DNA methylation

Examples from the literature

University of BRISTOL
BBSRC bioscience for the future
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Integrative Epidemiology Unit

Choosing training data: gestational age prediction

Gestational age

Preterm infants (< 37 weeks) have immediate risks:

e.g. respiratory distress, jaundice, infection

As well as long-term risks:

e.g. learning disabilities, retinopathy, behaviour problems

https://www.mayoclinic.org/diseases-conditions/premature-birth/basics/complications/con-20020050

Gestational age can be estimated from:

- Obstetric ultrasound (gold standard)
- Last menstrual period
- Neonatal estimation (R^2 = 0.58; Thawani et al. JGPN, 2013)
- Blood analyte (R^2 = 0.67; Wilson et al. AJOB, 2016)

Two DNA methylation clocks for gestational age

DNA methylation was measured in cord blood

Bohlin *et al. Genome Biology* (2016) 17:207
DOI 10.1186/s13059-016-1063-4

Genome Biology

RESEARCH Open Access

Prediction of gestational age based on genome-wide differentially methylated regions

J. Bohlin[1*], S. E. Håberg[1], P. Magnus[1], S. E. Reese[2], H. K. Gjessing[1], M. C. Magnus[1], C. L. Parr[1], C. M. Page[1], S. J. London[2†] and W. Nystad[1†]

Knight *et al. Genome Biology* (2016) 17:206
DOI 10.1186/s13059-016-1068-z

Genome Biology

RESEARCH Open Access

An epigenetic clock for gestational age at birth based on blood methylation data

Anna K. Knight[1], Jeffrey M. Craig[2], Christiane Theda[3], Marie Bækvad-Hansen[4], Jonas Bybjerg-Grauholm[4], Christine S. Hansen[4], Mads V. Hollegaard[4,5], David M. Hougaard[4,5], Preben B. Mortensen[6], Shantel M. Weinsheimer[7], Thomas M. Werge[7], Patricia A. Brennan[8], Joseph F. Cubells[1,9,10], D. Jeffrey Newport[11], Zachary N. Stowe[12], Jeanie L. Y. Cheong[2,3], Philippa Dalach[2], Lex W. Doyle[2,3], Yuk J. Loke[2], Andrea A. Baccarelli[13], Allan C. Just[14], Robert O. Wright[14], Mara M. Téllez-Rojo[15], Katherine Svensson[14], Letizia Trevisi[16], Elizabeth M. Kennedy[1], Elisabeth B. Binder[10,17], Stella Iurato[17], Darina Czamara[17], Katri Räikkönen[18], Jari M. T. Lahti[18,19,20], Anu-Katriina Pesonen[18], Eero Kajantie[21,22,23], Pia M. Villa[24], Hannele Laivuori[25,26], Esa Hämäläinen[27], Hea Jin Park[28], Lynn B. Bailey[28], Sasha E. Parets[10], Varun Kilaru[28], Ramkumar Menon[29], Steve Horvath[30,31], Nicole R. Bush[32,33], Kaja Z. LeWinn[32], Frances A. Tylavsky[34], Karen N. Conneely[1,9†] and Alicia K. Smith[1,10,28*†]

