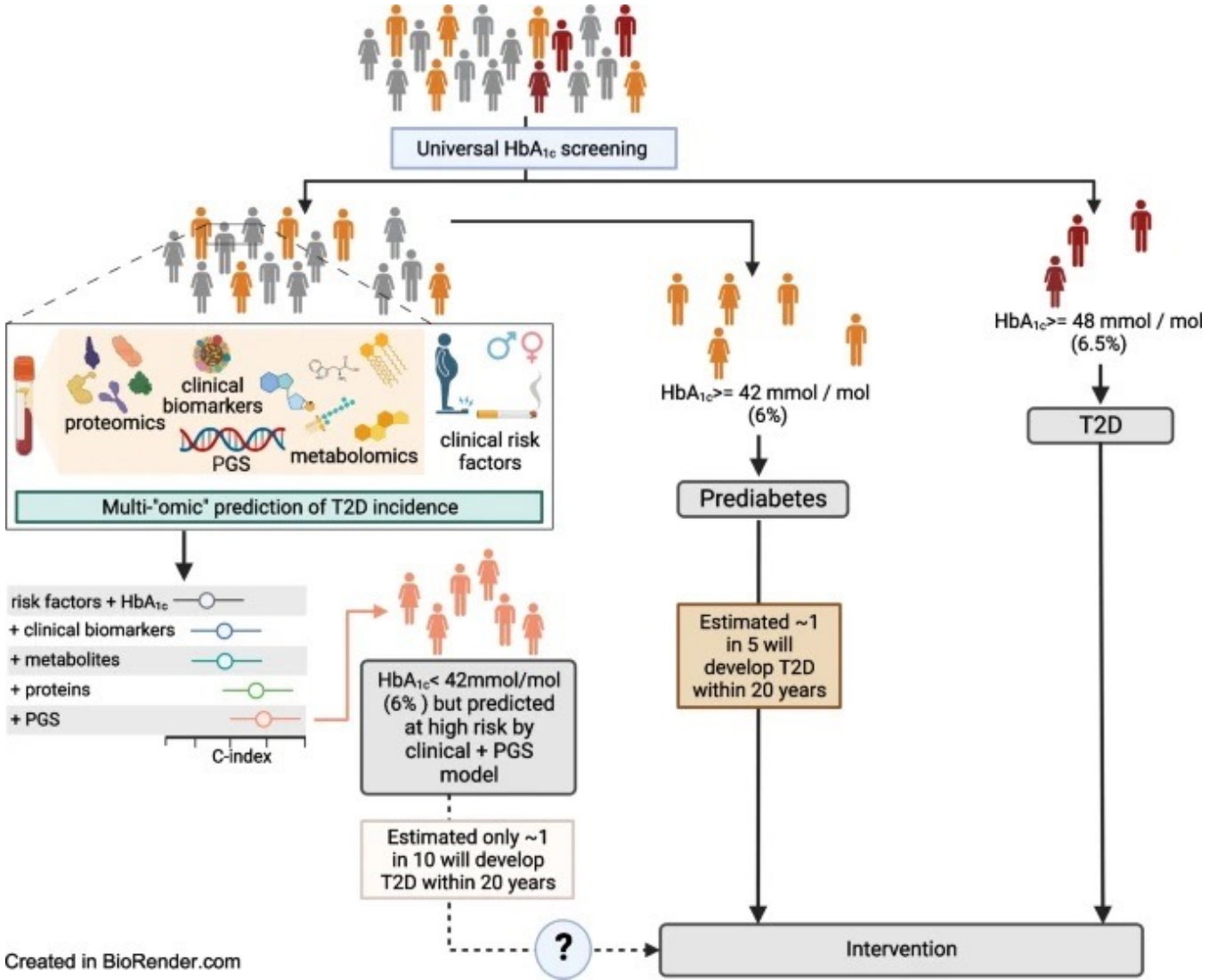
INTEGRATIVE GENOMICS

#10



LETS GET STARTED

AGENDA

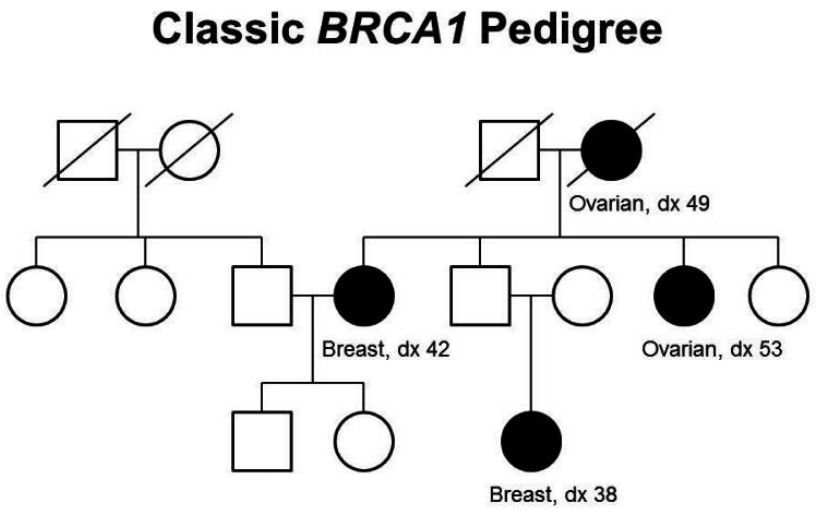
12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle

AGENDA

12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle

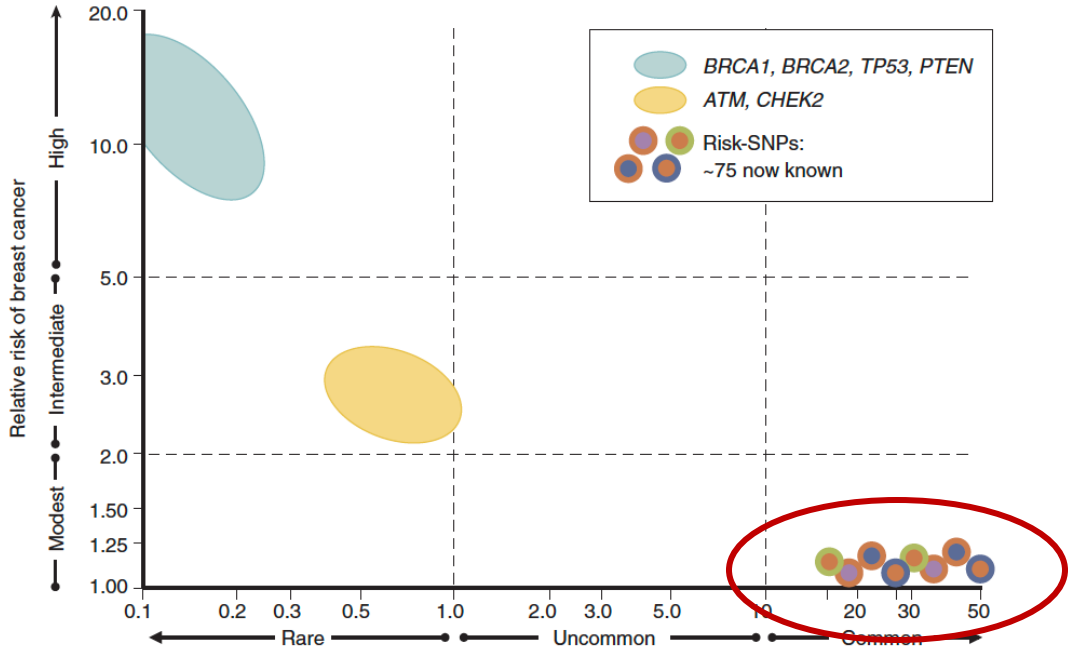
RARE CANCER MUTATIONS

- Can be germline or somatic
- Inherited driver mutations include BRCA1/2 and APC
- Nearly 10% of cancers are inherited
- Autosomal dominant pattern with incomplete penetrance
- Presents early and bilaterally
- BRCA1/2 follow the “two-hit” hypothesis
- Inherited cancers are recessive at the cellular level but dominant at the individual level

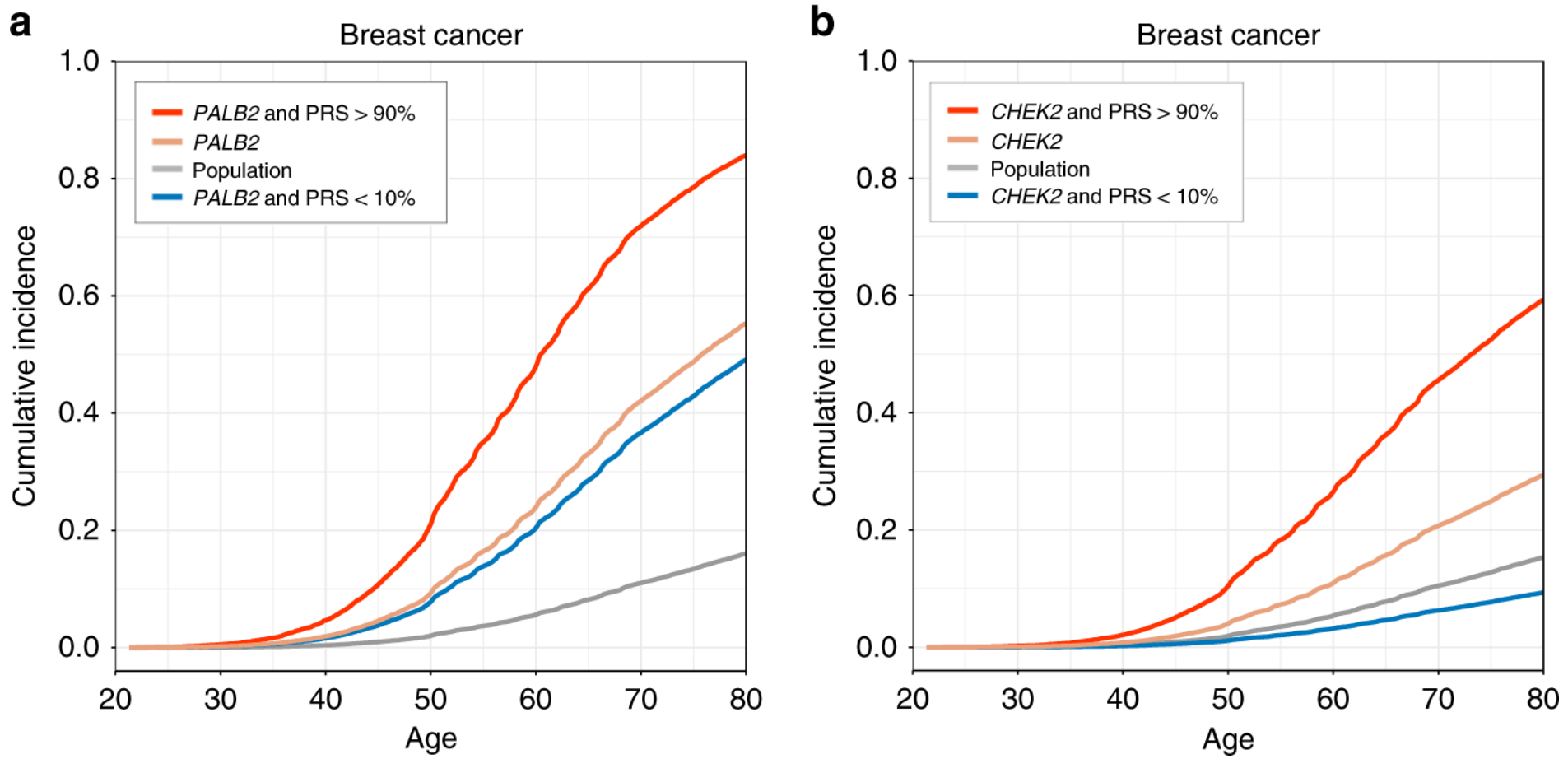


COMMON CANCER MUTATIONS

- Common variants are not strong enough to be considered driver mutations
- Identified by GWAS
- Identification of common SNPs relies on
  - sample size
  - allele frequency
  - effect size
  - phenotypic clarity
- Heritability estimates range from 4-26%
- Strength of PGS will vary between cancers



