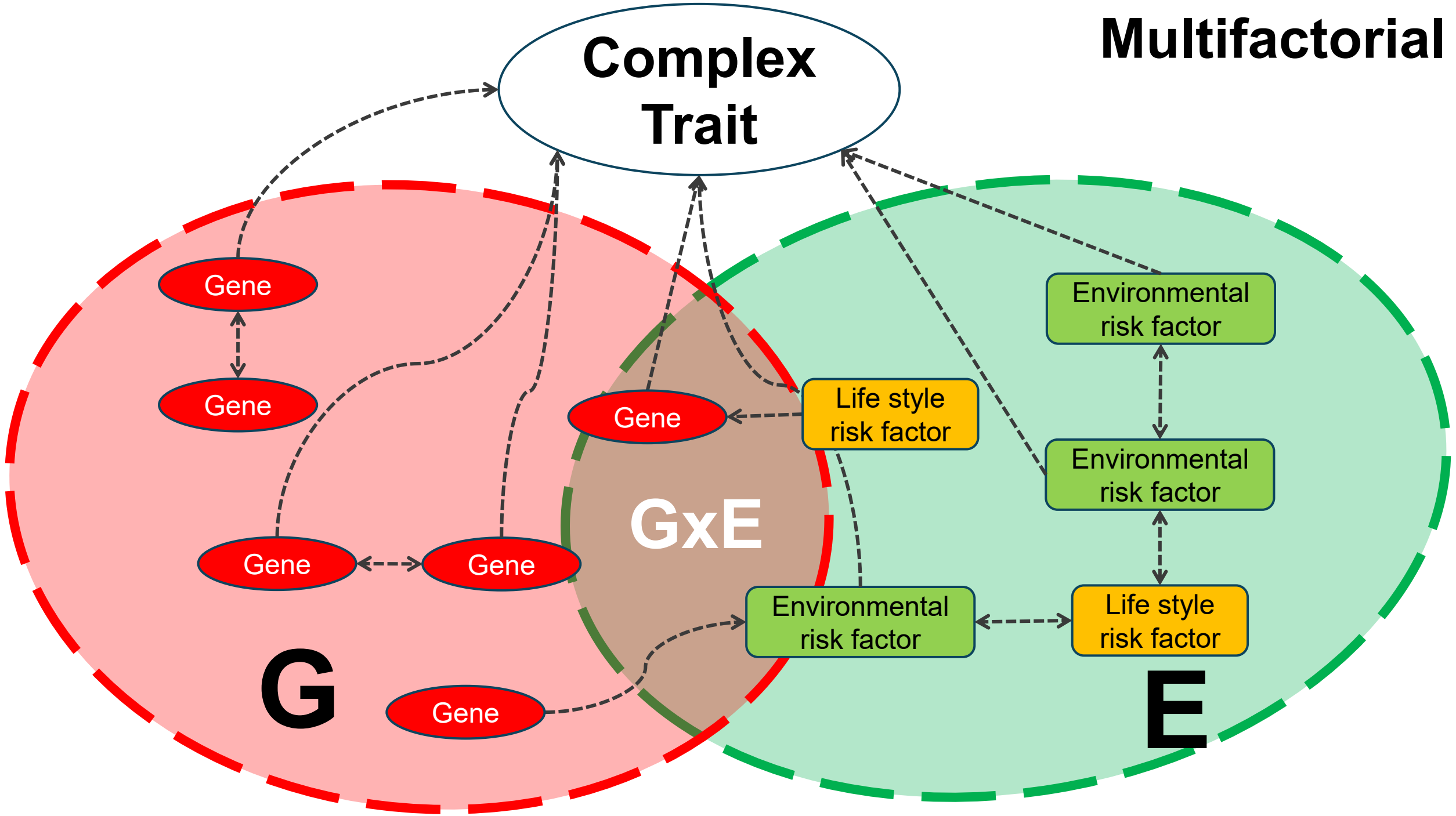
Since core genes only constitute a tiny fraction of all genes, most heritability comes from genes with indirect effects.



Boyle et al, Cell 2017

Multifactorial

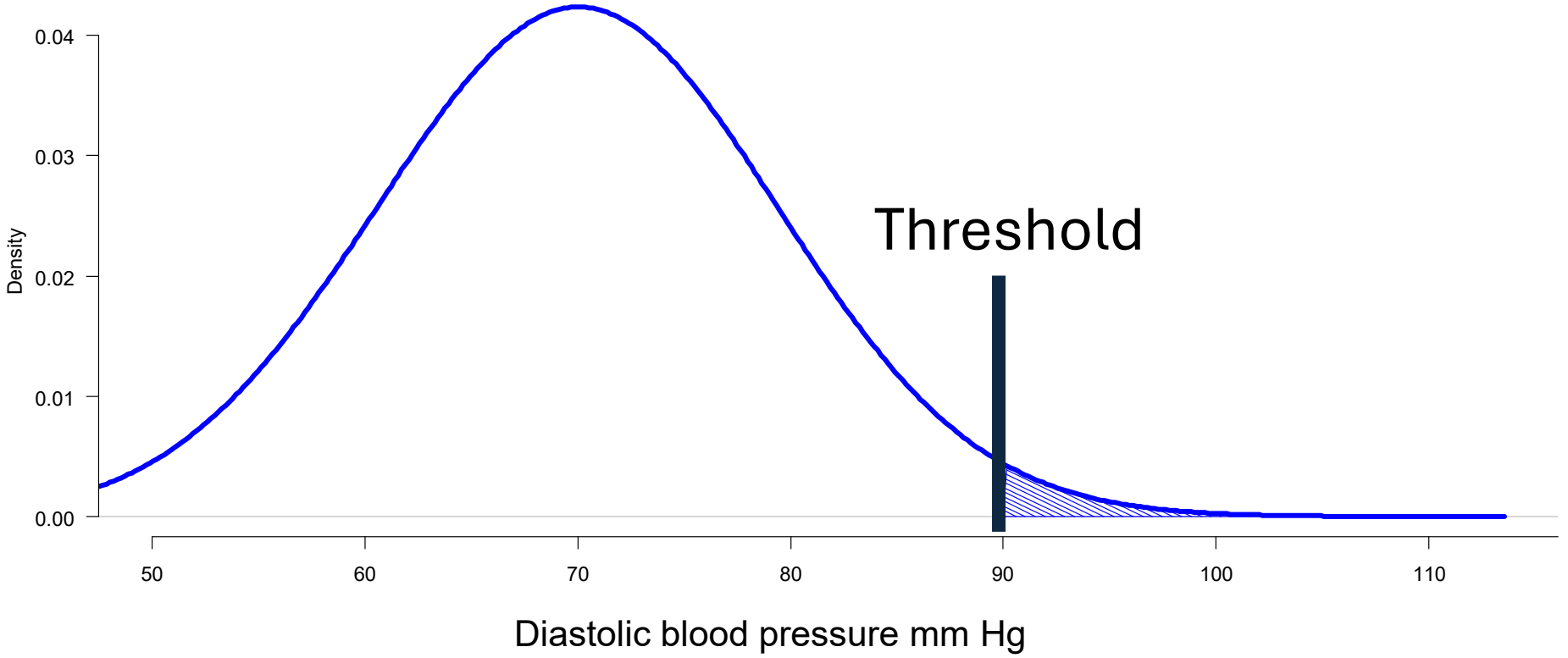



Multifactorial
Complex Trait
Gene
Gene
Gene
Gene
Gene
Gene
G
GxE
Life style risk factor
Environmental risk factor
Environmental risk factor
Environmental risk factor
Environmental risk factor
Life style risk factor
E

A complex disease can be seen as an adverse manifestation following extreme levels of one or more underlying continuous complex traits