University of BRISTOL

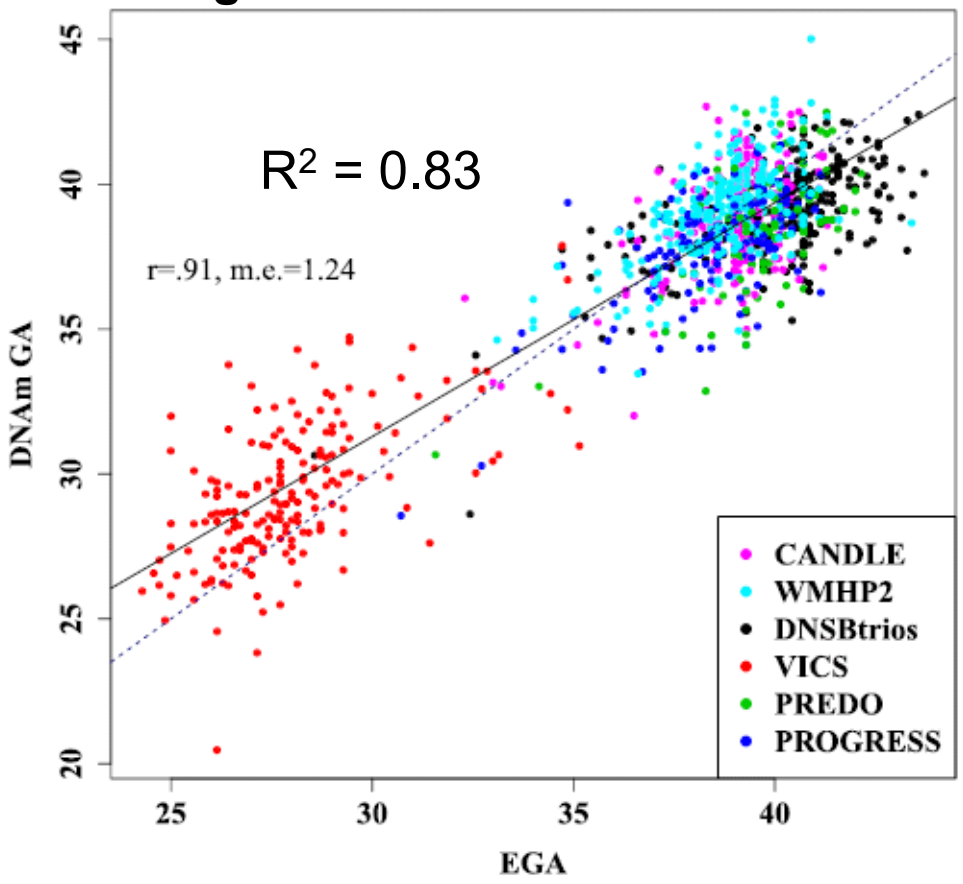
BBSRC bioscience for the future

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

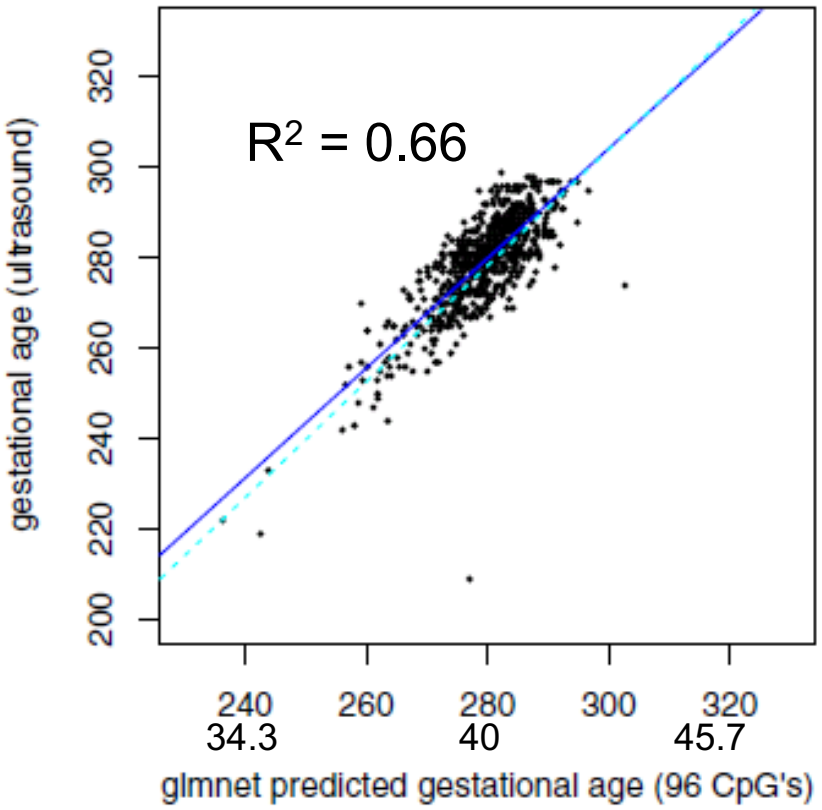
MRC Integrative Epidemiology Unit

Performance

Knight et al



Bohlin et al



University of BRISTOL

BBSRC bioscience for the future

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

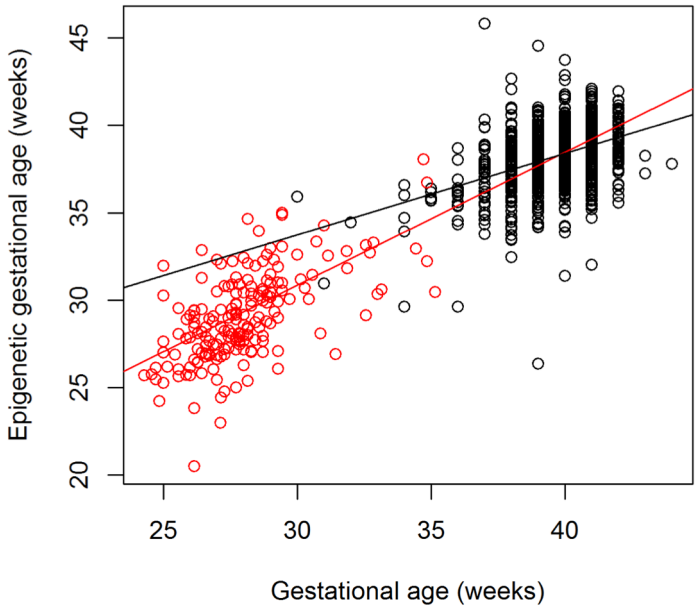
MRC Integrative Epidemiology Unit

An independent dataset (n=863)

Knight et al

R^2 = 0.15 (vs 0.83)

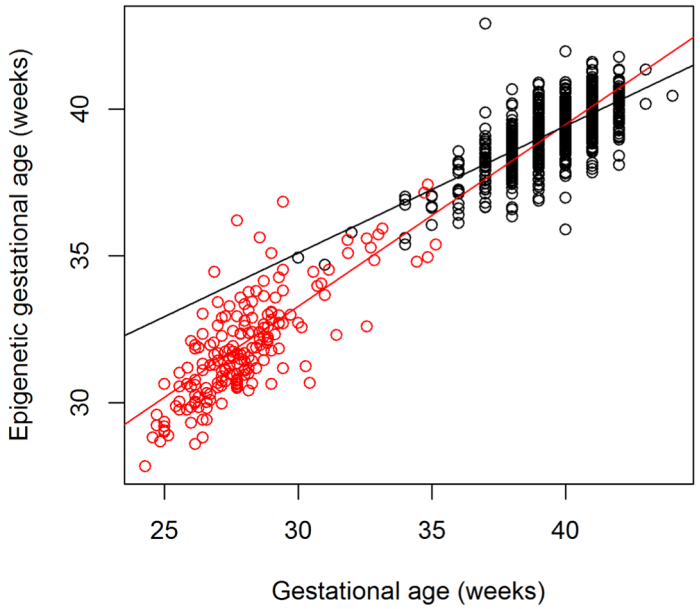
+ n=183 preterms R^2 = 0.79



Bohlin et al

R^2 = 0.46 (vs 0.66)