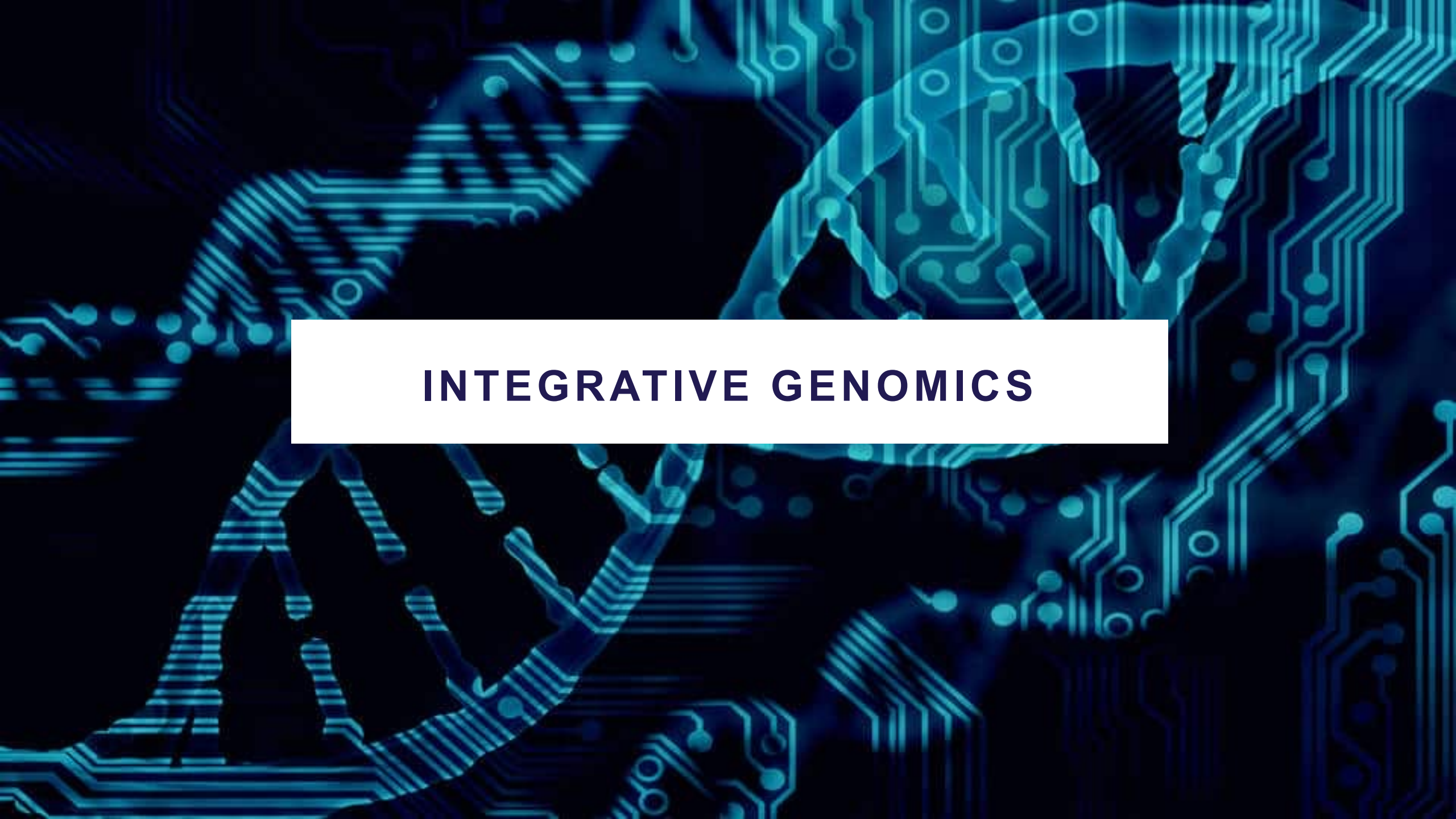
COMBINING CANCER MUTATIONS




BREAK + QUESTIONS

AGENDA

12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle



INTEGRATIVE GENOMICS

integrative

adjective

UK /ˈɪn.tə.grə.tɪv/ **US** /ˈɪn.ţə.greɪ.ţɪv/

combining two or more things in order to make them more effective:

- *The new system will allow more efficient and integrative management of our data.*
- *Our patients might benefit if we took a more integrative approach to their care.*

integrative

adjective

What do we combine?

UK /ˈɪn.tə.grə.tɪv/ US /ˈɪn.ţə.greɪ.ţɪv/

combining two or more things in order to make them more effective:

- *The new system will allow more efficient and integrative management of our data.*
- *Our patients might benefit if we took a more integrative approach to their care.*

integrative

adjective

UK /ˈɪn.tə.grə.tɪv/ **US** /ˈɪn.ţə.greɪ.ţɪv/

What do we combine?

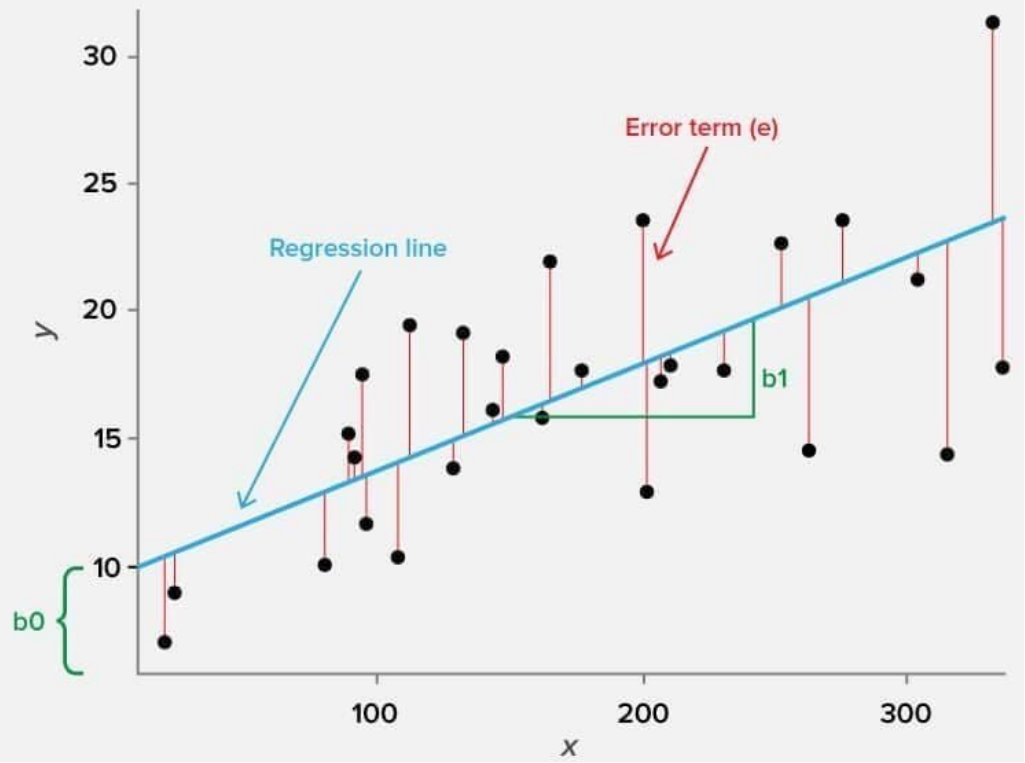
combining two or more things in order to make them more effective:

How do we combine it?

- *The new system will allow more efficient and integrative management of our data.*
- *Our patients might benefit if we took a more integrative approach to their care.*

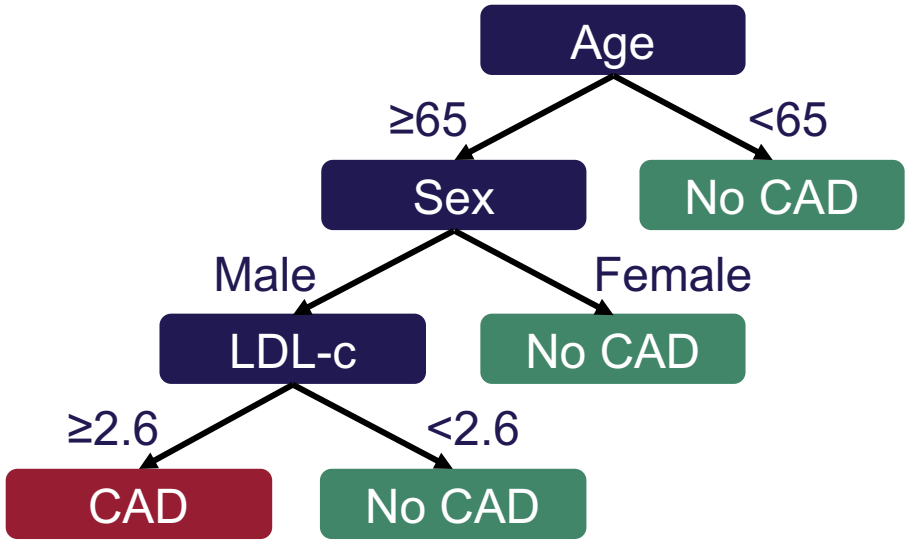
STARTING WITH THE “HOW”

- Regression models
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$
  - Can be penalized to reduce overfitting
  - Always interpretable



STARTING WITH THE “HOW”

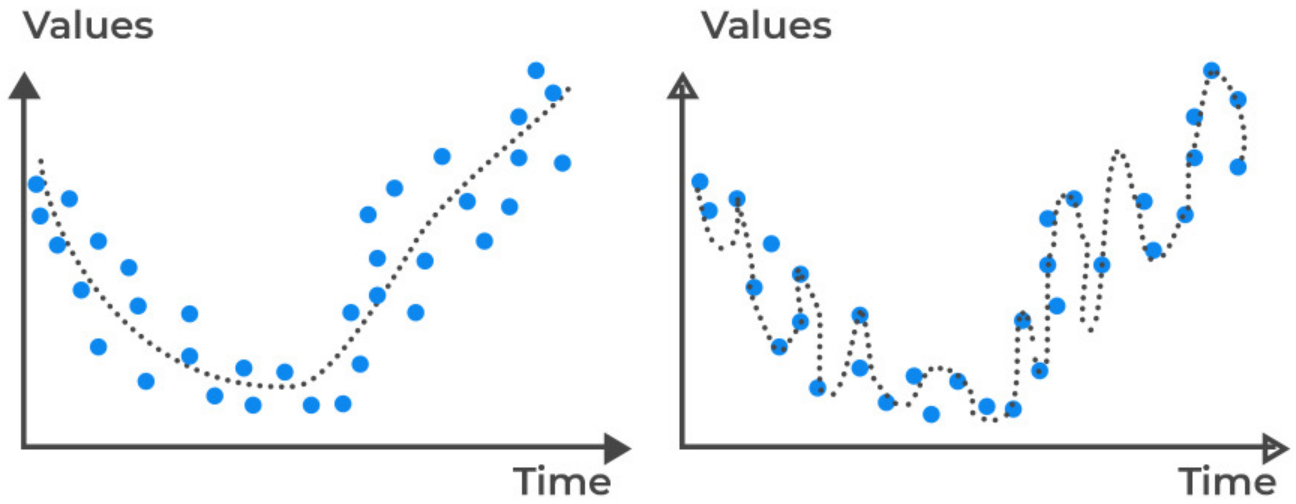
- Regression models
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$
  - Can be penalized to reduce overfitting
  - Always interpretable
- Decision trees
  - Can be combined in “forests” to reduce overfitting
  - Each tree contains random feature subsets
  - Large feature sets increase complexity
  - Becomes ”blackbox” at scale



FITTING A LOGISTIC REGRESSION

- Minimize the cost function
  - $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Risk of overfitting

- Extend cost function with penalty term
  - $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|\widehat{\beta}_j|$
  - λ is a tuning parameter
  - p is the number of predictors
  - Penalty term increases with many predictors and large coefficients

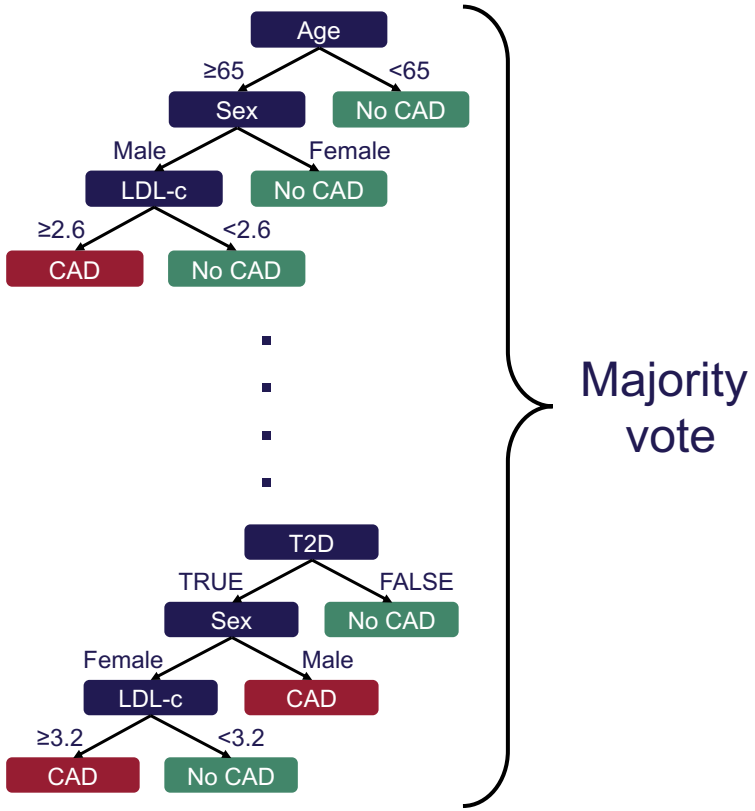


FITTING A RANDOM FOREST

- Sample random subset of data
- Train decision tree on data subset

} Repeat N times

- Summarise decision trees by majority vote



THINGS TO CONSIDER

- Classification or regression
- Interpretability
- Model size
- Non-linear interactions
- Over-/Underfitting
- Computational costs

WHAT DO WE COMBINE?