We can establish the relationship

Complex disease → **Complex trait**

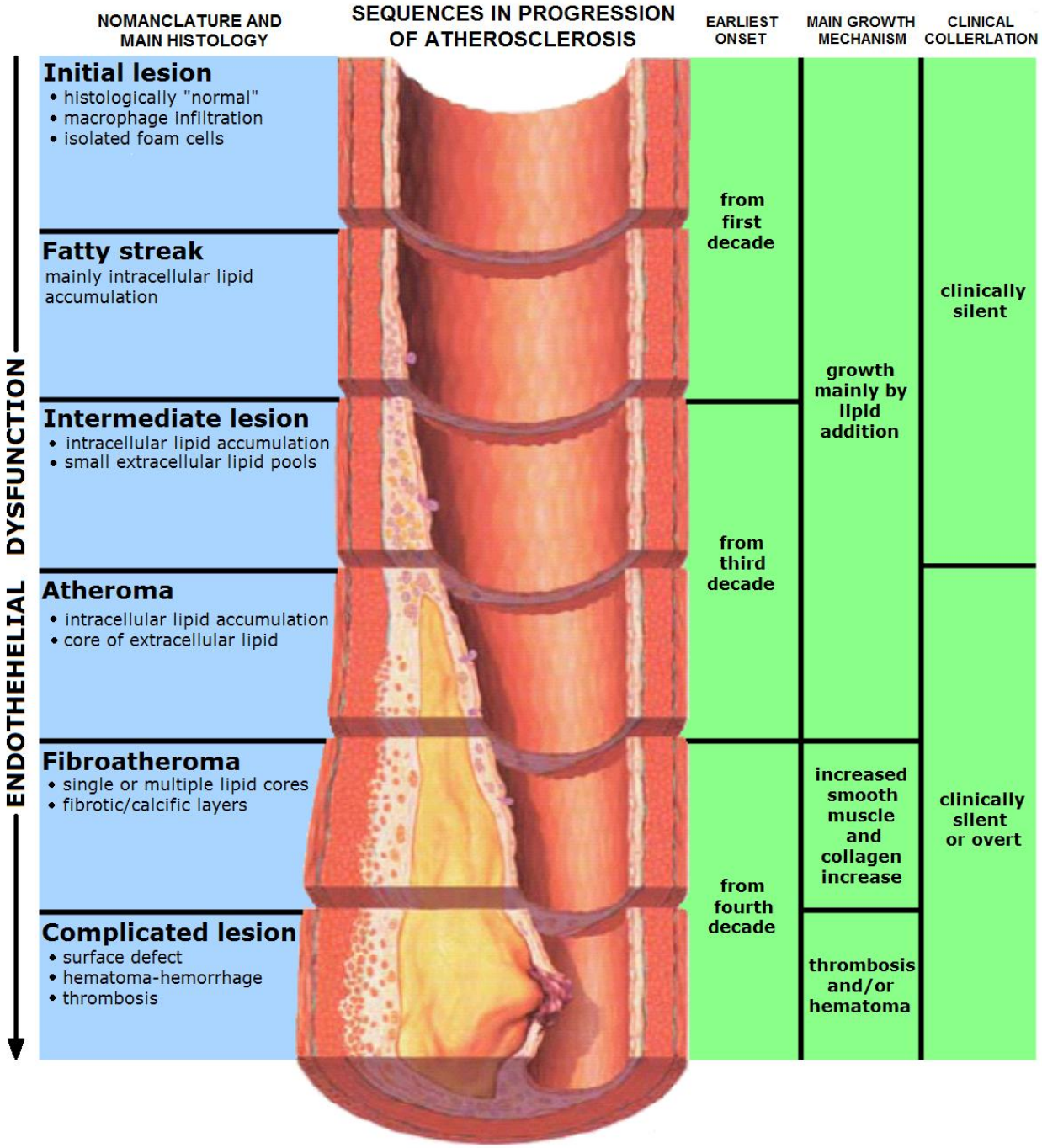

0.04
0.03
0.02
0.01
0.00
Density
Threshold
50
60
70
80
90
100
110
Diastolic blood pressure mm Hg

We can probably treat *most binary as liability threshold phenotypes*

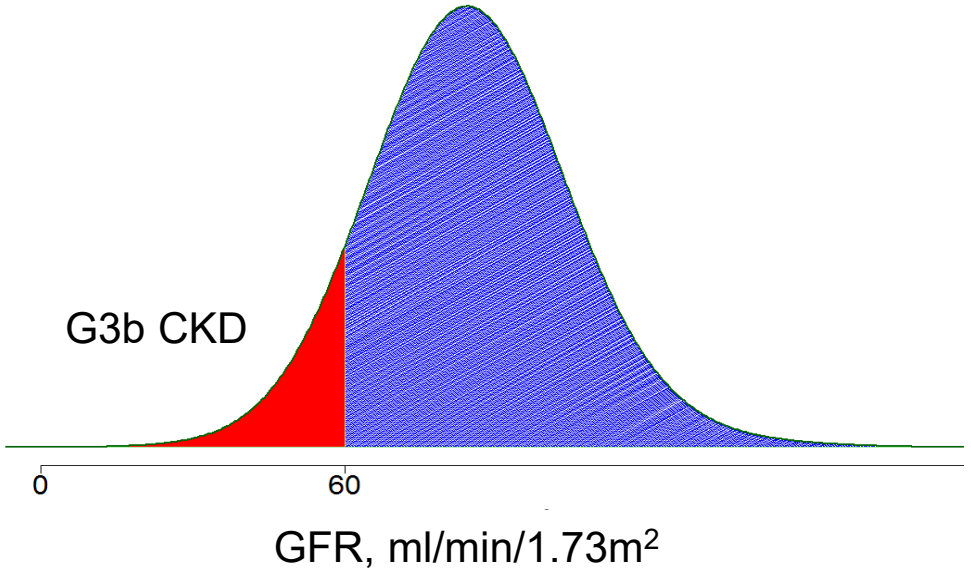
…and they even have some quantitative measurable biomarkers that approximates the underlying latent phenomenon

Example:

atherosclerosis = thickness of blood vessel walls = measurable. Progressive thickening of the walls, due to accumulation of fatty components such as cholesterol or tryglicerides → **thromboembolism**.



https://en.wikipedia.org/wiki/Atherosclerosis#/media/File:Endo_dysfunction_Athero.PNG