+ n=183 preterms R^2 = 0.92



University of BRISTOL

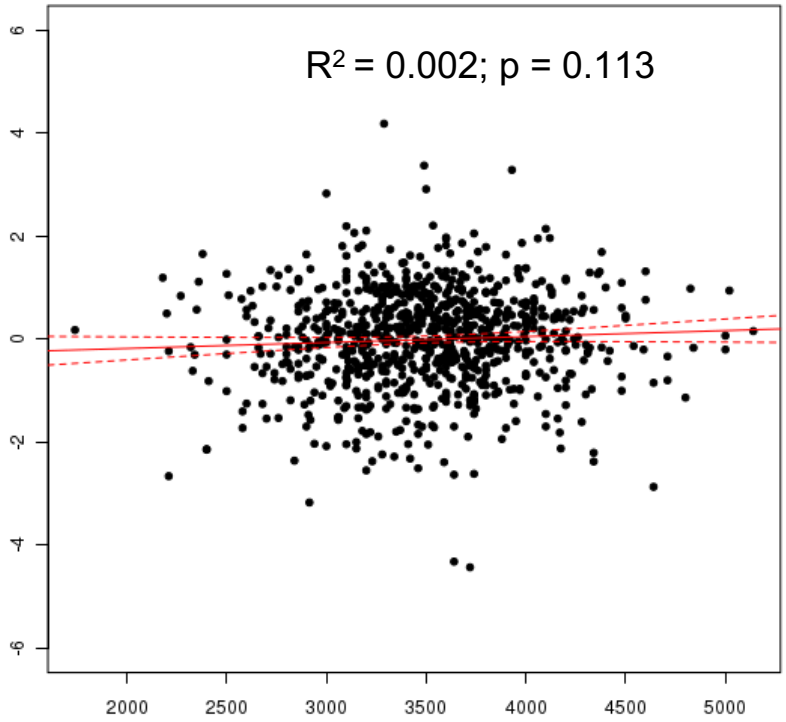
BBSRC bioscience for the future

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

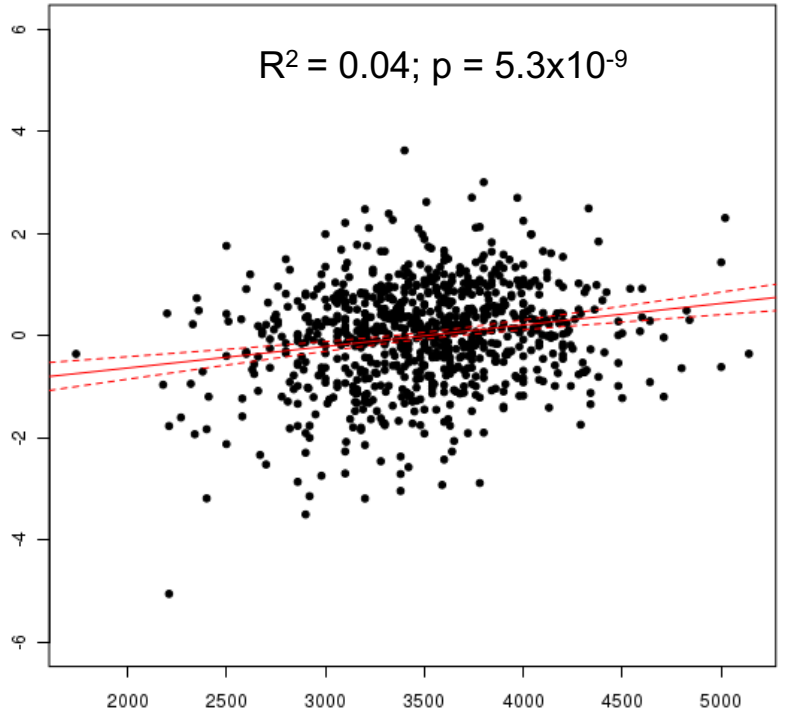
MRC Integrative Epidemiology Unit

GAA and birthweight

Knight et al



Bohlin et al



R^2(Knight GAA, Bohlin GAA) = 0.13

Bohlin datasets

Table 1
Covariates used in the preliminary regression models—MoBa 1

Covariate	Occurrence/mean value	N
Child's sex, male	53.2 %	568/1068
Mean age of mother at birth	29.9 (95 % CI 29.7–30.2)	1068
Maternal smoking during pregnancy	14.6 %	156/1068
Caesarian section	11.5 %	123/1068
Asthma at 3 years	32.9 %	351/1068
Ultrasound estimated GA	279.6 (95 % CI 279–280.3)	1048
LMP estimated GA	282.3 (95 % CI 281.6–283.00)	1030

CI confidence interval

Training dataset
n = 1068

Table 2
Study population —MoBa 2

Covariate	Occurrence/mean value	N
Child's sex, male	56.1 %	384/685
Mean age of mother at birth	30.0 (95 % CI 29.7–30.3)	685
Maternal smoking during pregnancy	10.2 %	70/685
Caesarian section	13 %	89/685
Asthma at 3 years	21.3 %	104/489
Ultrasound estimated GA	279.4 (95 % CI 278.5–280.2)	644
LMP estimated GA	281.5 (95 % CI 280.7–282.4)	615