WHAT IS AVAILABLE?

- Clinical data – age, sex, behaviour, comorbidities
- Genomics

- Epigenomics
- Microbiomics
- Lipidomics
- Proteomics
- Glycomics
- Transcriptomics
- Metabolomics

genomics

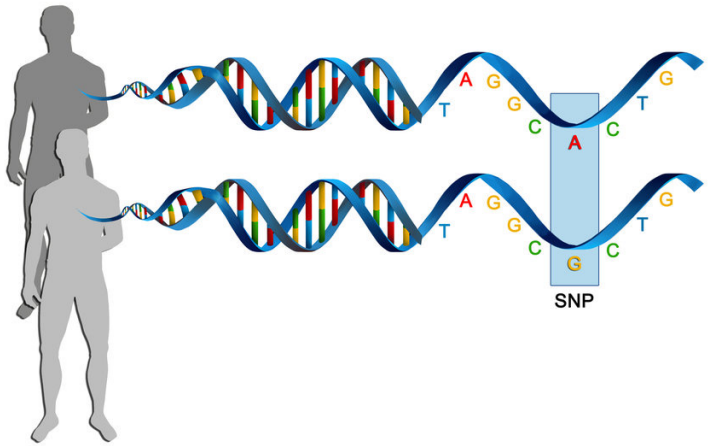
noun [U]

UK /dʒəˈnəʊm.ɪks/ **US** /dʒəˈnoʊm.ɪks/

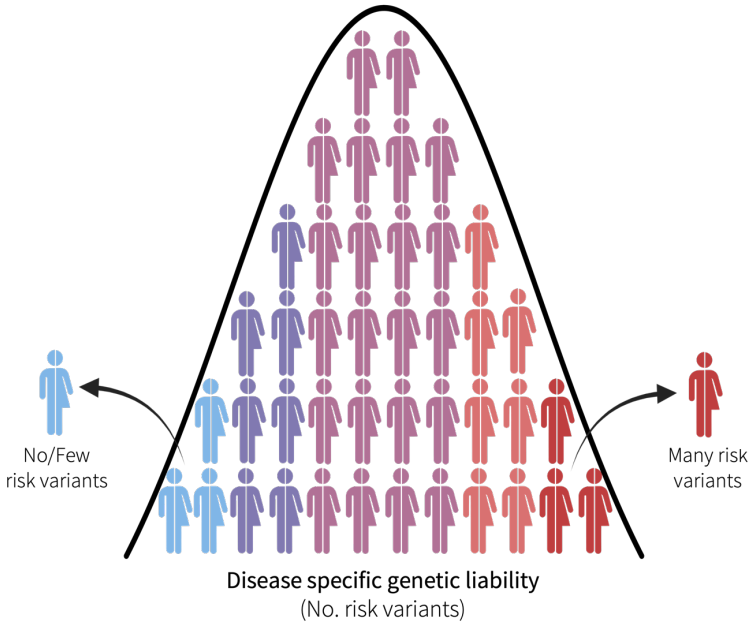
the study of the genomes of living things:

- *She is a specialist in animal genomics.*

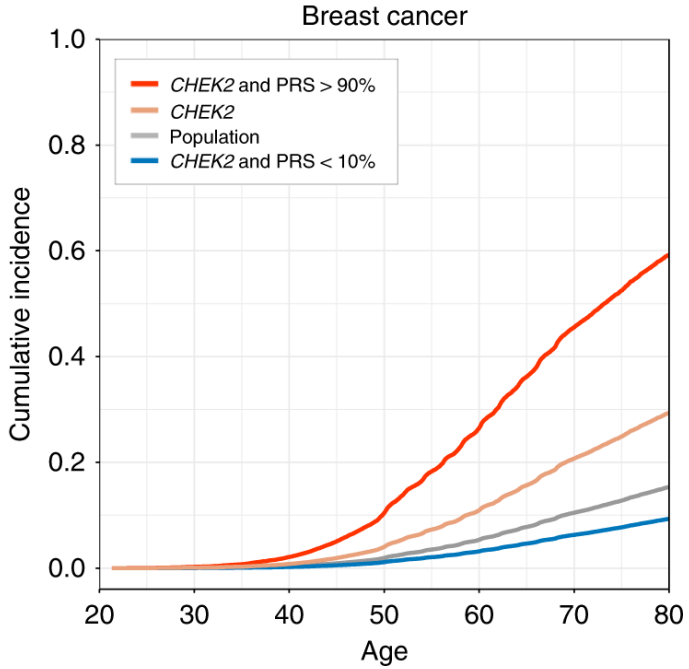
THE GENOME SO FAR



Monogenic effects

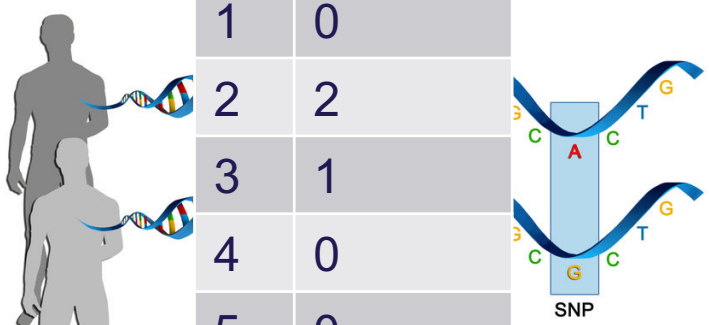


Polygenic effects



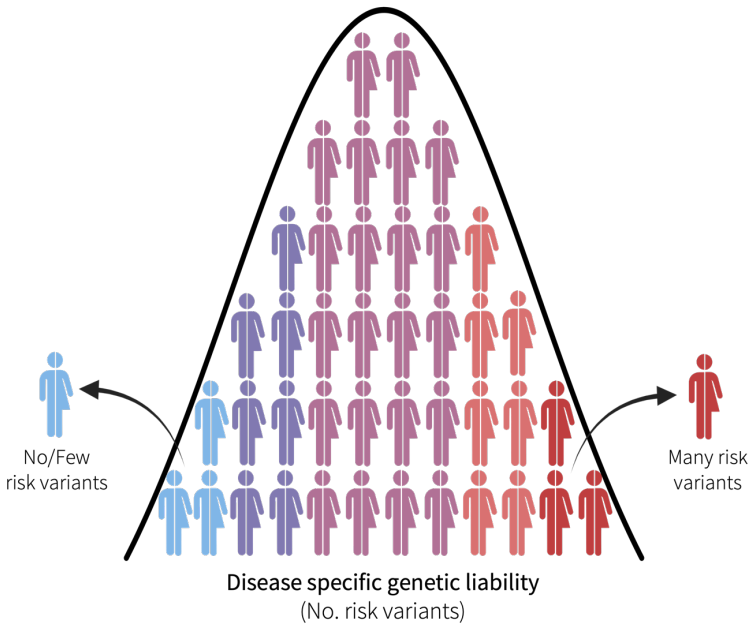
Combined effects

THE GENOME SO FAR

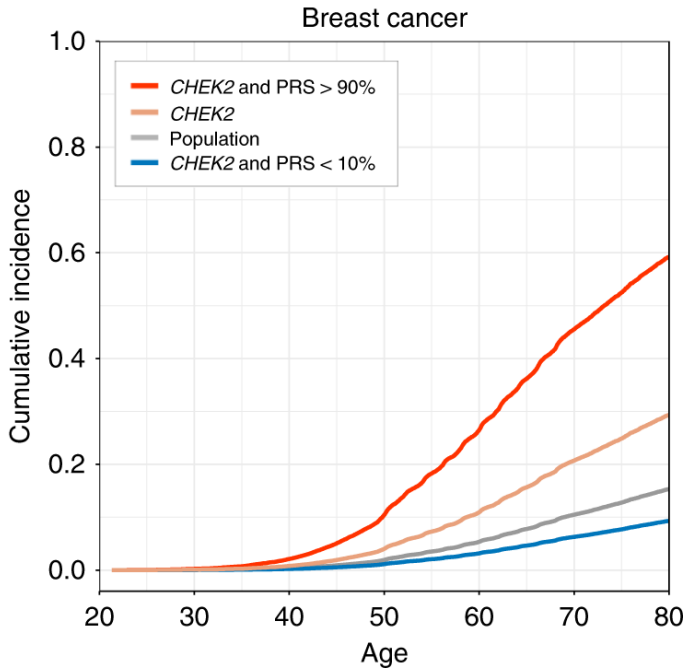


ID	Gene A
1	0
2	2
3	1
4	0
5	0
6	1
7	2

Monogenic effects



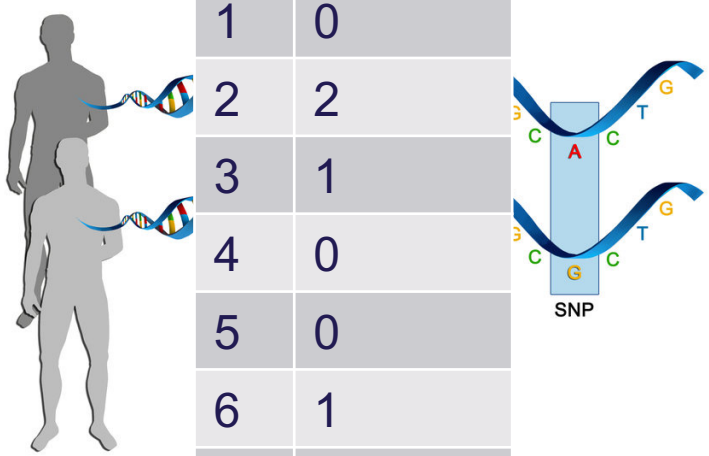
Polygenic effects



Combined effects

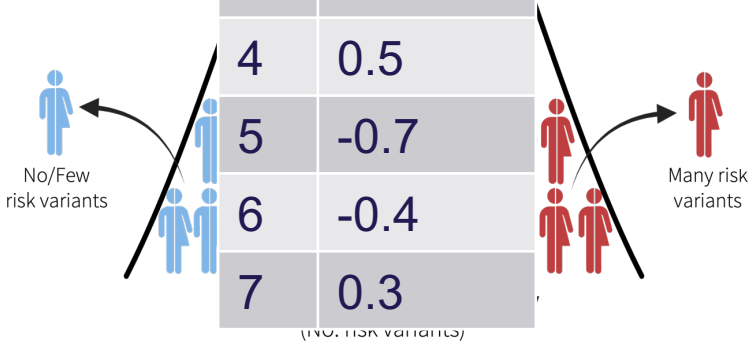
THE GENOME SO FAR

ID	Gene A
1	0
2	2
3	1
4	0
5	0
6	1
7	2

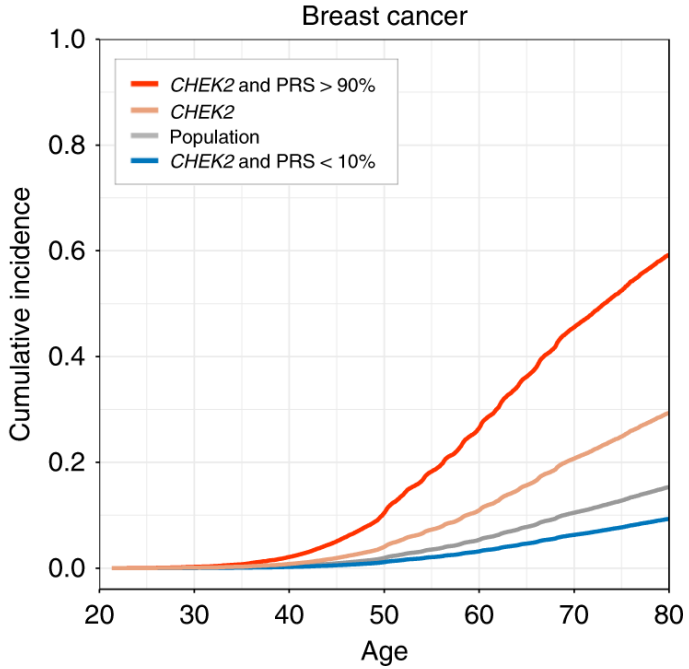


Monogenic effects

ID	PGS
1	0.1
2	-0.2
3	1.3
4	0.5
5	-0.7
6	-0.4
7	0.3



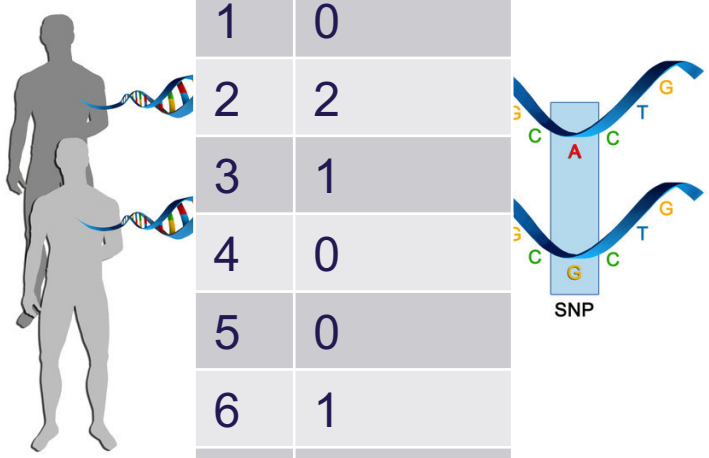
Polygenic effects



Combined effects

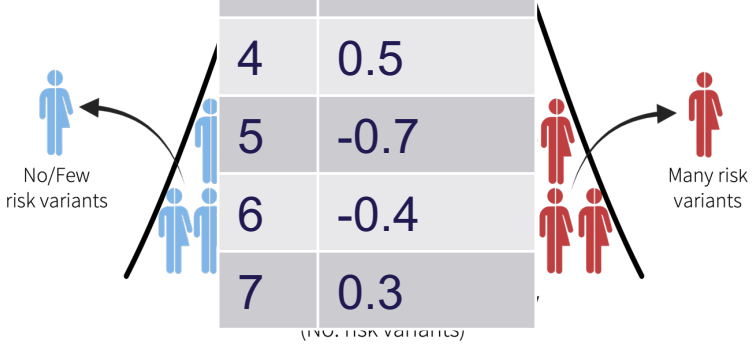
THE GENOME SO FAR

ID	Gene A
1	0
2	2
3	1
4	0
5	0
6	1
7	2



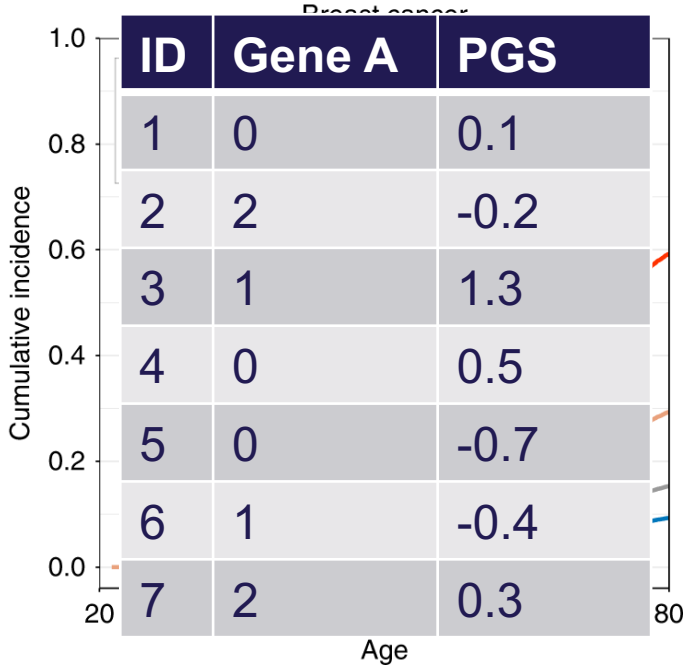
Monogenic effects

ID	PGS
1	0.1
2	-0.2
3	1.3
4	0.5
5	-0.7
6	-0.4
7	0.3



Polygenic effects

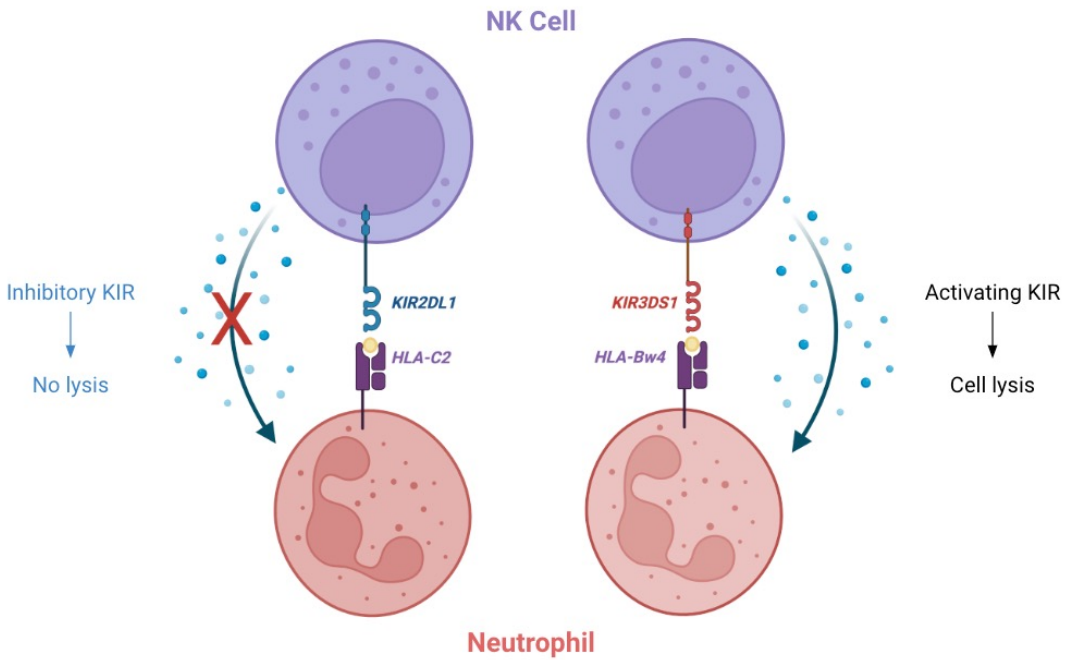
ID	Gene A	PGS
1	0	0.1
2	2	-0.2
3	1	1.3
4	0	0.5
5	0	-0.7
6	1	-0.4
7	2	0.3



Combined effects

OTHER ASPECTS OF THE GENOME

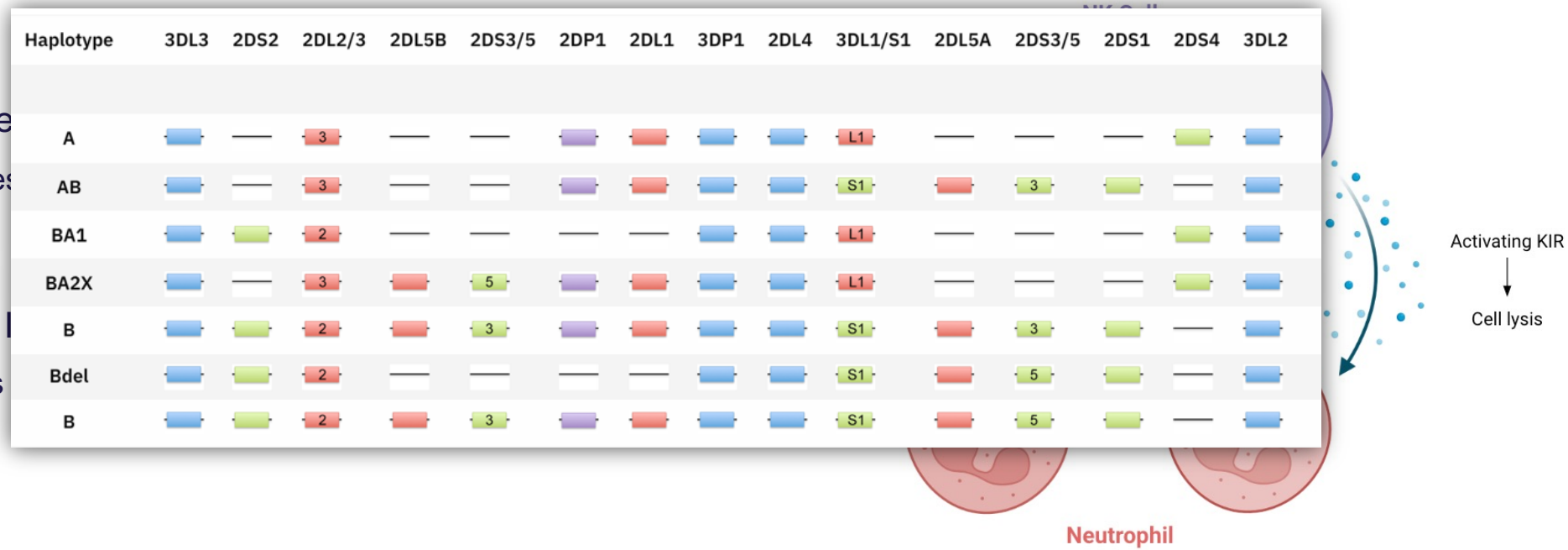
- The Human Leukocyte Antigens (HLA)
  - 41560 alleles across 20 genes
- The Killer-cell Immunoglobin-like Receptors (KIR)
  - 2219 alleles across 17 genes



OTHER ASPECTS OF THE GENOME

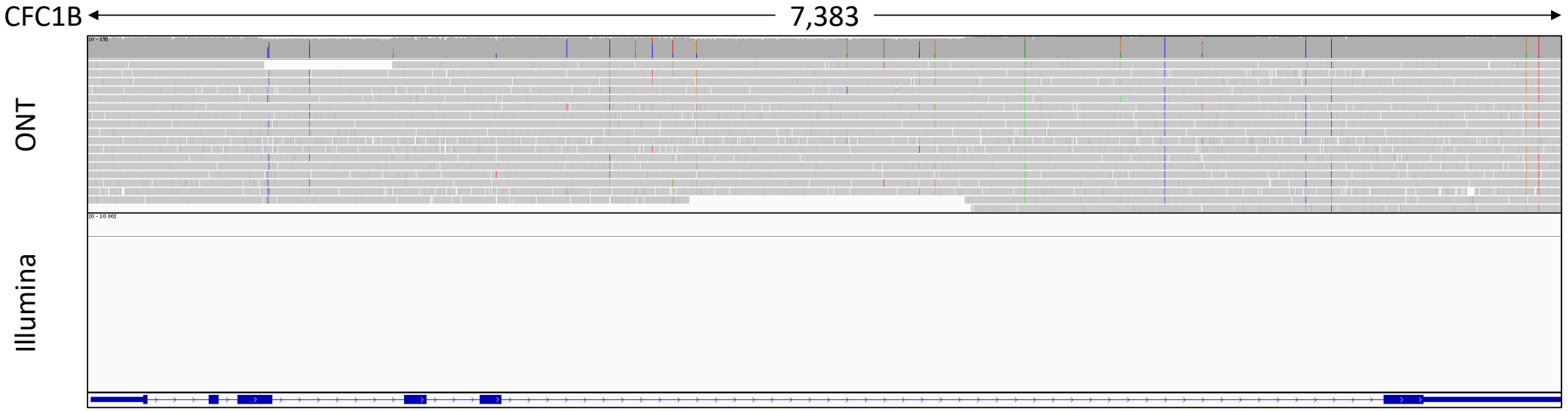
- The Human Le
  - 41560 alleles
- The Killer-cell I
  - 2219 alleles

Haplotype	3DL3	2DS2	2DL2/3	2DL5B	2DS3/5	2DP1	2DL1	3DP1	2DL4	3DL1/S1	2DL5A	2DS3/5	2DS1	2DS4	3DL2
A			3							L1					
AB			3							S1		3			
BA1			2							L1					
BA2X			3		5					L1					
B			2		3					S1		3			
Bdel			2							S1		5			
B			2		3					S1		5			



OTHER ASPECTS OF THE GENOME

- Dark genomic regions
  - 36794 regions across 6054 genes



THINGS TO CONSIDER

- Genomics is a very broad term
- What is the technology underlying your data?
- What are the limitations of that technology?
- Is the immune system implicated?
- How much complexity have you removed to facilitate analysis?