Prognosis of CKD by GFR and albuminuria categories: KDIGO 2012

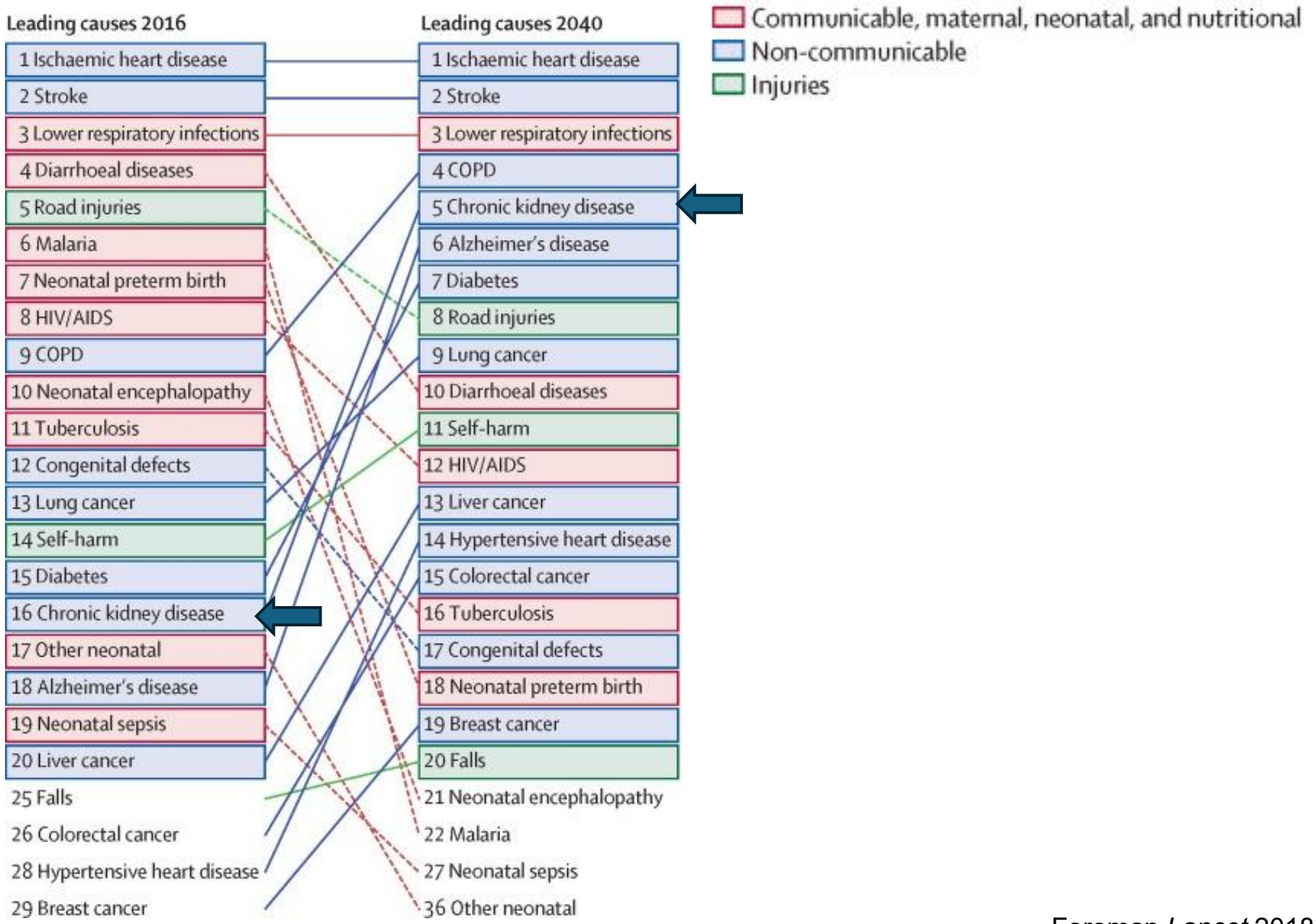
GFR categories (ml/min/1.73 m²) Description and range				Persistent albuminuria categories Description and range		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				< 30 mg/g < 3 mg/mmol	30–300 mg/g 3–30 mg/mmol	> 300 mg/g > 30 mg/mmol
	G1	Normal or high	≥ 90			
	G2	Mildly decreased	60–89			
	G3a	Mildly to moderately decreased	45–59			
	G3b	Moderately to severely decreased	30–44			
	G4	Severely decreased	15–29			
	G5	Kidney failure	< 15			

Overall	Urine albumin-creatinine ratio, mg/g					Urine albumin-creatinine ratio, mg/g				
eGFRcr	<10	10–29	30–299	300–999	1000+	<10	10–29	30–299	300–999	1000+
	All-cause mortality: 82 cohorts 26 444 384 participants; 2 604 028 events					Myocardial infarction: 64 cohorts 22 838 356 participants; 451 063 events				
105+	1.6	2.2	2.9	4.3	5.8	1.1	1.4	2.0	2.7	3.8
90–104	ref	1.3	1.8	2.6	3.1	ref	1.3	1.6	2.2	3.2
60–89	1.0	1.3	1.7	2.2	2.8	1.1	1.3	1.6	2.2	3.1
45–59	1.3	1.6	2.0	2.4	3.1	1.4	1.7	2.0	2.8	3.7
30–44	1.8	2.0	2.5	3.2	3.9	1.9	2.0	2.4	3.2	4.3
15–29	2.8	2.8	3.3	4.1	5.6	2.7	3.1	3.1	4.2	5.1
<15	4.6	5.0	5.3	6.0	7.0	4.6	5.6	4.8	6.0	6.0
	Cardiovascular mortality: 76 cohorts 26 022 346 participants; 776 441 events					Stroke: 68 cohorts 24 746 436 participants; 461 785 events				
105+	1.4	2.0	3.0	4.1	5.4	1.2	1.6	2.2	3.1	4.3
90–104	ref	1.3	1.9	2.7	3.6	ref	1.3	1.6	2.4	3.1
60–89	1.0	1.4	1.7	2.4	3.2	1.1	1.3	1.7	2.2	3.0
45–59	1.4	1.7	2.2	2.8	3.8	1.4	1.6	1.9	2.3	2.9
30–44	2.0	2.3	2.8	3.7	4.6	1.6	1.7	2.0	2.4	3.0
15–29	3.2	3.1	3.5	5.0	6.5	1.8	2.1	2.1	2.7	3.0
<15	6.1	6.4	6.4	7.3	8.2	3.2	2.8	2.9	3.2	3.8
	Kidney failure with replacement therapy: 57 cohorts 25 466 956 participants; 158 846 events					Heart failure: 61 cohorts 24 603 016 participants; 1 132 443 events				
105+	0.5	1.2	2.9	7.7	25	1.2	1.7	2.7	4.2	6.9
90–104	ref	1.8	4.3	12	43	ref	1.3	2.0	2.8	4.2
60–89	2.3	4.9	10	27	85	1.1	1.4	1.9	2.7	4.2
45–59	13	19	37	89	236	1.6	1.8	2.4	3.4	5.0
30–44	50	58	115	240	463	2.2	2.5	3.1	4.2	6.5
15–29	283	301	443	796	1253	3.6	3.5	4.1	5.8	8.1
<15	770	1040	1618	2297	2547	5.1	5.7	5.8	7.9	9.9
	Acute kidney injury: 49 cohorts 23 914 614 participants; 1 408 929 events					Atrial fibrillation: 50 cohorts 22 886 642 participants; 1 068 701 events				
105+	1.0	1.6	2.4	3.7	5.5	1.1	1.3	1.7	2.4	3.5
90–104	ref	1.4	2.1	3.2	5.0	ref	1.2	1.5	1.9	2.3
60–89	1.6	2.2	3.1	4.3	6.7	1.0	1.2	1.4	1.7	2.2
45–59	3.5	4.0	5.1	6.9	9.0	1.2	1.3	1.5	1.8	2.4
30–44	5.6	5.9	6.8	8.6	11	1.4	1.5	1.7	2.0	2.4
15–29	8.3	8.0	8.5	9.9	10	1.9	1.8	2.0	2.6	3.0
<15	8.5	11	7.9	5.5	5.7	2.6	2.5	3.1	3.6	4.2
	Hospitalization: 49 cohorts 25 426 722 participants; 8 398 637 events					Peripheral artery disease: 54 cohorts 24 830 794 participants; 378 924 events				
105+	1.4	1.7	2.1	2.1	2.3	0.9	1.4	1.9	2.8	5.0
90–104	ref	1.1	1.3	1.5	1.7	ref	1.3	1.9	2.8	4.3
60–89	1.0	1.1	1.3	1.5	1.8	1.0	1.3	1.8	2.5	3.8
45–59	1.3	1.3	1.5	1.7	2.1	1.5	1.7	2.1	2.9	4.2
30–44	1.5	1.5	1.6	1.9	2.3	2.0	1.9	2.5	3.6	5.0
15–29	1.8	1.8	1.9	2.4	2.8	3.3	3.3	3.8	5.7	8.1
<15	2.7	2.8	3.0	3.2	3.8	9.1	9.0	9.6	13	14

Stevens et al. **KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease** Kidney Int 105(4) S117-314

those causes that are climbing the ranking the fastest are the **NCDs** that result from the **gradual degradation of an underlying quantitative process**