CI confidence interval

Testing dataset
n = 685

Knight datasets

Table 1 Description of cohorts

Dataset	N	GA range (weeks)	GA mean ± SD	Male (%)
Training datasets				
GSE36642	51	32–38	36.3 ± 1.7	56.9
WMHP1	40	31–41	37.9 ± 2.3	47.5
GSE62924	38	34–41	39.1 ± 1.4	42.1
NBC	36	24–41	36.0 ± 5.4	47.2
GSE51180	23	25–42	32.7 ± 6.6	69.6
GSE30870	19	34–41	38.9 ± 2.1	NA
Test datasets				
DNSBtrios	264	28–44	40.3 ± 1.9	64.9
WMHP2	251	33–43	38.7 ± 1.4	51.0
CANDLE	198	32–41	39 ± 1.3	52.0
VICS	183	24–35	28.0 ± 2.1	42.1
PROGRESS	148	30–43	38.6 ± 1.7	52.0
PREDO	91	31–42	39.6 ± 1.5	54.9

Training datasets and test datasets were chosen to represent a similar range of gestatio
NA not available, *SD* standard deviation

Training dataset
n = 207

148 CpG sites in the prediction model compared to 96 in the Bohlin model

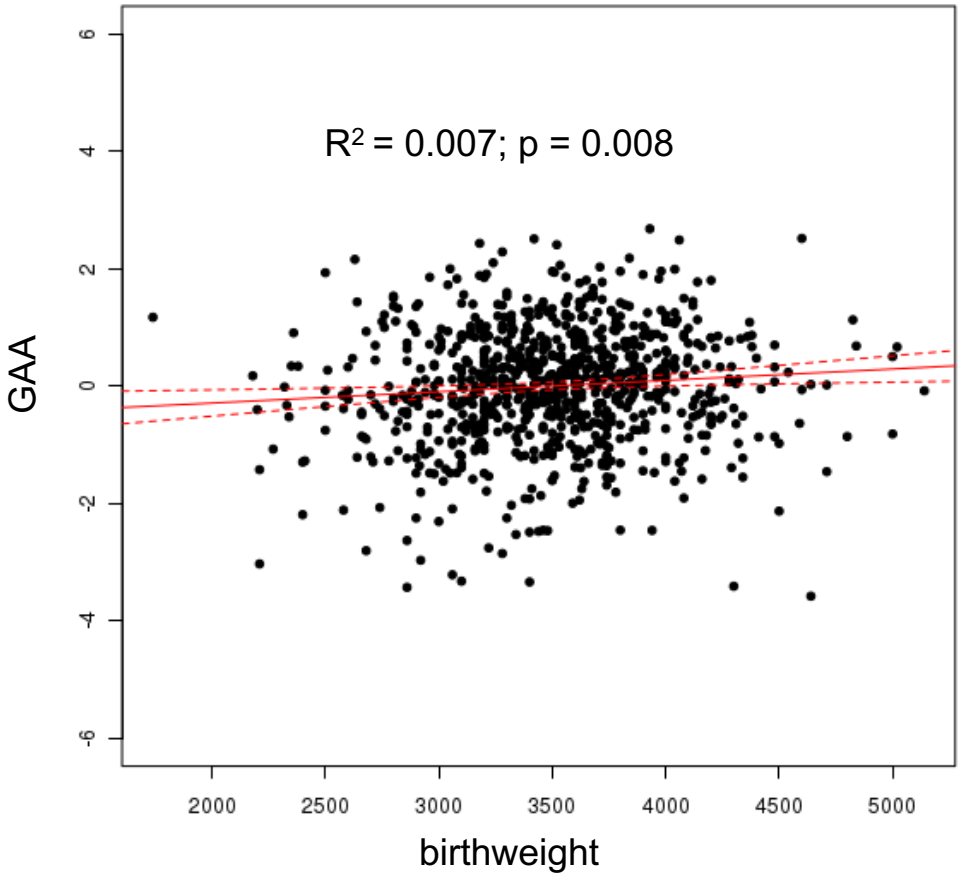
Testing dataset
n = 1135

A new clock based on Knight data

N = 400

50 CpG sites (elastic net constrained otherwise it would have selected 193)

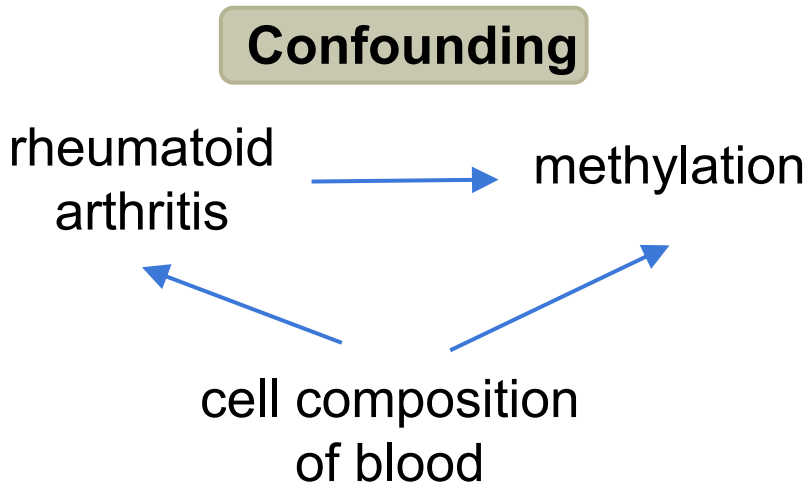
Very similar to both the Knight and Bohlin predictors but …