ABSENCE CAD PREDICTION

- GLMNET algorithm, combining ridge and lasso regularization
- Ensuring balanced dataset by focusing on CAD absence group
- Leveraging proteomics, genomics and clinical data
- Proteomics are age- and sex-corrected

- Finds proteomics to be ineffective
- PRS_{CAD} significantly improves prediction of CAD absence

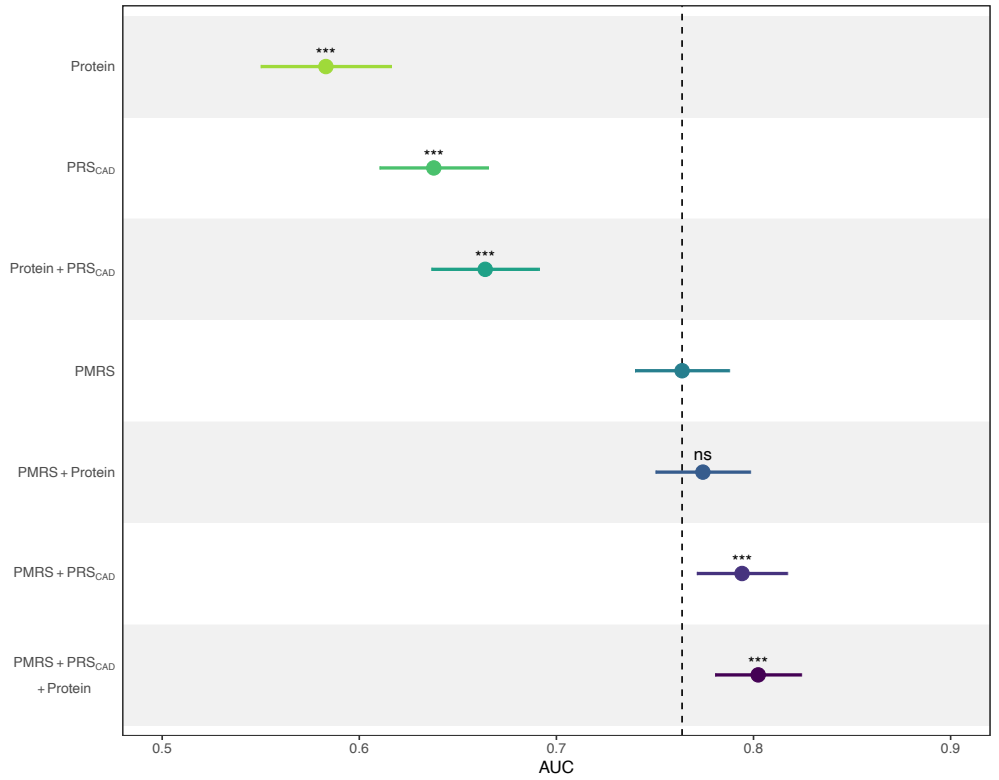
RESEARCH ARTICLE | Originally Published 27 September 2023 |

Check for updates

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc, Palle Duun Rohde, MSc, PhD, Jonathan Nørtoft Dahl, MD, Laust Dupont Rasmussen, MD, PhD, Samuel Emil Schmidt, MSc, PhD, Louise Nissen, MD, PhD, Victoria McGilligan, PhD, ... SHOW ALL ... , and Mette Nyegaard, MSc, PhD AUTHOR INFO & AFFILIATIONS

Circulation: Genomic and Precision Medicine • Volume 16, Number 5 • https://doi.org/10.1161/CIRCGEN.123.004053



ABSENCE CAD PREDICTION

RESEARCH ARTICLE | Originally Published 27 September 2023 |

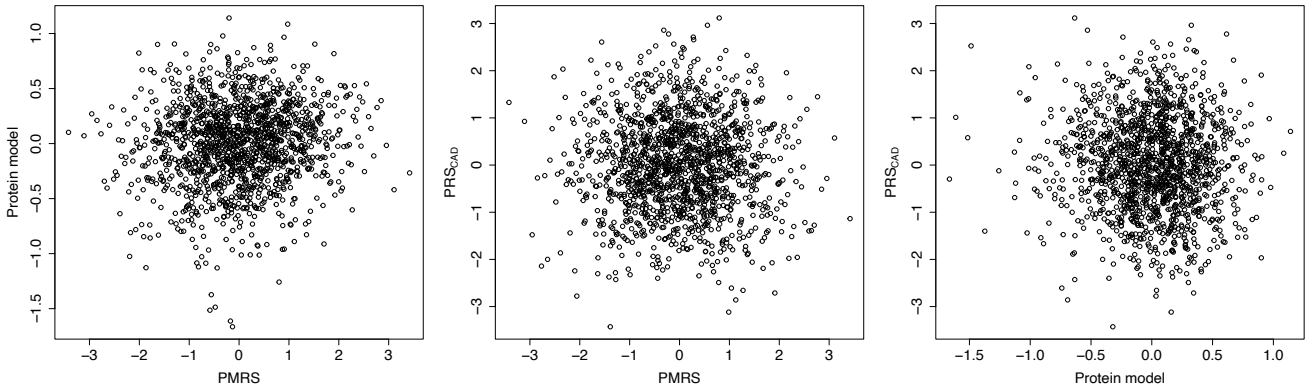
Check for updates

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc, Palle Duun Rohde, MSc, PhD, Jonathan Nørtoft Dahl, MD, Laust Dupont Rasmussen, MD, PhD, Samuel Emil Schmidt, MSc, PhD, Louise Nissen, MD, PhD, Victoria McGilligan, PhD, ... SHOW ALL ... , and Mette Nyegaard, MSc, PhD | AUTHOR INFO & AFFILIATIONS

Circulation: Genomic and Precision Medicine • Volume 16, Number 5 • https://doi.org/10.1161/CIRCGEN.123.004053

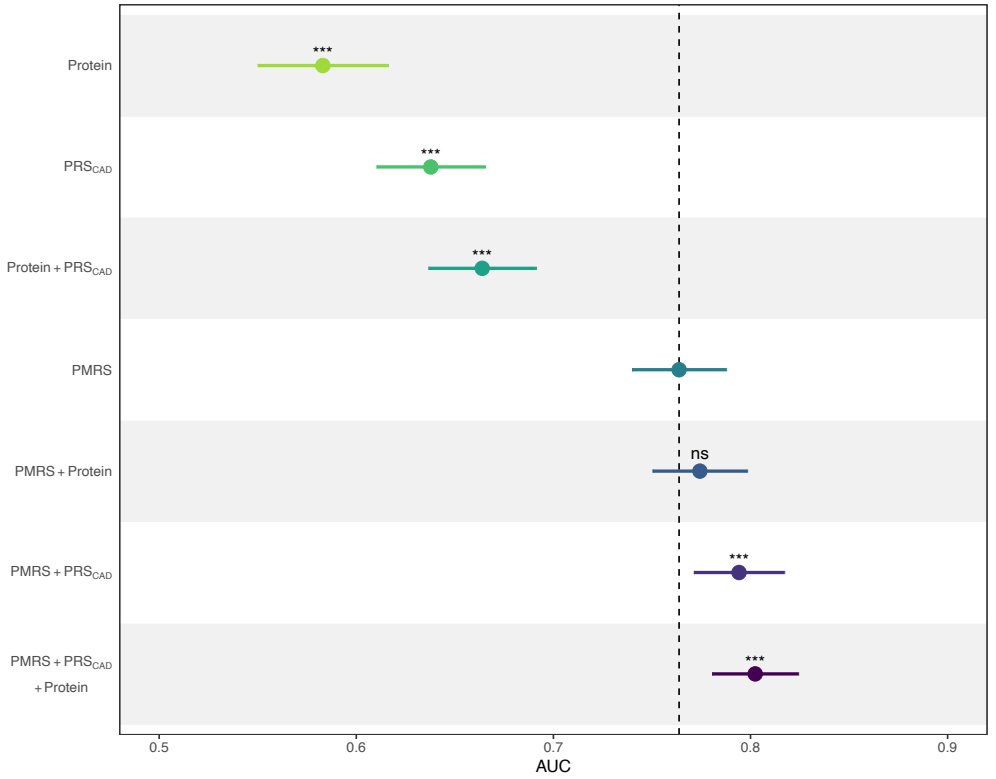
- GLMNET algorithm, combining ridge and lasso regularization



Pearson's correlations = 0.10
P < 0.001

Pearson's correlations = -0.01
P = 0.57

Pearson's correlations = -0.04
P = 0.13



HIGH-RISK PLAQUE PREDICTION

Research | Open access | Published: 20 March 2024

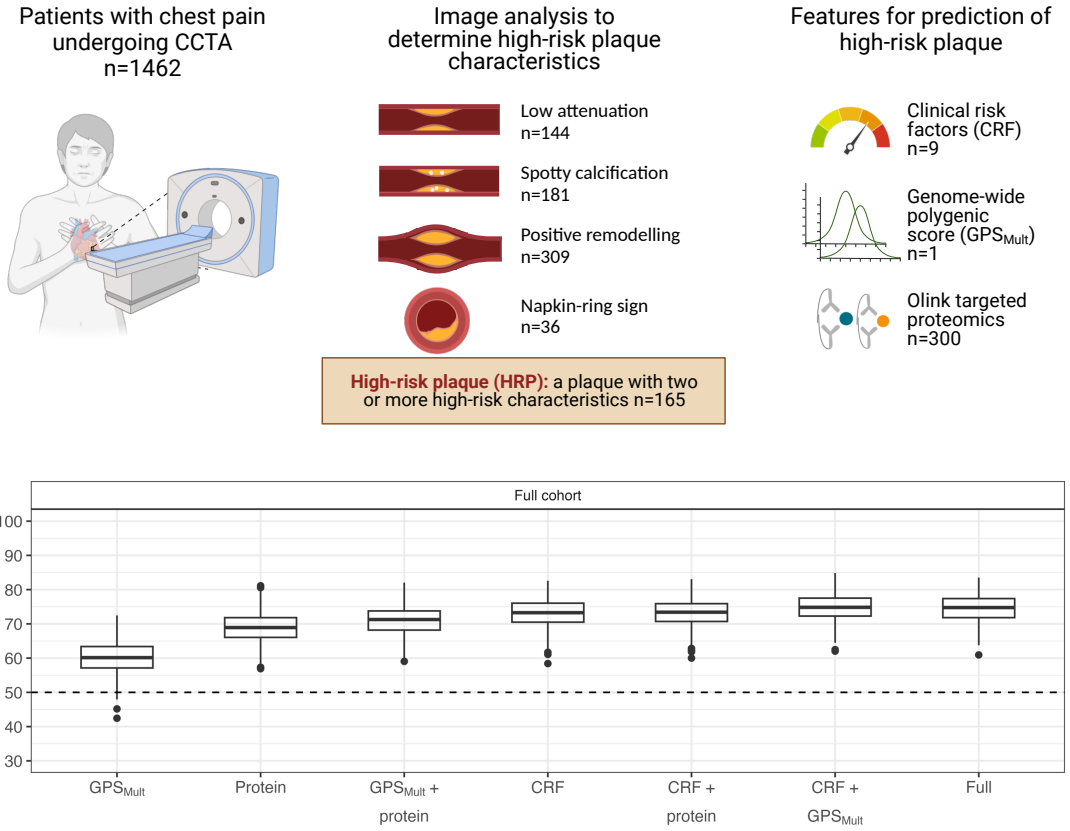
Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Peter Loof Møller, Palle Duun Rohde, Jonathan Nørtoft Dahl, Laust Dupont Rasmussen, Louise Nissen, Samuel Emil Schmidt, Victoria McGilligan, Daniel F. Gudbjartsson, Kari Stefansson, Hilma Holm, Jacob Fog Bentzon, Morten Bøttcher, Simon Winther & Mette Nyegaard

Genome Medicine **16**, Article number: 40 (2024) | Cite this article

- GLMNET algorithm
- Unbalanced dataset
- Leveraging proteomics, genomics and clinical data
- Proteomics are not age- and sex-corrected

- Estimates protein to be almost on par with clinical risk factor, but the combination does not improve prediction significantly



HIGH-RISK PLAQUE PREDICTION

Research | Open access | Published: 20 March 2024

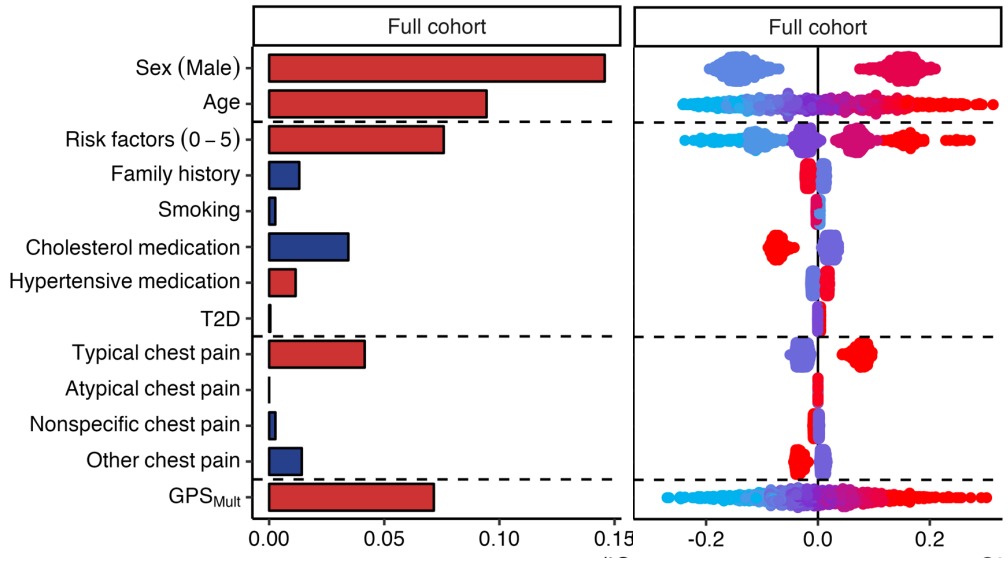
Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Peter Loof Møller, Palle Duun Rohde, Jonathan Nørtoft Dahl, Laust Dupont Rasmussen, Louise Nissen, Samuel Emil Schmidt, Victoria McGilligan, Daniel F. Gudbjartsson, Kari Stefansson, Hilma Holm, Jacob Fog Bentzon, Morten Bøttcher, Simon Winther & Mette Nyegaard

