Molecular biomarkers

Cañadas-Garre, M., Smyth, L., Neville, C., Kee, F., Woodside, J., & McKnight, A. J. (2021). Molecular biomarkers. In F. Kee, C. Neville, B. McGuinness, & R. Hogg (Eds.), *Objective measures of health and wellbeing of older adults in Northern Ireland: the NICOLA Study Wave 1* (pp. 128-147). Queen's University Belfast.

Published in:
Objective measures of health and wellbeing of older adults in Northern Ireland: the NICOLA Study Wave 1

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Objective Measures of Health and Wellbeing of Older Adults in Northern Ireland

The NICOLA Study Wave 1

June 2021

Editors
Frank Kee, Charlotte Neville,
Bernadette McGuinness, Ruth Hogg

Northern Ireland Cohort for the Longitudinal Study of Ageing
... Understanding today for a healthier tomorrow

8 Molecular Biomarkers

Authors: Marisa Cañadas-Garre, Laura Smyth, Charlotte Neville, Frank Kee, Jayne Woodside, Amy Jayne McKnight (Research Group Lead)

Citation

Cañadas-Garre M*, Smyth LJ*, Neville CE, Kee F, Woodside JV, McKnight AJ (2021). Chapter 8, Molecular Biomarkers. In: NICOLA Health Assessment Report. 2021. (*joint first authors)

Key Findings

- NICOLA has a strong focus on molecular biomarkers so there is complementary genetic, epigenetic and transcriptomic data available for a subset of individuals.
- We inherit much of our DNA from our parents while a small amount of this material changes as we get older. In NICOLA we have 551,839 directly genotyped and 18,148,478 imputed Single Nucleotide Polymorphisms (SNPs) currently available for 2969 participants.
- Summary statistics for the association of these gene polymorphisms with ~30 phenotypes were generated.
- Epigenetics provides a link between our inherited DNA and environmental influences from a person's diet, medication and lifestyle. NICOLA has the epigenetic quality controlled profiles of 1984 individuals arising from variations in DNA methylation at 862,927 genetic sites.
- Transcriptomics (i.e. measures of the expression of genes) may help explain how genetic and epigenetic changes lead to disease, and we have complementary gene expression data for a subset of individuals. Reference data for men and women at different ages has been generated, enabling the cohort to contribute to international collaborations focussing on a range of conditions including renal and eye disease.

8.1 Introduction

Multiple biomarker studies have been performed using this early baseline data with biological material safely stored for future molecular biomarker studies (Table 8.1). This chapter presents the findings from the molecular analysis of the samples and is divided into several complementary sections - Genetic biomarkers, Epigenetic biomarkers, and Transcriptomic biomarkers. The chapter describes the genotyping and quality control (QC) procedures applied to the genotype data in the first two batches of the NICOLA data release, which contains 3,266 unique samples genotyped at 551,839 single nucleotide polymorphisms (SNPs). We also describe characteristics of the generated genotype data, both in terms of content and quality. This document is relevant to researchers accessing and using the genotype data available in the NICOLA resource.

Table 8.1: Overview of genetic, epigenetic and transcriptomic biomarkers and derived phenotypes currently available in the NICOLA cohort.

Biomarker	Derived variables
Infinium CoreExome-24 Array	Imputation to the 1KGP3 reference panel Imputation to the HRC reference panel Clinically actionable variants
	Annotated and Filtered VCF files (Batches 1 & 2; 1KGP3 & HRC Panels): VCF files with only polymorphic variants Lists of monomorphic variants after imputation Quality Control information of the imputation process VCF files with polymorphic variants annotated with RS VCF files with polymorphic variants with R2>0.3 VCF files with polymorphic variants with R2>0.3 and MAC≥5 VCF files with polymorphic variants with R2>0.3 and MAC≥5 annotated with RS and gene
	Kinships for relatedness in association analysis: Kinship matrix for autosomes, 1KGP3 reference panel Kinship matrix for chrX, 1KGP3 reference panel Kinship matrix for autosomes, HRC reference panel Kinship matrix for chrX, HRC reference panel
	Software specific files (Batches 1 & 2; 1KGP3 & HRC Panels): Concatenated VCF files with polymorphic variants for kinship generation Concatenated VCF files with polymorphic variants and duplicated SNPs removed for clumping pgen files for input in PLINK 2.00 alpha

	Individual Genome-Wide Association Summary statistics for: Total cholesterol HDL-cholesterol LDL-cholesterol Non-HDL cholesterol Triglycerides Height Body mass index Waist-to-hip ratio Early age-related macular degeneration eGFR Serum creatinine Chronic kidney disease Serum urea Subretinal drusenoid deposits (reticular pseudodrusen) Subretinal drusenoid deposits (reticular colour) Macular pigment: peak height Macular pigment: peak volume Naevi Arterial calibre Venular calibre Arteriovenous ratio Arteriolar fractal dimension Venular fractal dimension Arteriolar tortuosity Venular tortuosity
	Lists of dosage information on SNPs provided for candidate SNPs and Genetic Risk Score projects: 55 SNPs previously associated with age-related macular degeneration 64 SNPs previously associated with arsenic levels
	Genome-Wide Association Meta-Analysis Summary statistics for: Subretinal drusenoid deposits (reticular pseudodrusen) Subretinal drusenoid deposits (reticular colour) Macular pigment: peak height Macular pigment: peak volume Naevi Arterial calibre Venular calibre Arteriovenous ratio Arteriolar fractal dimension Venular fractal dimension Arteriolar tortuosity Venular tortuosity