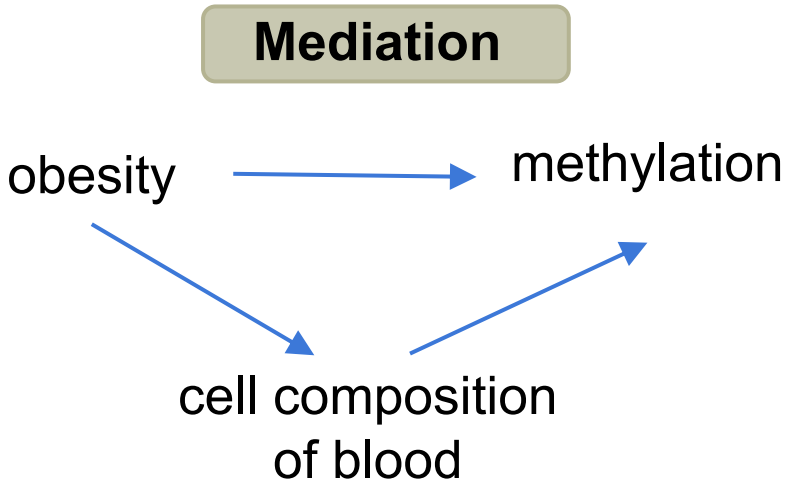


Case study: handling cellular heterogeneity

Handling cellular heterogeneity

Cellular heterogeneity can confound or mediate associations:





Whether you want to remove it from the data or not, depends on the specific question and goals.

Horvath *Genome Biology* 2013, **14**:R115
http://genomebiology.com/2013/14/10/R115

Genome Biology

RESEARCH Open Access

DNA methylation age of human tissues and cell types

Steve Horvath[1,2,3]

Abstract

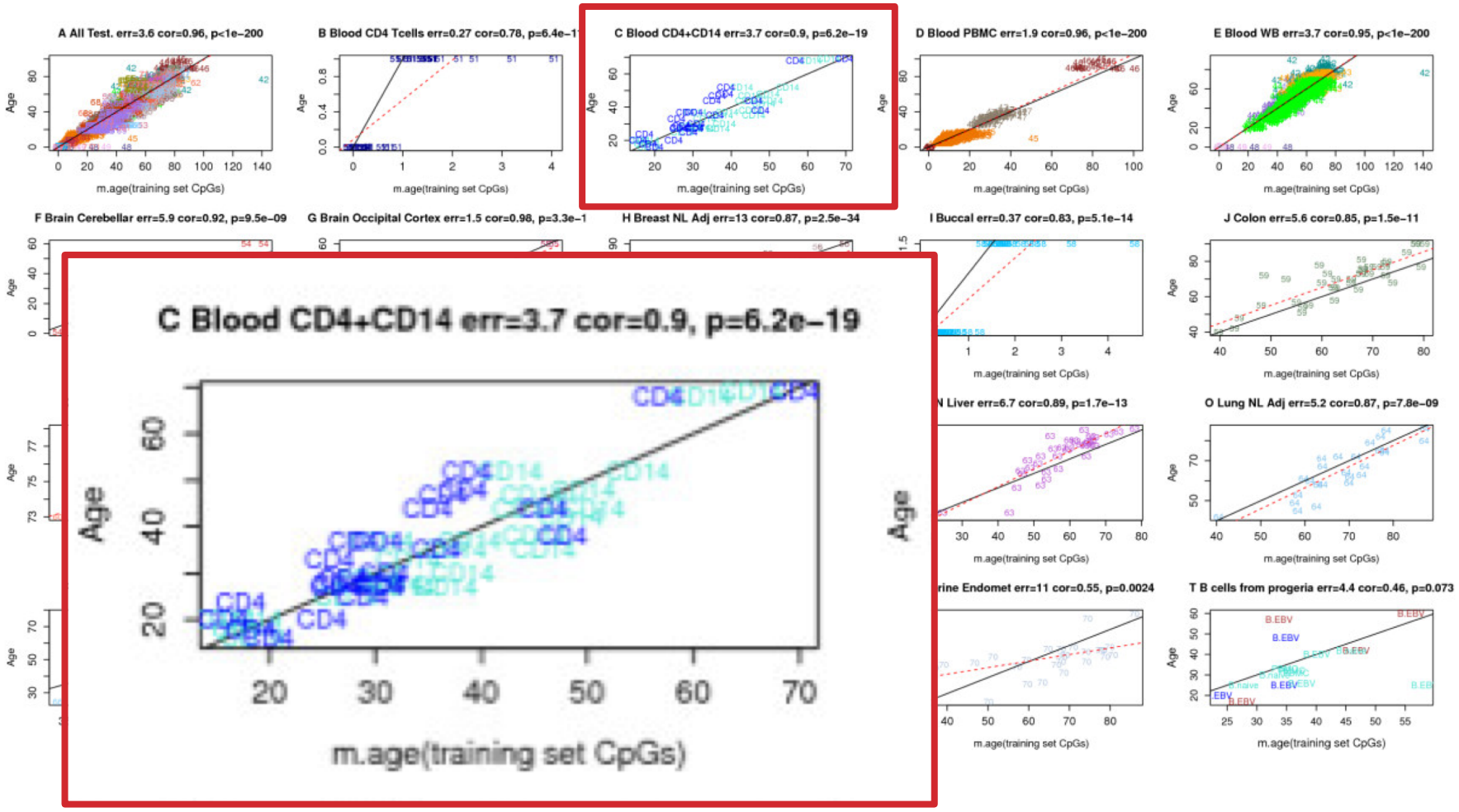
Background: It is not yet known whether DNA methylation levels can be used to accurately predict age across a broad spectrum of human tissues and cell types, nor whether the resulting age prediction is a biologically meaningful measure.