Leading causes of mortality worldwide



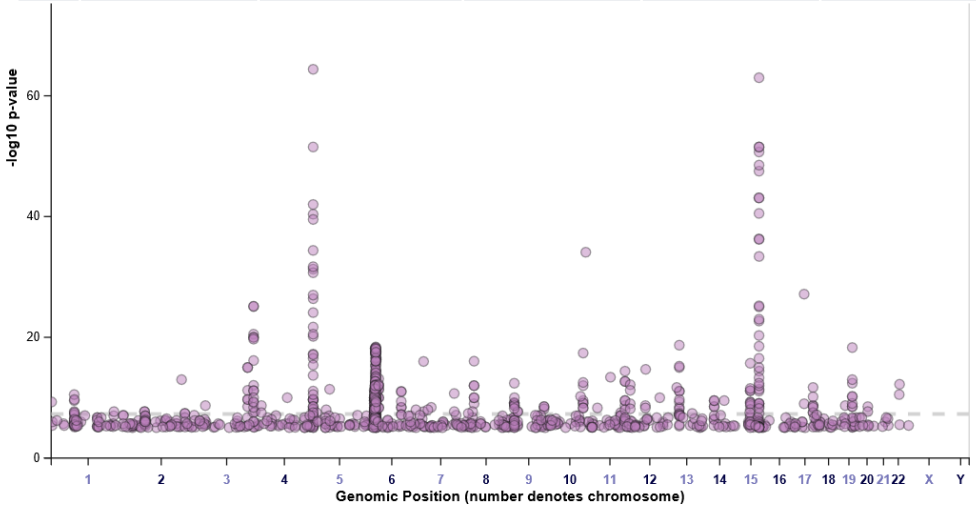
Foreman *Lancet* 2018

Lung cancer

Genetic heritability ~ 8.3% (SE=1.3%)

Pettit, Sci Rep 2021

Chr	Pos	Mapped Gene	RSID	AF	P-VALUE	OR
15	78,565,644	CHRNA5	rs55781567	0.37	3.00E-103	1.30
5	1,285,859	TERT	rs7705526		4.00E-65	1.31
15	78,674,313	CHRNB4	rs11639372	0.41	3.00E-52	1.20
15	78,601,997	CHRNA3	rs1051730		3.00E-52	1.34
10	121,575,416	FGFR2	rs11200014		8.00E-35	
15	78,782,176	ADAMTS7	rs4886591	0.44	4.00E-34	1.17
5	1,320,132	CLPTM1L	rs380286	0.58	2.00E-32	1.15
17	37,741,642	HNF1B	rs12601991		7.00E-28	
5	1,286,401	TERT	rs2736100	0.41	1.00E-27	1.27
3	189,638,472	TP63	rs4488809	0.47	7.00E-26	1.26
13	32,398,489	BRCA2	rs11571833	0.01	2.00E-19	1.83
6	31,466,334	HCP5	rs116822326	0.16	4.00E-19	1.25
19	40,847,202	CYP2A6	rs56113850	0.56	5.00E-19	1.13



Cigarette smoking

Fraction of lung cancer deaths attributable to smoking in UK:
85% among males
80% among females

Br J Cancer. 2011; 105(Suppl 2): S6–13.

Credit: Amanda Eller, NC

Occupational exposures:

Asbestos, coal, silica

The variance component model

$$Y_{KL} = \mu + G_K + E_L + u$$

$$Y_{KL} = \mu + \sum_1^K g_k + \sum_1^L e_l + u$$

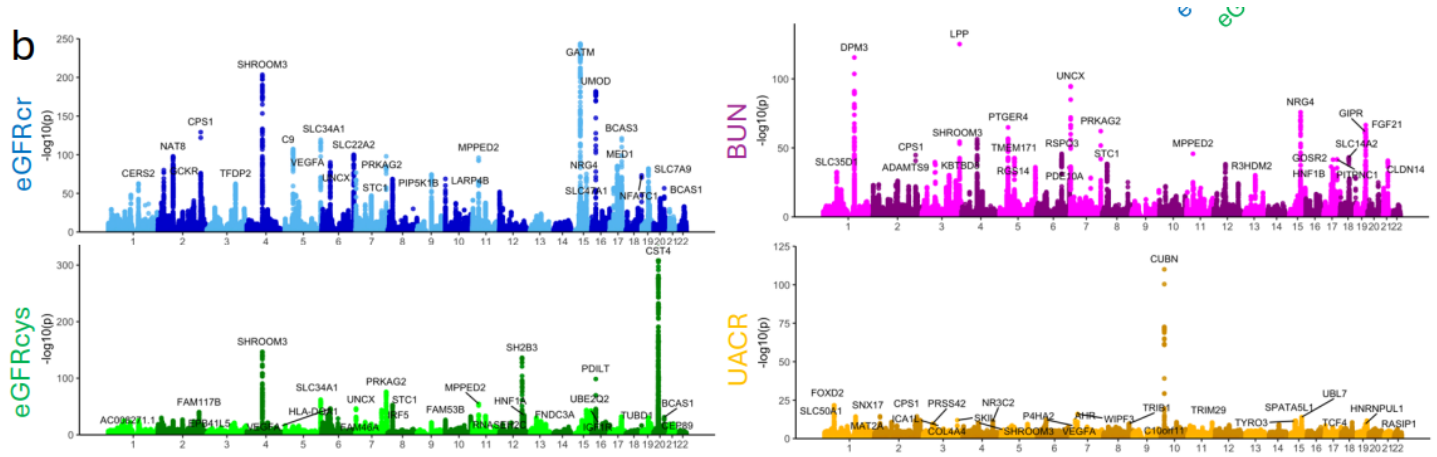
Example: a „multifactorial“ (complex) disease

CKD

A **polygenic** component
(hundreds of genetic variants)

+

An **environmental** component
(several non-genetic risk factors)



Emmett A, et al. medRxiv. 2026
https://www.medrxiv.org/content/10.64898/2026.01.15.26344229v1

- ✓ salt rich diet
- ✓ infections
- ✓ toxicity
- ✓ dehydration
- ✓ etc.

If genotypes and environmental factors are independent,

the total variance of Y can be written as:

$$Var(Y) = \sigma^2 = \sigma_G^2 + \sigma_E^2 + \sigma_u^2$$

Broad sense heritability

$$\boldsymbol{H^2 = \frac{\sigma_G^2}{\sigma^2}}$$

proportion of phenotypic variation attributable to all genetic factors

Directly analogous to the R^2 coefficient of determination, if there are no shared environmental factors affecting the phenotype of parents and offsprings.

The relationship holds if no hidden **Gene X Environment interaction** is in place. Otherwise:

$$Var(Y) = \sigma^2 = \sigma_G^2 + \sigma_E^2 + \underbrace{2cov(GE)} + \sigma_u^2$$

In the presence of G x E interaction, the genetic heritability estimator H^2 might be biased, because part of variability might be erroneously attributed to genetic or environmental components

The **GENETIC component** can be further decomposed

$$Var(Y) = \sigma^2 = \sigma_G^2 + \sigma_E^2 + \sigma_u^2$$

$$= (\sigma_A^2 + \sigma_{other\ genetic\ effects}^2) + \sigma_E^2 + \sigma_u^2$$

Additive genetic variance =
ADDITIVE POLYGENIC EFFECT

Narrow-sense heritability (genetic heritability) $h^2 = \frac{\sigma_A^2}{\sigma^2}$

proportion of variance explained by the additive polygenic component

Genetic heritability (h^2) should be estimated with methods that remove the shared-environment effect:

e.g. Relatedness Disequilibrium Regression by Young et al.