Genome Medicine **16**, Article number: 40 (2024) | Cite this article

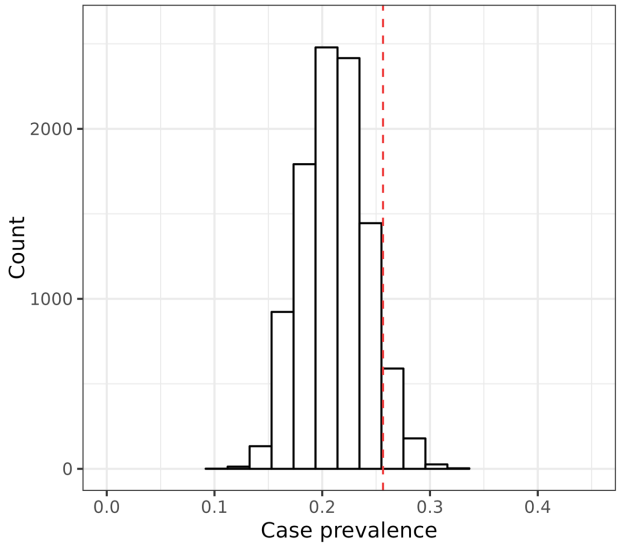
- GLMNET algorithm
- Unbalanced dataset
- Leveraging proteomics, genomics and clinical data
- Proteomics are not age- and sex-corrected
- Estimates protein to be almost on par with clinical risk factor, but the combination does not improve prediction significantly

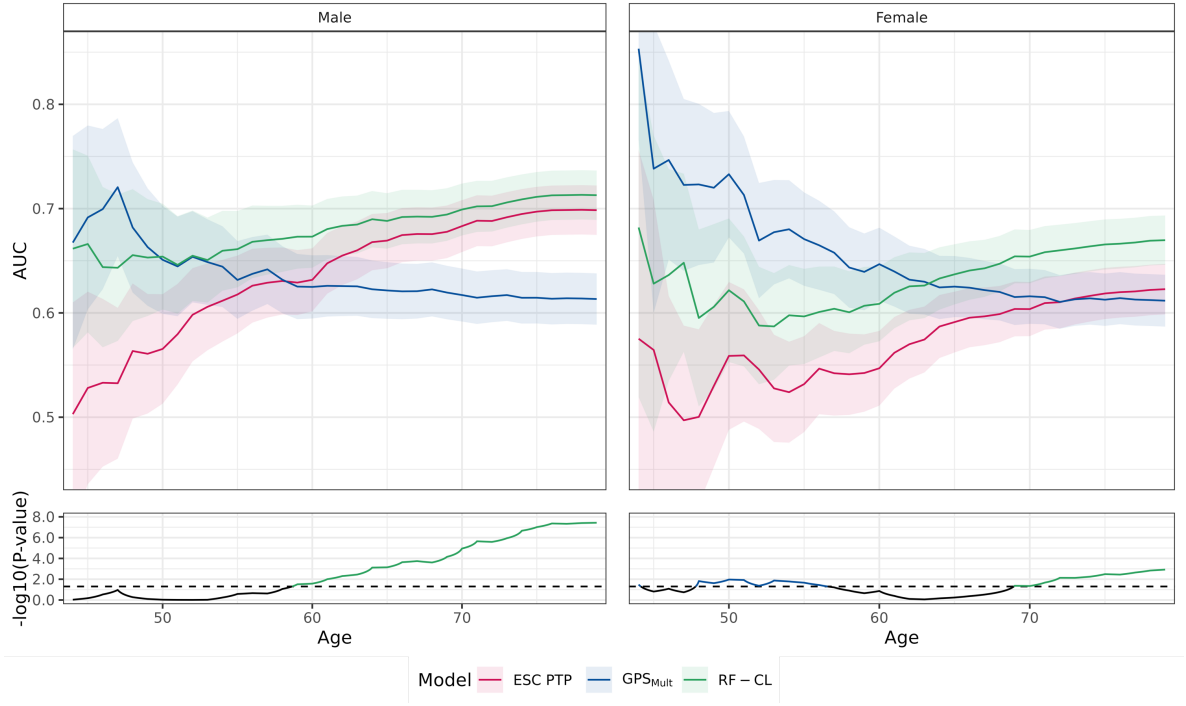
Feature importance estimates



CONDITIONAL PGS APPLICATION

- PTP (clinical model) referral of women is no better than chance
- Age-wise analysis of models reveal PGS to be the best predictor in young people, especially women





CONDITIONAL PGS APPLICATION

CAD risk	Sex	Age	PTP*	RF-CL*	PGS*
High risk (>15%)	Female	<55	0 (0%)	0 (0%)	43 (12%)
		≥55	156 (10%)	57 (7%)	136 (12%)
	Male	<55	189 (12%)	100 (16%)	138 (18%)
		≥55	661 (22%)	404 (28%)	368 (31%)
Intermediate risk (5-15%)	Female	<55	185 (5%)	53 (8%)	32 (3%)
		≥55	484 (4%)	348 (7%)	227 (7%)
	Male	<55	146 (10%)	172 (10%)	117 (5%)
		≥55	72 (10%)	307 (12%)	198 (13%)
Low risk (<5%)	Female	<55	97 (2%)	229 (3%)	207 (2%)
		≥55	49 (2%)	284 (3%)	326 (2%)
	Male	<55	60 (5%)	123 (5%)	140 (6%)
		≥55	0 (0%)	22 (9%)	167 (7%)

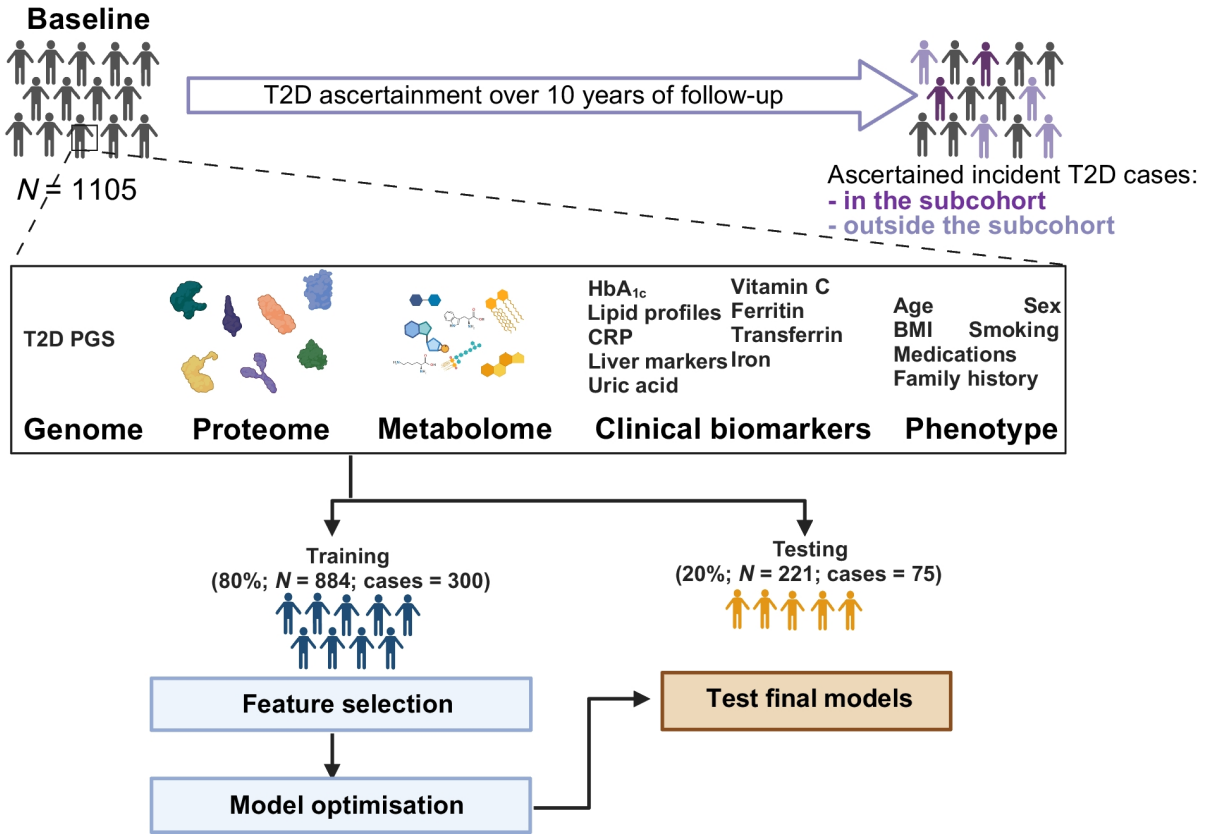
*Percentages denote the observed obstructive CAD prevalence within groups.


BREAK

AGENDA

12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle

MULTI-OMIC PREDICTION OF INCIDENDT TYPE 2 DIABETES



What is HbA1c and why do they use it for stratifying their cohort?

What is HbA1c and why do they use it for stratifying their cohort?

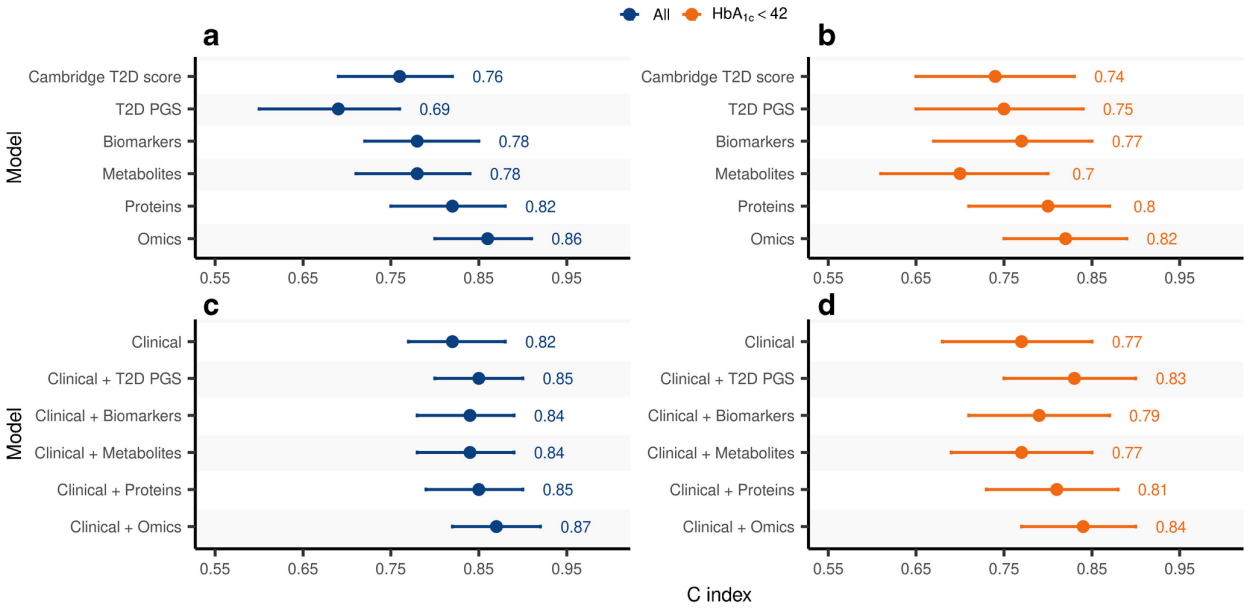
- Hemoglobin A1c
- Used for diagnosis and monitoring of diabetes patients
- Reflects average glucose levels over a longer period (10-12 weeks)
- ≥48 mmol/mol indicates diabetes
- ≥42 mmol/mol indicates prediabetes
- <42 mmol/mol indicates normoglycaemia

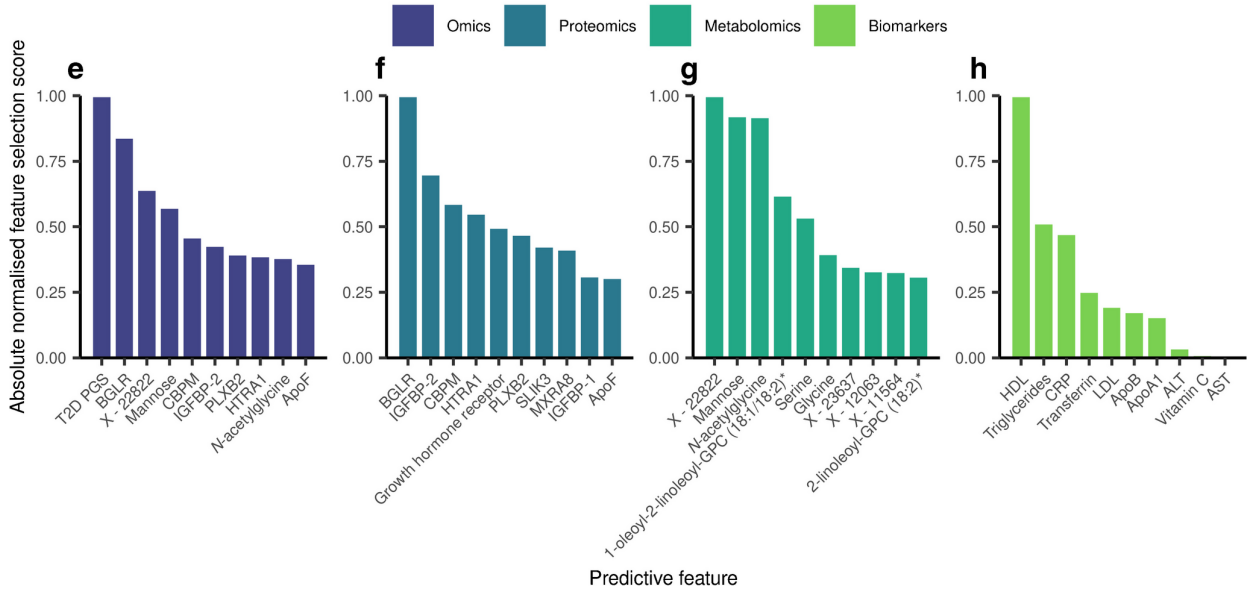
Why is it valuable to identify individuals at risk of type 2 diabetes?