Results: I developed a multi-tissue predictor of age that allows one to estimate the DNA methylation age of most tissues and cell types. The predictor, which is freely available, was developed using 8,000 samples from 82 Illumina DNA methylation array datasets, encompassing 51 healthy tissues and cell types. I found that DNA methylation age has the following properties: first, it is close to zero for embryonic and induced pluripotent stem cells; second, it is independent of cell passage number; third, it gives rise to a highly heritable measure of age acceleration; and fourth, it is applicable to chimpanzee tissues. Analysis of 6,000 cancer samples from 32 datasets showed that all of the cancer types exhibit significant age acceleration, with an average of 36 years. Low age-acceleration of cancers is associated with a high number of somatic mutations and *TP53* mutations, while mutations in steroid receptors are associated with DNA methylation age in breast cancer. Finally, I characterize the 353 CpG sites that together form the predictor with respect to chromatin states and tissue variance.

Conclusions: I propose that DNA methylation age measures the cumulative effect of an epigenetic maintenance system. This novel epigenetic clock can be used to address a host of questions in developmental biology and beyond. I discuss its relevance to research.

"The elastic net regression model automatically selected 353 CpGs"

Age versus DNAm age



C Blood CD4+CD14 err=3.7 cor=0.9, p=6.2e−19
Age
m.age(training set CpGs)

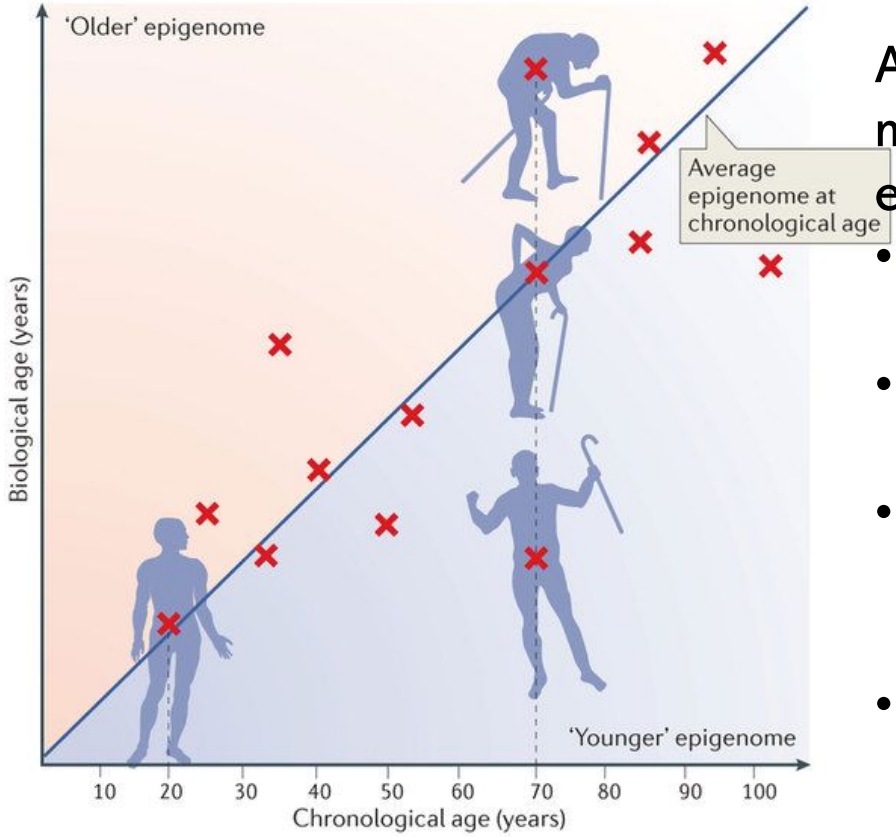
University of BRISTOL

BBSRC bioscience for the future

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MRC Integrative Epidemiology Unit

DNAm age 'acceleration'



Nature Reviews | Molecular Cell Biology

Benayoun BA, Pollina EA, Brunet A. Nat Rev Mol Cell Biol. 2015

Acceleration is associated with many phenotypes and exposures, e.g.

- **All-cause mortality** (Marioni et al. Genome Biol., 2015)
- **Physical and cognitive fitness** (Marioni et al. IJE, 2015)
- **Trauma** (Boks et al. Psychoneuroendocrinology, 2015)
- **Obesity in liver** (Horvath et al. PNAS, 2014)