„**Relatedness disequilibrium regression estimates heritability without environmental bias**„ Nat Genet 2018

Additional readings:

The Infinitesimal

The missing heritability question is now (mostly) answered

Not with a bang but with a whimper

SASHA GUSEV

NOV 21, 2025

https://theinfinitesimal.substack.com/p/the-missing-heritability-question

Genetica
https://doi.org/10.1007/s10709-022-00149-7

ORIGINAL PAPER

Heritability: What's the point? What is it not for? A human genetics perspective

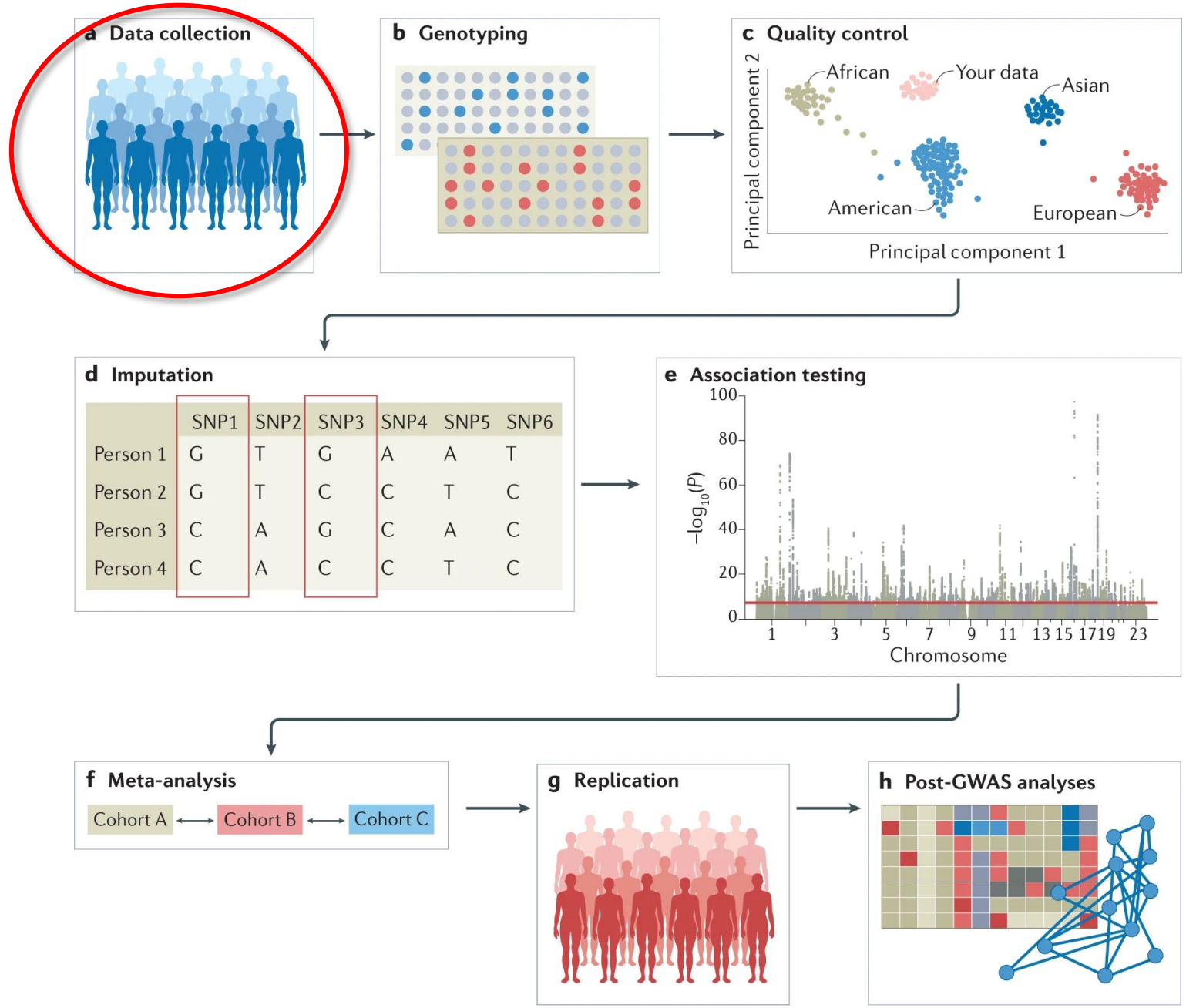
Nicolas Robette[1] · Emmanuelle Génin[2] · Françoise Clerget-Darpoux[3]

$$H^2 = \frac{\sigma_G^2}{\sigma^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

PDF reader: https://t.co/LZS2ywJXCG

Take home messages

1. GWAS are hypothesis-free screenings
2. Complex traits are multifactorial and critically depend on both G and E
3. Complex diseases can be seen as the extreme manifestations of underlying quantitative processes
4. GWAS of complex traits rely on the assumption of a polygenic or omnigenic model



Uffelmann et al, Nat Rev Methods Primers 2021; 1:59

Different kinds of study designs, depending on the purpose:

- Case-control studies (retrospective)

  Disease cases compared vs control individuals without the disease

- Cohort studies (prospective)

  A group of individuals is followed up over time to assess whether people exposed to a certain risk factor have an increased risk of disease

Clinically-based

Population-based

Family-based

CKD-related phenotypes – experiences from the CKDGen Consortium

The CKDGen Consortium: ten years of insights into the genetic basis of kidney function

Kidney International (2020) **97,** 236–242; https://doi.org/10.1016/j.kint.2019.10.027

Anna Köttgen
University of Freiburg

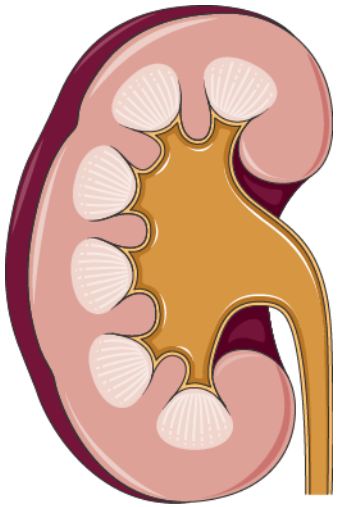
- Analyst group (~20 members) meeting weekly since 2009
- >120 GWAS studies
- >2 Million individuals
- Aim: uncovering genes/gene products that can become targets to prevent/treat CKD
- Main traits: eGFRcrea/cys, UACR, CKD, microalbuminuria, urate, gout, calcium, phosphorus, albumin
- Philosophy: advancing science while promoting junior scientist‘s careers

Table 1 | Kidney-related outcomes assessed by the CKDGen Consortium in (epi-)genome-wide discovery studies

Dimension of kidney function	Phenotype	Description	PubMed ID
Kidney filtration function and eGFR-based kidney disease definitions	eGFRcrea	Creatinine-based estimation of GFR[a]	19430482,[b] 20383146, 22479191, 22962313, 24029420, 26831199, 27920155, 28452372, 29097680,[b] 30315176,[b] 31152163
	eGFRcys	Cystatin C–based estimation of GFR	19430482,[b] 20383146, 22479191, 24029420, 26831199, 27920155, 28452372
	BUN	Blood urea nitrogen	31152163
	Annual decline of eGFR	Change in eGFR, in ml/min per 1.73 m^2 decline per year	25493955
	CKD	eGFRcrea <60 ml/min per 1.73 m^2	19430482,[b] 20383146, 22479191, 26831199, 30315176,[b] 31152163
	CKD45	eGFRcrea <45 ml/min per 1.73 m^2	22479191
	Rapid eGFR decline	Annual eGFR decline of ≥3 ml/min per 1.73 m^2	25493955
	CKDi	Incident CKD defined as new onset of eGFR <60 ml/min per 1.73 m^2 during follow-up	25493955
	CKDi25	Incident CKD with ≥25% eGFR decline from baseline	25493955
Kidney damage	UACR	Urinary albumin-to-creatinine ratio	21355061, 26631737, 27920155, 30315176,[b] 31511532
	MA	(Micro- and macro-) albuminuria[c]	21355061, 26631737, 31511532
Urate metabolism (tubular function)	Serum urate levels		30315176,[b] 31578528
	Gout	Self-report, intake of antigout medication, ICD codes for gout	31578528

Köttgen & Pattaro. The CKDGen Consortium: ten years of insights into the genetic basis of kidney function. *Kidney Int* 2020

Chronic Kidney Disease



Decreased GFR

Markers of kidney damage

GFR

Albuminuria

Persistent haematuria

Urine sediment abnormalities

Electrolyte and other abnormalities due to tubular disorders (proteinuria, acido-basic disorders, …)