Why is it valuable to identify individuals at risk of type 2 diabetes?

- High prevalence of disease in the general population
- Early treatment reduces risk of complications
- Prediabetic individuals can often be ”treated” with preventative behavioural interventions

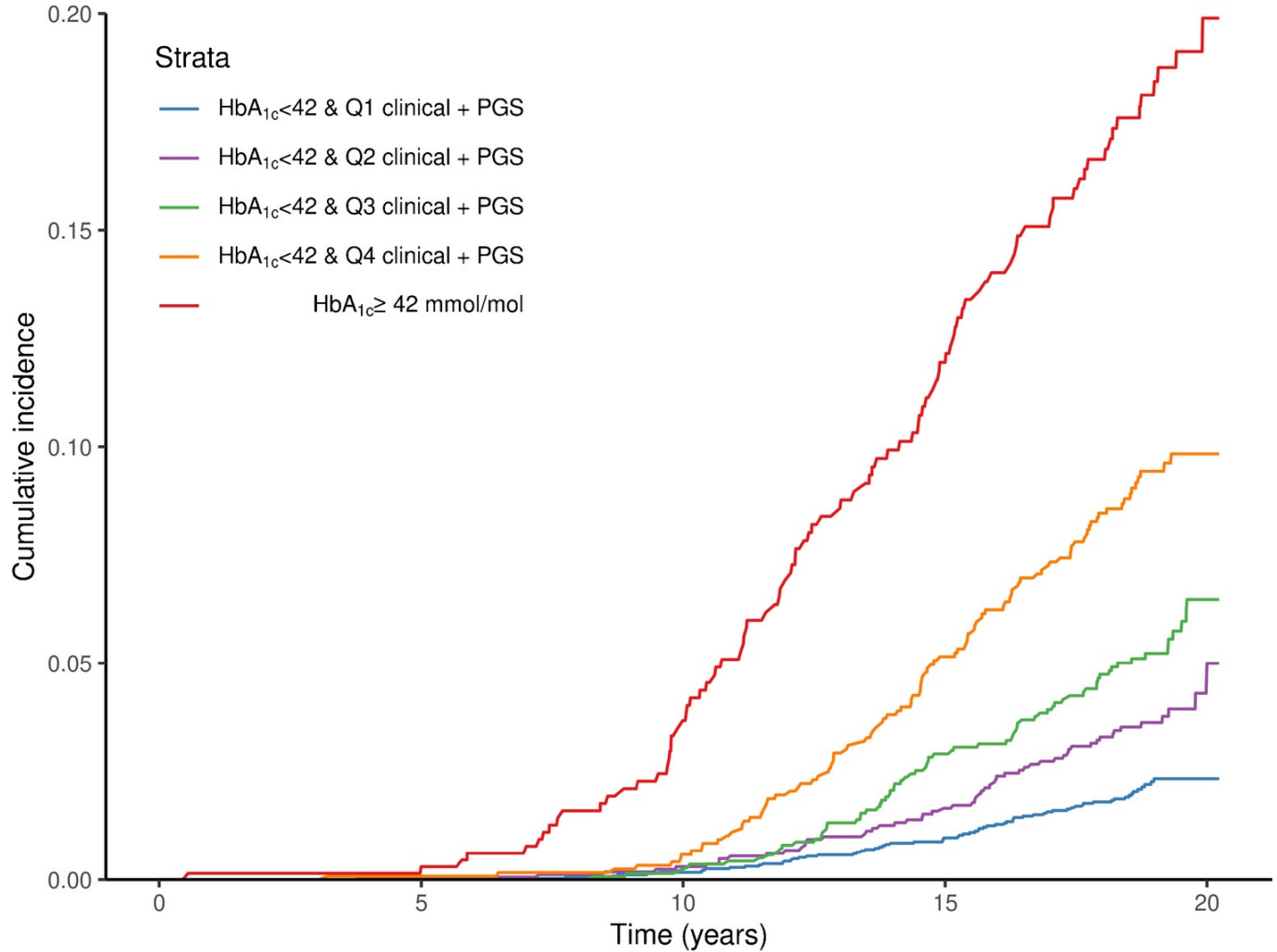
Which type of omics would you prioritize, if you could only have one?





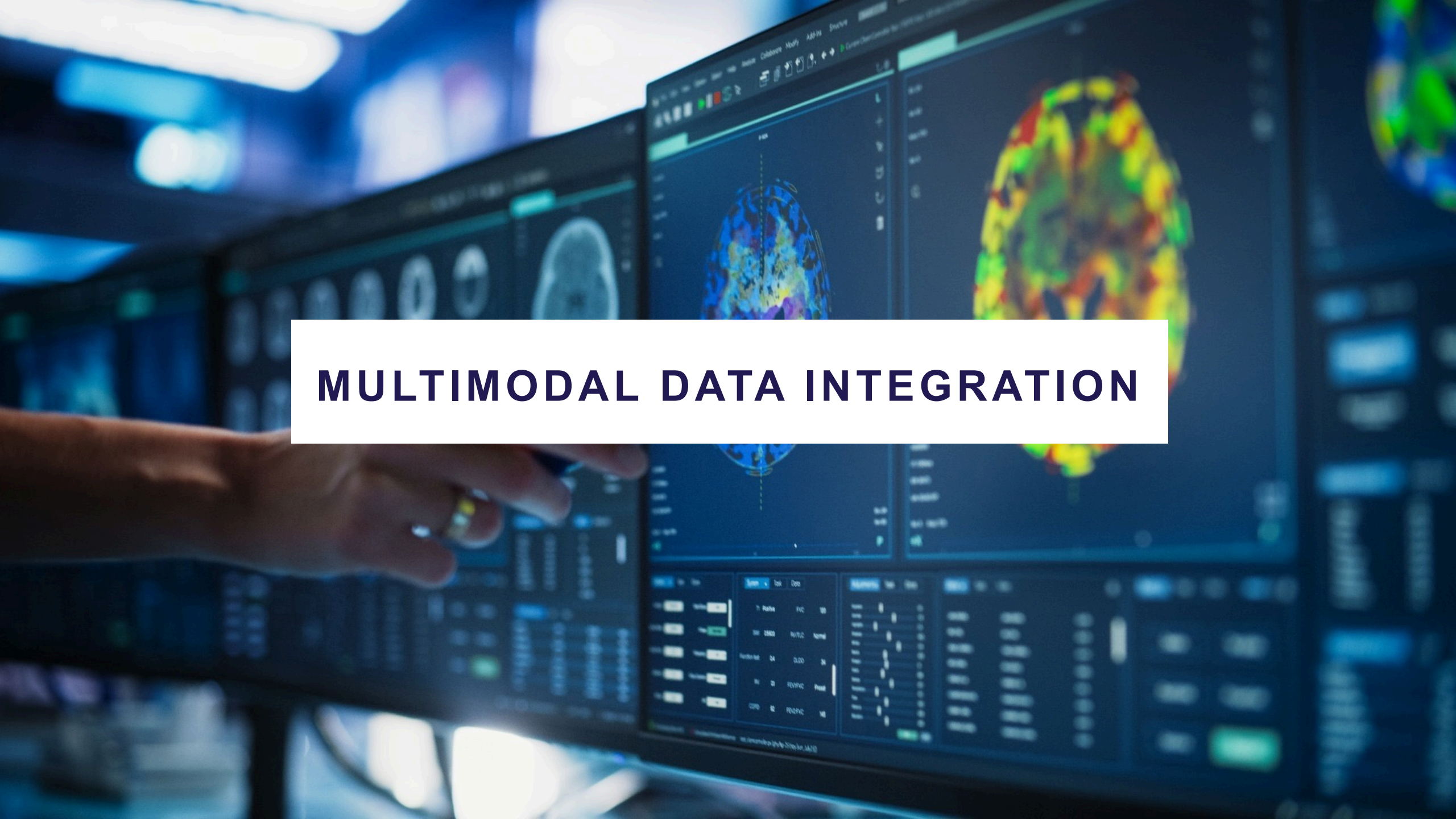
Looking at figure 3, could the researchers have done anything differently?

- “However, individuals at high predicted polygenic risk were at a substantially lower absolute risk than people with prediabetes, suggesting limited potential value in targeted genetic screening for preventative interventions”



AGENDA

12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle



MULTIMODAL DATA INTEGRATION

DATA SO FAR

- Tabular data
- Variables can be

Table 1 Sample medical dataset

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

DATA SO FAR

- Tabular data
- Variables can be
  - Continuous

Table 1 Sample medical dataset

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

DATA SO FAR

- Tabular data
- Variables can be
  - Continuous
  - Categorical

Table 1 Sample medical dataset

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

OTHER TYPES OF DATA

- Text – Electronic health records
- Images – CT/MRI scans
- Speech – Operative dictation
- Video – Endoscopy
- 3D scans – Ultrasound

OTHER TYPES OF DATA

- Text – Electronic health records
- Images – CT/MRI scans
- Speech – Operativ
- Video – Endoscopy
- 3D scans – Ultrasound

Is it feasible to convert to tabular data?

IMAGE ANALYSIS

nature › nature biomedical engineering › articles › article

Article | Published: 19 February 2018

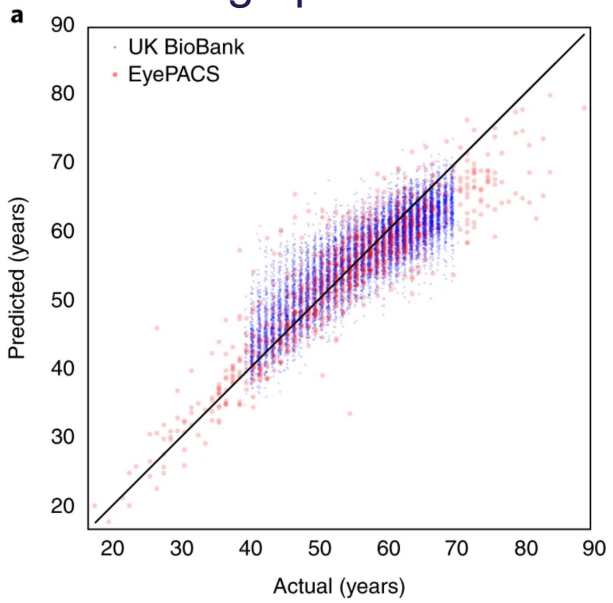
Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng ✉ & Dale R. Webster