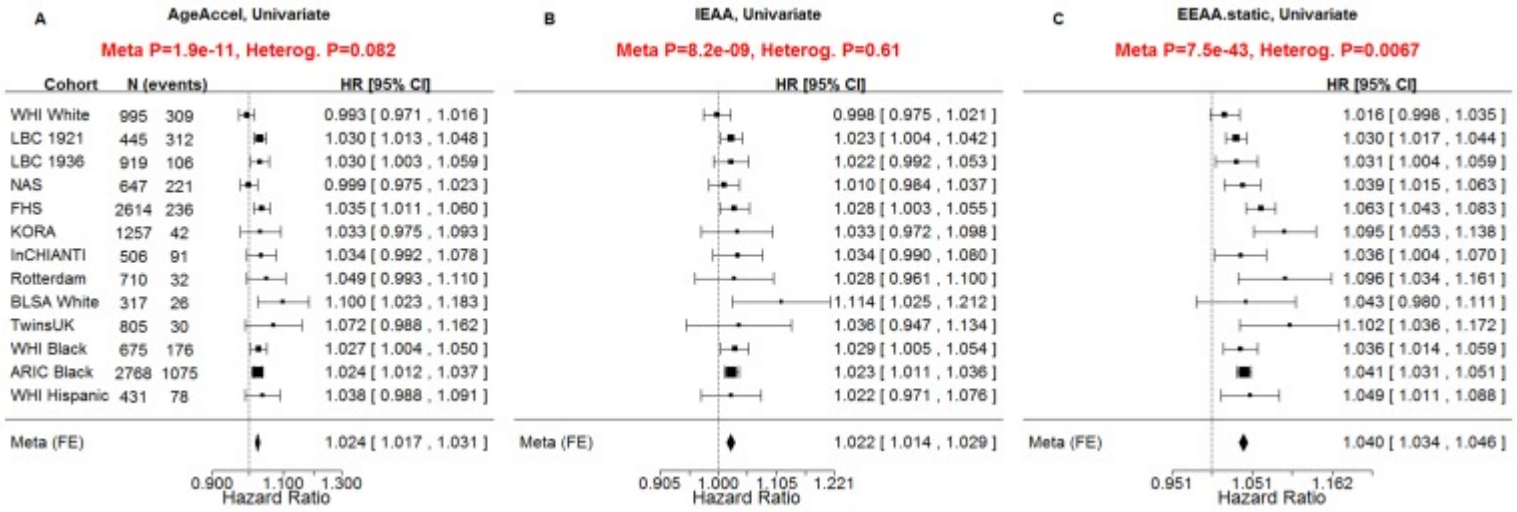
DNAm age and cell heterogeneity

Measure of age acceleration	Short name of measure	Epigenetic age estimate	Correlation with blood counts
Epigenetic age acceleration	*AgeAccel*	Horvath: 353 CpGs	weak
Intrinsic epigenetic age acceleration	*IEAA*	Horvath: 353 CpGs	very weak
Extrinsic epigenetic age acceleration	*EEAA*	Hannum with cell count associated DNAm	strong

Adapted from Chen, et al. Aging (Albany NY). 2016

DNAm age and cell heterogeneity

"the extrinsic measure EEAA out-performs previous measures of age acceleration when it comes to predicting all-cause mortality"



But: "IEAA but not EEAA is predictive of lung cancer
… only IEAA and AgeAccel relate to centenarian status"

Chen, et al. Aging (Albany NY). 2016

Prediction using multiple omics

University of BRISTOL
BBSRC bioscience for the future
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Integrative Epidemiology Unit

The simplest approach

1. Calculate DNAm score for BMI
2. Calculate genetic score for BMI
3. Derive a model for BMI that takes as input the DNAm and genetic scores

   e.g. BMI ~ genetic score + DNAm score

Shah *et al.* (Am J Hum Genet. 2015) found that the two scores were mostly independent:

- Genetic score explained 8% of BMI variance
- DNAm score explained 7%
- Together they explained 14%