Histology abnormalities

Structural abnormalities detected by imaging

History of kidney transplantation

Pattaro and Ponte, *Kidney Int 2026*

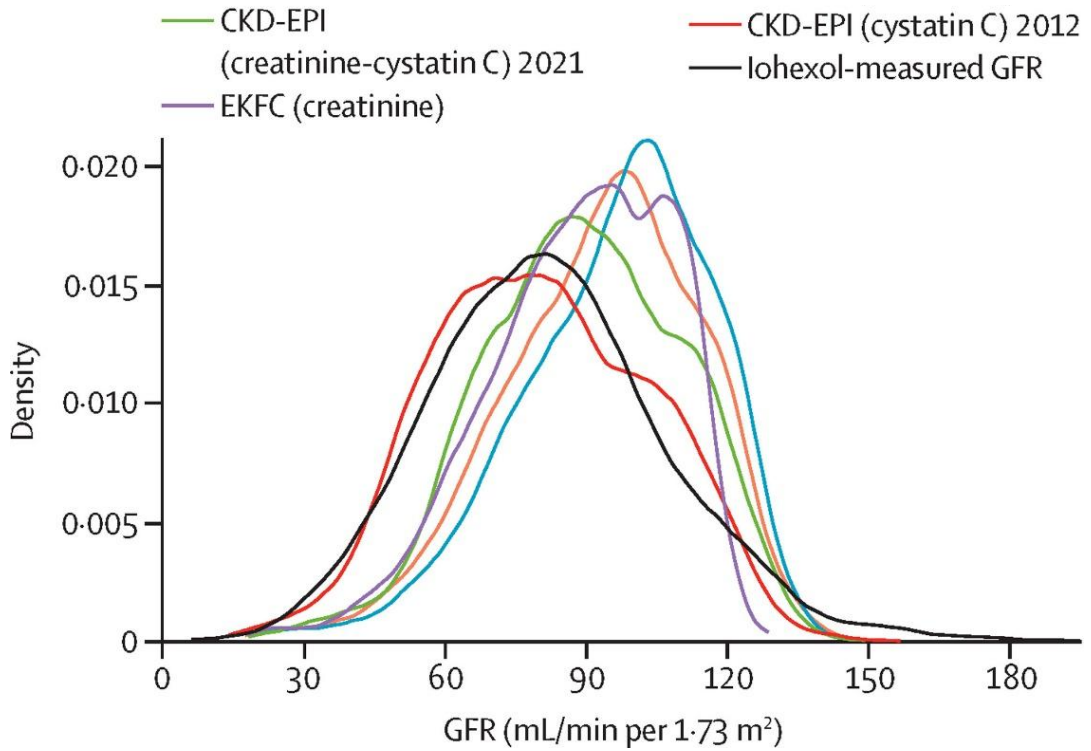
eGFR based on creatinine

$$\mathrm{eGFR} = 142 \times \min(\mathrm{Scr}/\kappa, 1)^{\alpha} \times \max(\mathrm{Scr}/\kappa, 1)^{-1.200} \times 0.9938^{\mathrm{Age}} \times 1.012 \ [\text{if female}]$$

Scr = serum creatinine in mg/dL **κ** = 0.7 for females, 0.9 for males **α** = -0.241 for females, -0.302 for males

eGFR based on cystatin C

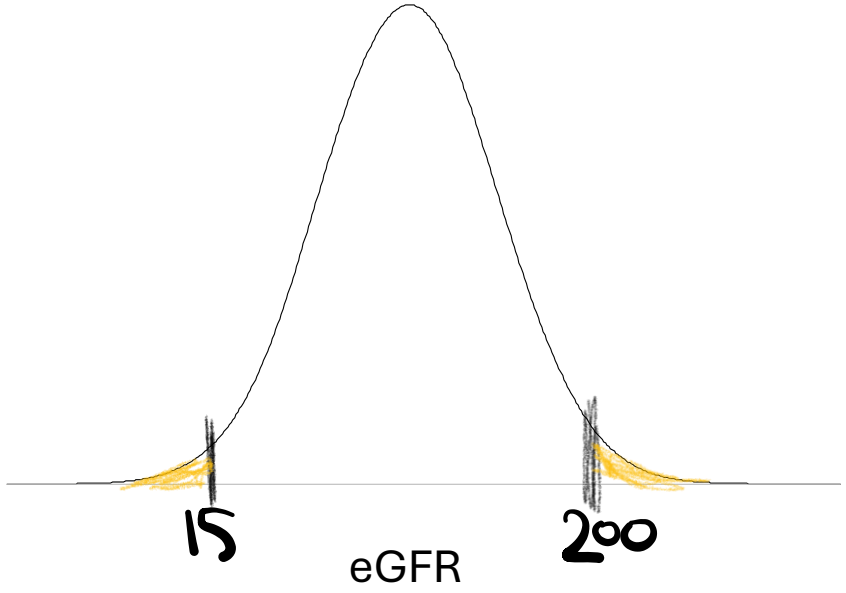
$$\mathrm{eGFR} = 133 \times \min(\mathrm{Scys}/0.8, 1)^{-0.499} \times \max(\mathrm{Scys}/0.8, 1)^{-1.328} \times 0.996^{\mathrm{Age}} \times 0.932 \ [\text{if f}$$



Fabian J et al, *Lancet Glob Health* 2022 - Measurement of kidney function in Malawi, South Africa, and Uganda: a multicentre cohort study

eGFR vs serum creatinine, transformations

- Using eGFR or creatinine?



- Until 2016
  - Winsorize eGFR at 15 and 200 ml/min
  - ln(eGFR) ~ age + sex + <design variables> → residuals → GWAS

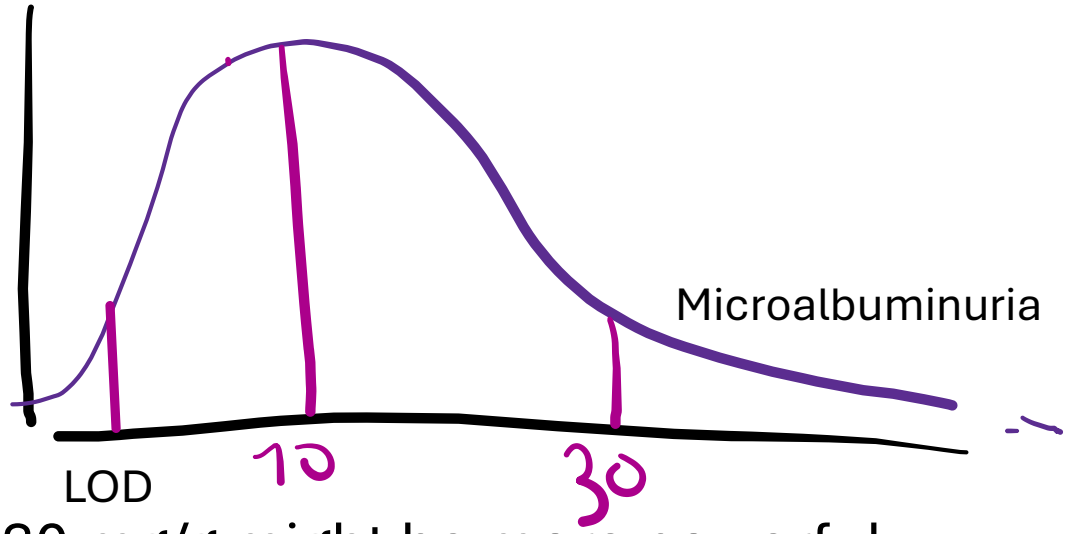
- After 2016 → inverse normal transformation
  1. INT(eGFR) ~ age + sex + <design variables> → GWAS

Pascal Schlosser

University of Freiburg

McCaw ZR, et al. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*. 2020;76:1262–1272.