Nature Biomedical Engineering **2**, 158–164 (2018) | Cite this article

Input

Age prediction

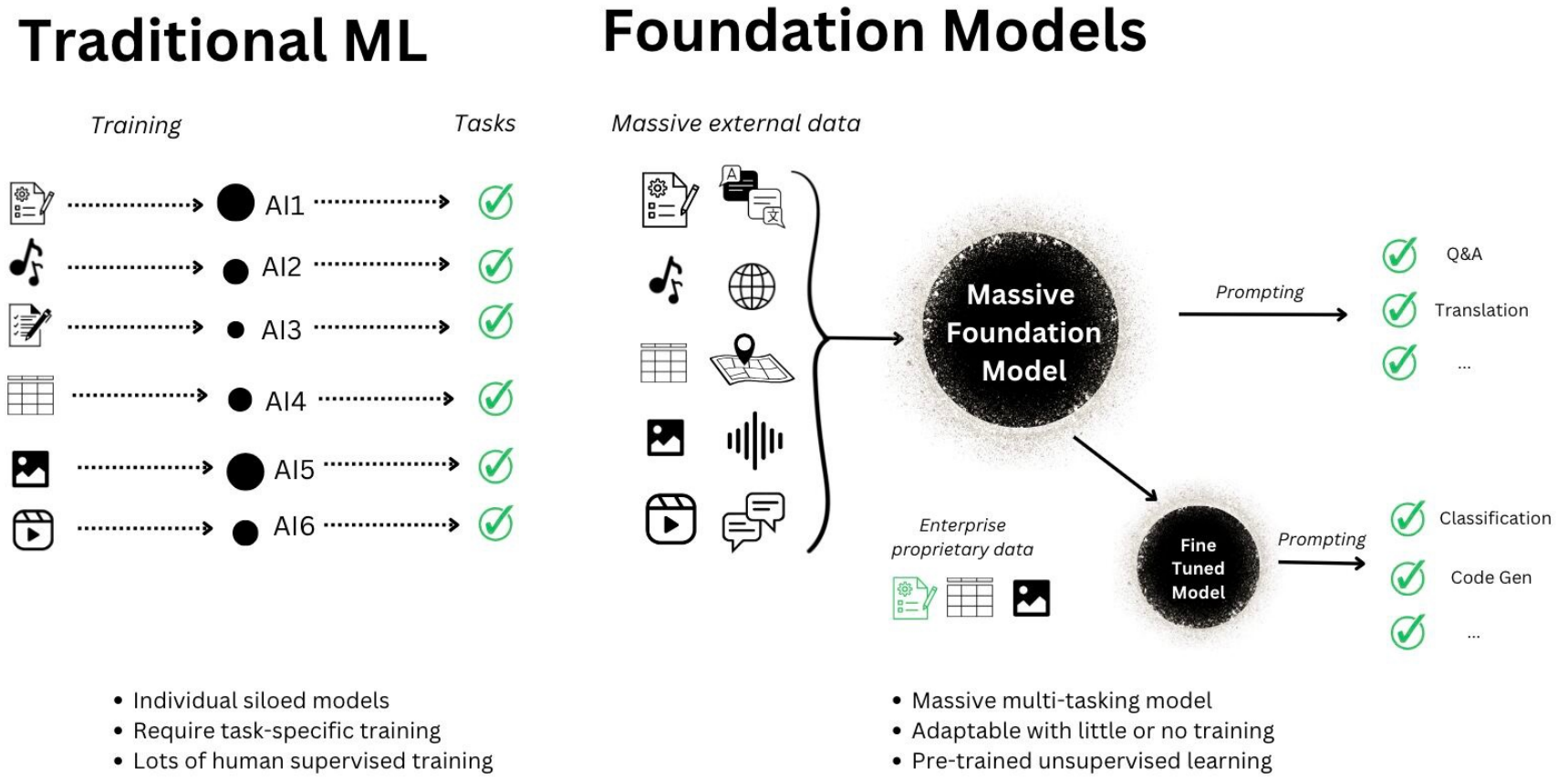


MAE = 3.26; R^2 = 0.74

Five-year MACE prediction

Risk factor(s) or model used for the prediction	AUC (95% CI)
Age only	0.66 (0.61,0.71)
SBP only	0.66 (0.61,0.71)
BMI only	0.62 (0.56,0.67)
Gender only	0.57 (0.53,0.62)
Current smoker only	0.55 (0.52,0.59)
Algorithm only	0.70 (0.65,0.74)
Age + SBP + BMI + gender + current smoker	0.72 (0.68,0.76)
Algorithm + age + SBP + BMI + gender + current smoker	0.73 (0.69,0.77)
SCORE[6,7]	0.72 (0.67,0.76)
Algorithm + SCORE	0.72 (0.67,0.76)

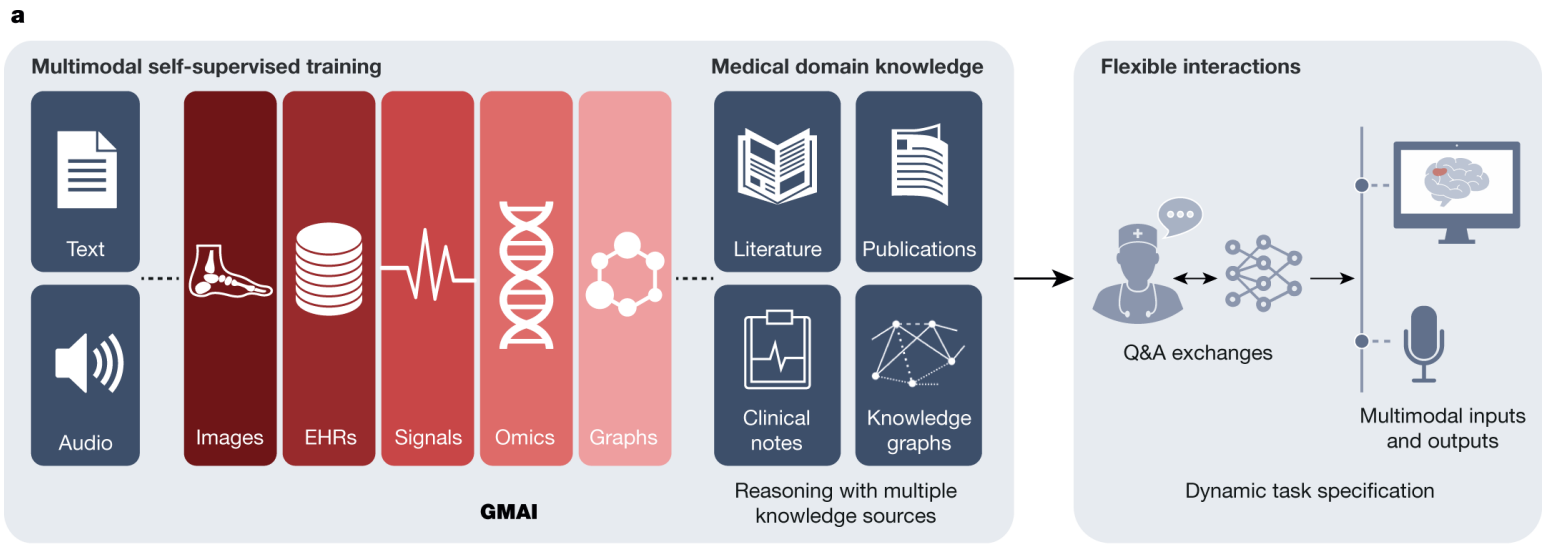
FOUNDATION MODELS



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

FOUNDATION MODELS IN MEDICINE

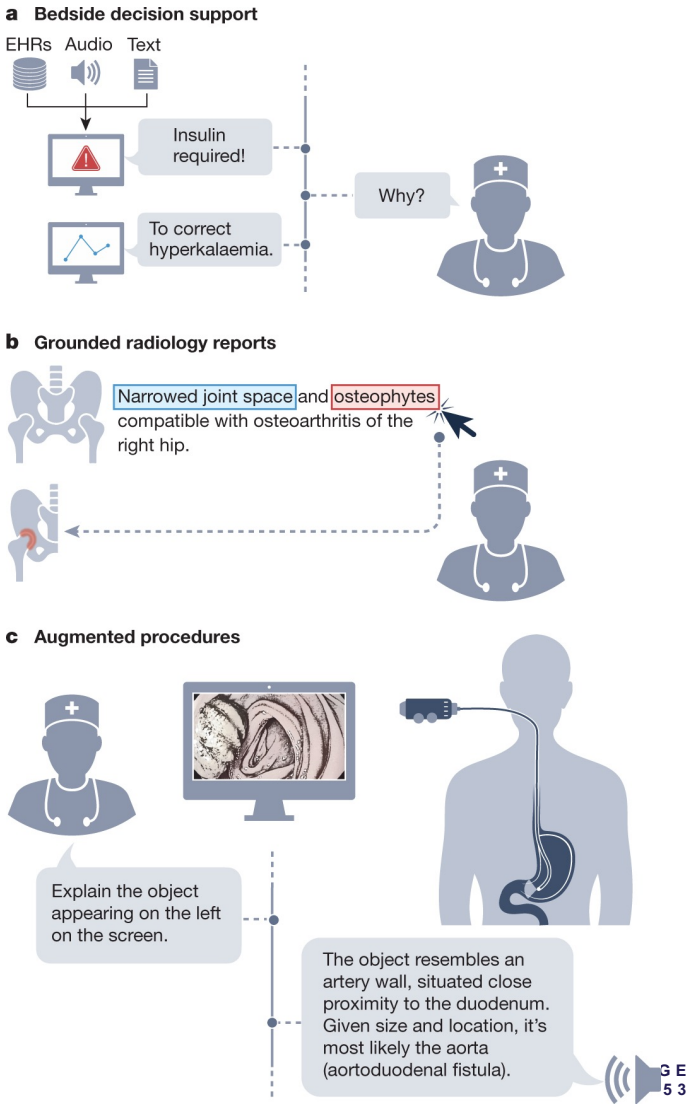


Perspective | Published: 12 April 2023

Foundation models for generalist medical artificial intelligence

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol & Pranav Rajpurkar

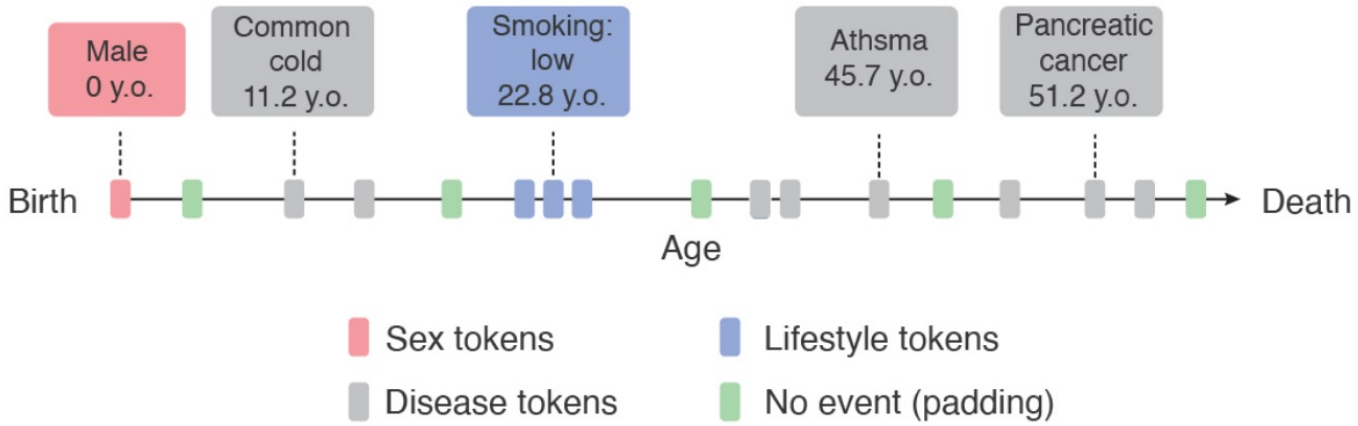
Nature **616**, 259–265 (2023) | Cite this article



FOUNDATION MODELS

Learning the natural history of human disease with generative transformers

Artem Shmatko[1,2,3*], Alexander Wolfgang Jung[2,4,5*], Kumar Gaurav[2*], Søren Brunak[4], Laust Mortensen[5], Ewan Birney[2#], Tom Fitzgerald[2#] and Moritz Gerstung[1,2,6,7,8,9#]



Input:
Age: Token
0.0: Male
2.0: B01 Varicella [chickenpox]
3.0: L20 Atopic dermatitis
5.0: No event
10.0: No event
15.0: No event
20.0: No event
20.0: G43 Migraine
21.0: E73 Lactose intolerance
22.0: B27 Infectious mononucleosis
25.0: No event
28.0: J11 Influenza, virus not identified
30.0: No event
35.0: No event
40.0: No event
41.0: Smoking low
41.0: BMI mid
41.0: Alcohol low
42.0: No event

Output:
43.2: No event
43.5: M54 Dorsalgia
44.6: I86 Varicose veins of other sites
50.4: K52 Other non-infective gastro-enteritis and colitis
52.2: H83 Other diseases of inner ear
53.9: J22 Unspecified acute lower respiratory infection
54.5: L30 Other dermatitis
55.3: No event
57.5: L50 Urticaria
59.4: K62 Other diseases of anus and rectum
...
69.8: J90 Pleural effusion, not elsewhere classified
70.0: K21 Gastro-oesophageal reflux disease
70.1: K76 Other diseases of liver
70.3: I10 Essential primary hypertension
70.4: M85 Other disorders of bone density and structure
70.7: M81 Osteoporosis without pathological fracture
71.2: J98 Other respiratory disorders
72.1: J80 Adult respiratory distress syndrome
72.2: No event
72.7: Death


BREAK

AGENDA

12:30 – 12:45	Recap [*Germline cancer genomics*]
12:45 – 13:00	Break + questions
13:00 – 13:30	Lecture 1 [*Integrative genomics*]
13:30 – 13:40	Break
13:40 – 14:00	Discussion of article [*Multi-omic prediction of incident type 2 diabetes*]
14:00 – 14:20	Lecture 2 [*Multimodal data integration*]
14:20 – 14:30	Break
14:30 – 15:45	Group work + presentation
15:45 – 16:00	Evaluation at Moodle

GROUP WORK

1) Make groups of three-four individuals
   - ❑ For your type of "omics", give a short presentation of the methodology and perspectives, considering the following:
     - ❑ How does the method work?
     - ❑ How much data is generated?
     - ❑ Why is this type of data interesting?
     - ❑ Does it interact with other "omics"?

2) Presentation [5-7 min per group]





TYPES OF “OMICS”

- Proteomics
- Metabolomics
- Microbiomics
- Epigenomics
- Lipidomics
- Glycomics
- Transcriptomics

YOUR OPPINION MATTERS

MOODLE EVALUATION