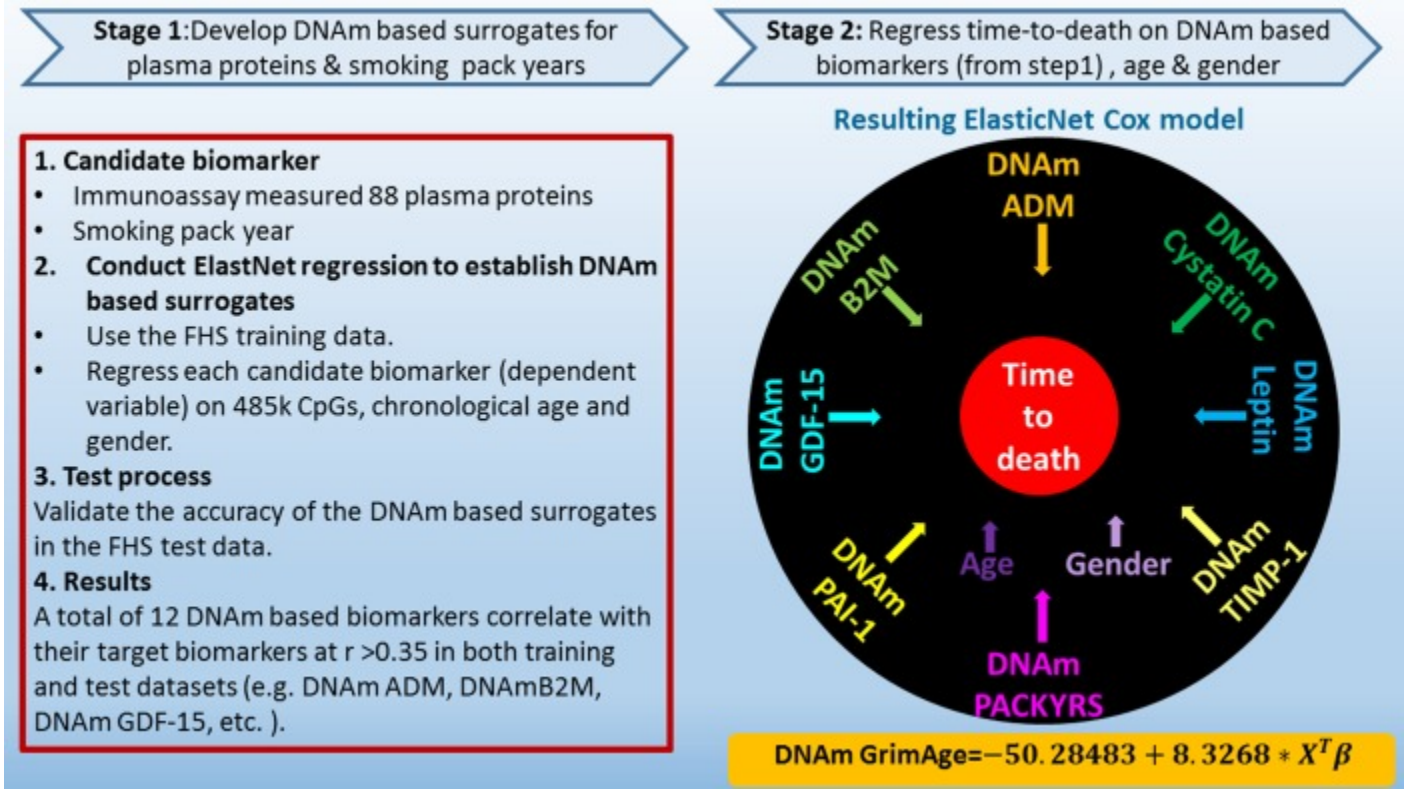
Limit search using other omics

With millions of CpG sites and SNPs in the genome, there are an extremely large number of possible models. How to choose?!

Idea: Assume that DNAm and genetic variation is only related to phenotype if it has measurable biological effects, e.g.

1. Consider only CpG sites associated with protein abundance
2. Consider only SNPs that are associated with DNAm variation

Grimage



The resulting model potentially includes CpG sites that would on their own be ignored by elastic net.

The model performs better than the same model based on protein abundances measured directly.

Lu, et al. Aging (Albany NY). 2019

Genetically predicted DNAm variation

Same approach

1. Derive genetic models for each CpG site
2. Train a model that combines genetic scores for CpG sites

Yang *et al*. (J Natl Cancer Inst. 2019) use this approach to estimate breast cancer risk. *They observe associations with 450 CpG site scores, of which 45 are in loci previously not linked to breast cancer.*

Omics prediction in practice

University of BRISTOL
BBSRC bioscience for the future
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Integrative Epidemiology Unit

Forbes Billionaires Innovation Leadership Money Consumer Industry Lifes

4,276 views | Oct 4, 2017, 10:50am

With The Swab Of A Cheek, This Company Knows When You're Likely To Die

Christopher Steiner Contributor

Jon Sabes' GWG Holdings has plans to disrupt the life insurance industry. gwg

The $635 billion life insurance business revolves around a staid set of practices that haven't evolved much in 40 years, even as technology has upended so many other industries. The big inputs for writing a policy have remained the same: a simple questionnaire, results from a trip to the doctor, and a mortality table.

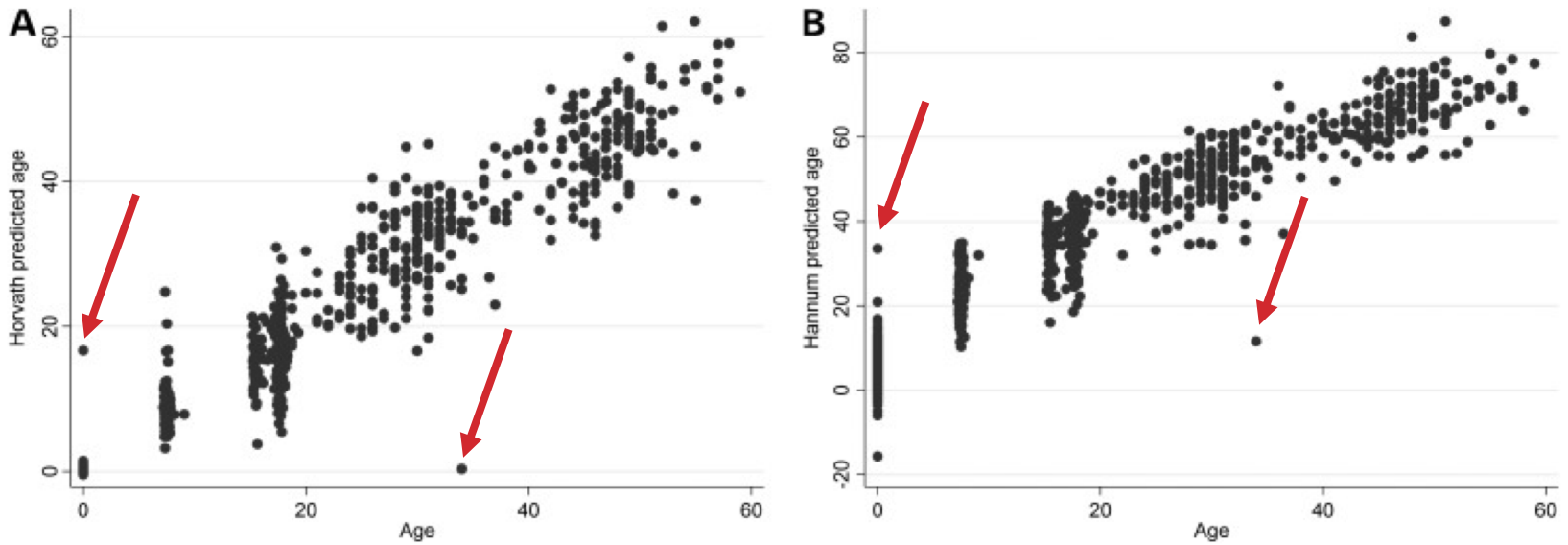
"The data reveal something of an expiration date for a person, and it can be uncannily accurate."

"the goal of building up a portfolio of life insurance policies by buying them from people who want to cash out or can no longer afford the premiums"

"The idea, of course, is to pay more for the policies held by those with advanced biological clocks, and pay less—or not buy at all—the policies belonging to people with biological clocks that lag their birthdays—people who are likely to live longer than normal."

Some challenges

The following are age predictions in a non-clinical population …



Simpkin, *et al*. Hum Mol Genet. 2016