UACR



- Do not remove values < limit of detection
- Set to the LOD or run multiple imputation
- UACR is not well calibrated across labs
- Sometimes, using microalbuminuria = UACR > 30 mg/g might be more powerful (less noisy)
  - Cases: UACR >30; Ctrls: UACR <= 30
  - Cases: UACR >30; Ctrls: UACR <= 10
- Since 2016:
  1. INT(UACR) ~ age + sex + <design variables> → GWAS

Matthias Wuttke

University of Freiburg

Alexander Teumer

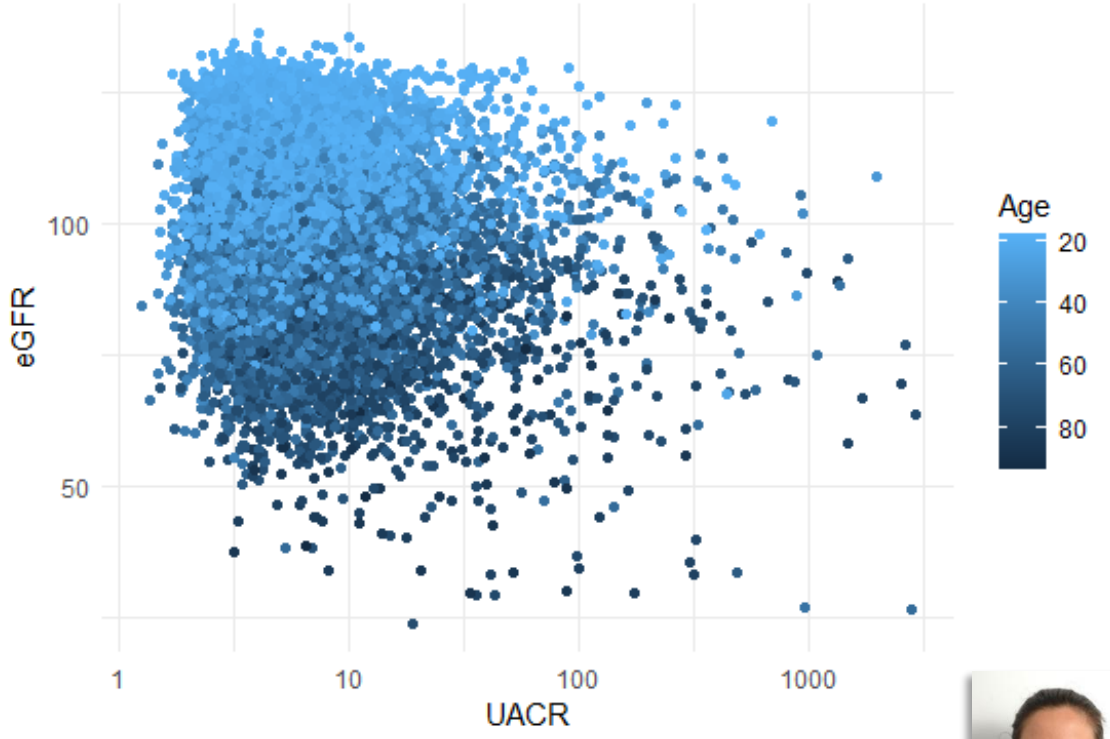
University of Greifswald

Minimal adjustment

- Traits are usually adjusted for
  - Age, sex
  - Design variables: study center; measurement method; batch; etc.
- Never adjust for variables that might belong to any causal or biological pathway
  - It might prevent identification of loci
    - eg: Gene → T2D → CKD
  - It might cause collider bias (false positives)
    - eg: Gene → CKD ← Infections

CKD = eGFR < 60 ml / min / 1.73 m^2 used just as a confirmatory trait / for clinical interpretation

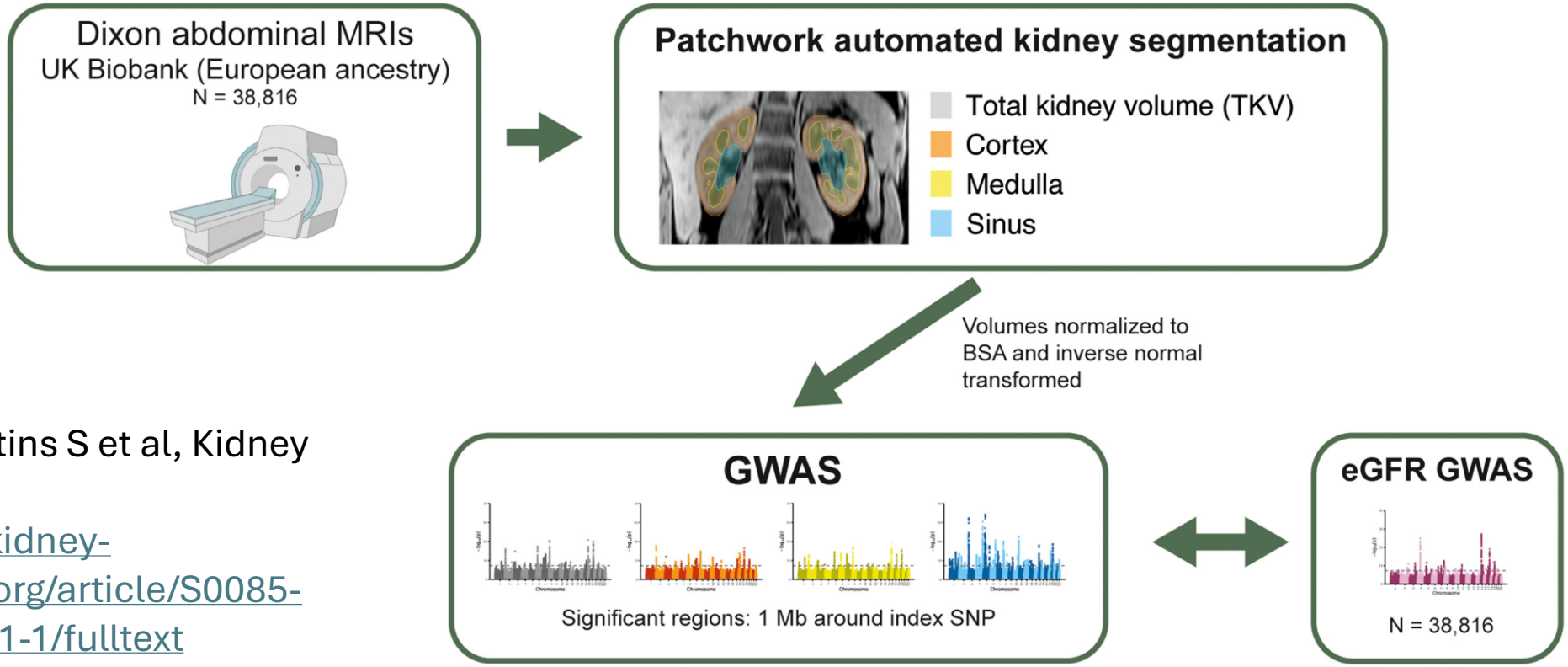
- Why not using as primary trait?
    1. Power loss vs eGFR
    2. Differential diagnosis by ancestry
- Using EHR-based CKD?
    1. Possible but, not available in most studies
    2. Heavy diagnostic bias
- KDIGO-based CKD, ie: eGFR < 60 ml/min OR UACR > 30 mg/g
    1. eGFR and UACR poorly correlated in general population samples
    2. Mixing up different processes (glomerular vs tubular)



Barbieri et al, *J Nephrol* 2025
https://link.springer.com/article/10.1007/s40620-024-02157-6

Other phenotypes

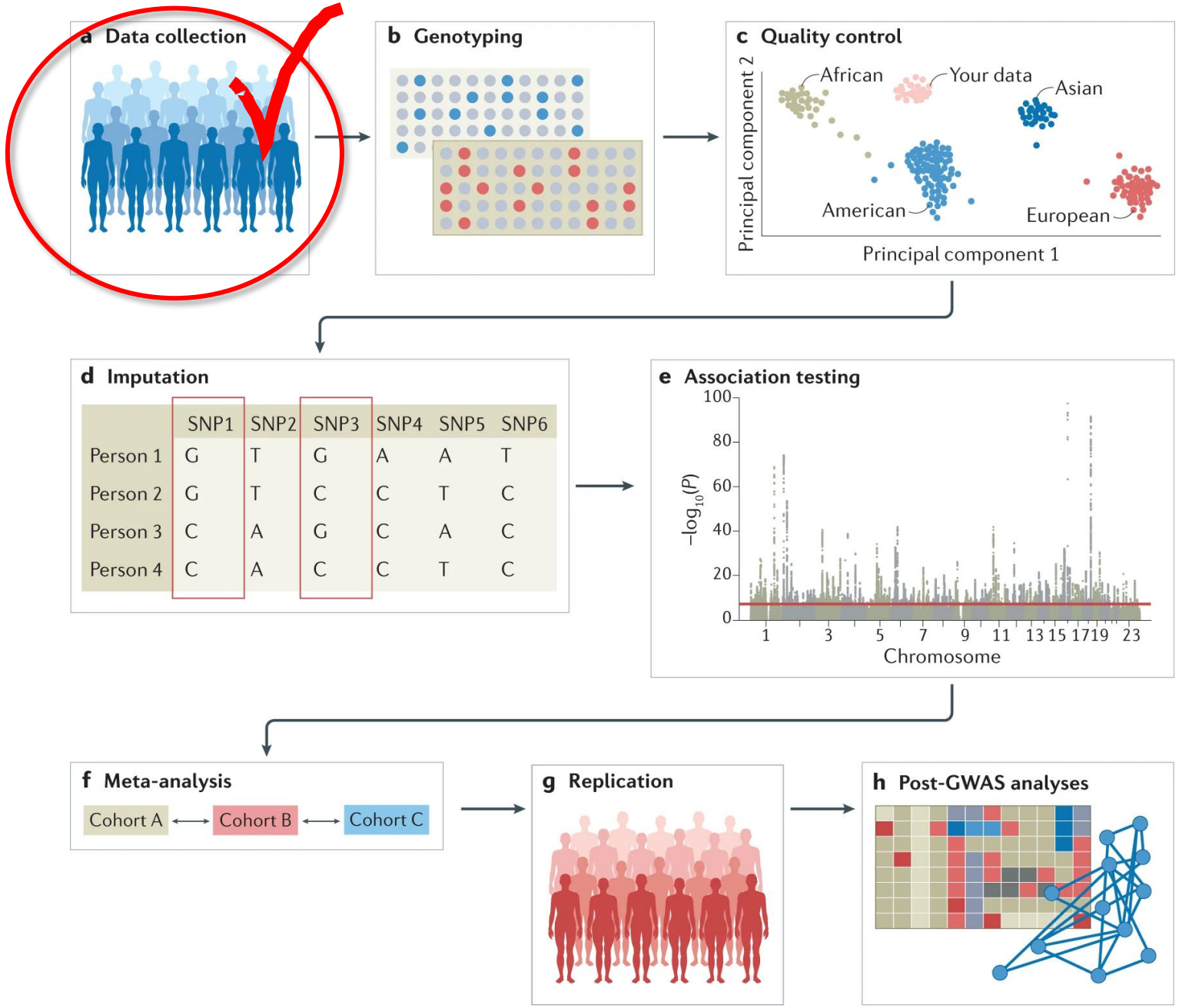
- Combined crea and cystatin-C CKD-EPI equation (not CKDGen; UKBB)
- Longitudinal : rapid eGFR decline and incident CKD
- Imaging (not CKDGen)



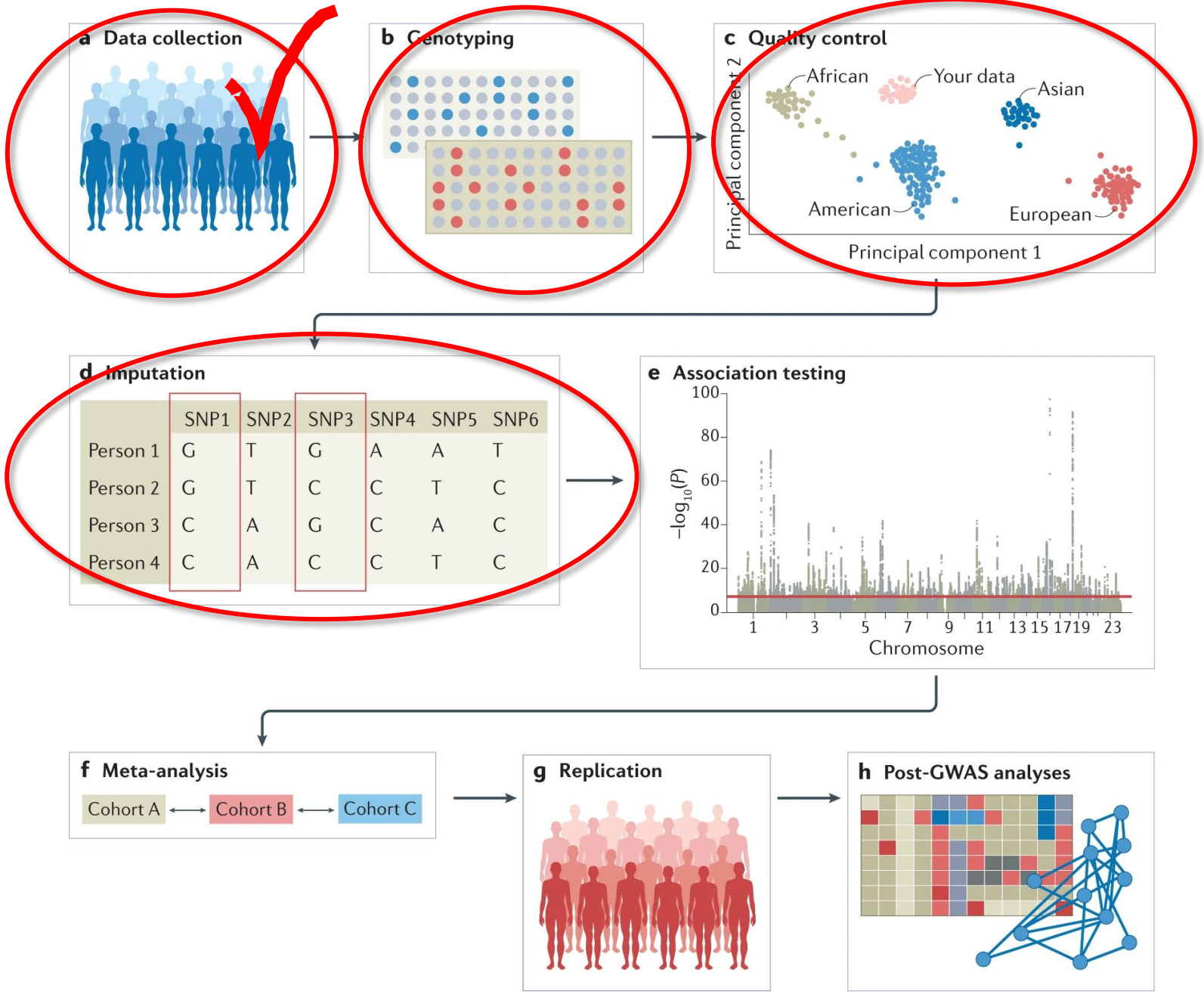
Monteiro-Martins S et al, Kidney Int 2026
https://www.kidney-international.org/article/S0085-2538(25)00781-1/fulltext

Take home messages

1. CKD dissected into its defining traits
2. Use quantitative traits if possible (more power)
3. Use INT to maximize power
4. Minimal adjustment



Uffelmann et al, Nat Rev Methods Primers 2021; 1:59



Uffelmann et al, Nat Rev Methods Primers 2021; 1:59