	Beta values (□)
	M values
	Proportional cell counts
	Epigenetic clocks
	Summary statistics for: Alcohol consumption Body-mass index Education level eGFR Naevi Physical activity Risk preference Serum urate Smoking Socioeconomic status Subretinal drusenoid deposits (reticular pseudodrusen) Subretinal drusenoid deposits (reticular colour) Time preference
	Gene expression counts
	Summary statistics for: Renal phenotypes
Abbreviations: 1KGP3: 1000 Genomes Phase3 v5; chr: chromosome; eGFR: estimated glomerular filtration ratio; HDL: High-density lipoproteins; HRC: Haplotype Reference Consortium; LDL: Low-density lipoproteins; VCF: Variant Call Format.	

8.2 Measurement of molecular biomarkers

Blood samples were collected from participants in EDTA tubes and processed within Belfast City Hospital to separate plasma and buffy coats. DNA was extracted from buffy coats with DNA quantified using PicoGreen and normalised to 200 ng/uL aliquots with 0.1 TE in 2D and readable barcode tubes. DNA was stored in multiple aliquots at -80°C to minimise freeze/thaw cycles and maximise the utility of high molecular weight DNA for molecular studies.

Samples were genotyped by Eurofins Scientific (Eurofins Genomics: https://www.eurofinsgenomics.eu). Genotype data (n = 551,839 markers directly typed) was generated using the Illumina Infinium CoreExome-24 for high-throughput screening on an iScan for two batches composed of 2799 and 467 participants respectively. For the purposes of analysis, each batch was processed separately. GenomeStudio® Genotyping Module was used as calling algorithm, using the Genome Reference Consortium Human Build 37 (GRCh37). Each SNP is analysed independently to identify genotypes. Seven control individuals, blinded to the genotyping lab, were included for internal QC. The Infinium CoreExome-24 BeadChip is a customizable

array designed to be used in large-scale genotyping studies which includes all the tag (SNPs) found on the Infinium Core-24 BeadChip, plus over 240,000 markers from the Infinium HumanExome BeadChip (1). The Infinium CoreExome-24 BeadChip can be used to obtain baseline sample data sets for various downstream applications quickly and easily. These applications include common variant, mitochondrial DNA (mtDNA), ancestry, sex confirmation, loss of-variant, and insertion/deletion (indel) detection studies.

Quality Control and Imputation

QC of the genotyped data was performed in PLINK 1.90 beta (2). QC included removal of samples with a call rate < 95%, heterozygosity (> median + 3 x interquartile range) and principal component (PC) analysis outliers; gender mismatches and duplicates or up to second-degree related individuals were also removed, by eliminating one individual from each pair with an Identity By Descent (IBD) value > 0.1875, which is halfway between the expected IBD for third- and second-degree relatives (3,4). Variants with a call rate ≥ 98%, Hardy-Weinberg equilibrium p>10-6 and Minor Allele Frequency (MAF) < 0.0001 were removed.

Files were prepared for imputation using the "HRC/1KG Imputation Preparation and Checking Tool", developed by Will Rayner (5), and then imputed to the 1000 Genomes Phase3 v5 (1KGP3) and Haplotype Reference Consortium (HRC) r1.1 2016 reference panels using the Michigan Imputation Server (6). A minor allele cut-off of 5 and imputation quality of 0.3 was applied to imputed files; monomorphic markers were removed.

Two empirical genomic relationship matrices (kinship matrix), one for the set of autosomes and one for the hemizygous region of X chromosome, were generated using rvtests, for each reference panel (7). Kinship matrices are used in association analyses to account for familial relatedness, cryptic relatedness, and population stratification. The matrices were created using the imputed VCF files after removing monomorphic markers.

Quality Control of directly genotyped SNPs

The NICOLA batch 1 consisted of 2799 individuals and batch 2 was composed of 467 individuals, after filtering by a minimum 94% sample call rate as a pre-QC control step in GenomeStudio. After QC, 2560 samples (352,061 SNPs) for batch 1 and 402 (307,743 SNPs) for batch 2 passed filters (Table 8.2).

Table 8.2: Genotyping Quality Control exclusions for NICOLA batches 1 and 2

	NICOLA Batch	
QC Step	**1**	**2**
Individuals (n)	**2799**	**467**
Sex Discordance	9	16
Sample Call Rate < 95%	63	11
Exclusion of heterozygosity > median ± 3 x IQR	73	34
Related Individuals (IBD < 0.185)	78	
Ancestry Outliers	21	
Final	**2560**	**402**
SNPs	**551,839**	**551,839**
Hardy-Weinberg Equilibrium p > 10-6	797	94
SNP Call Rate ≥ 98%	22,633	33,101
Minor Allele Frequency < 0.00001	176,210	210,793
No chromosome designation	138	108
Final	**352,061**	**307,743**
Abbreviations: IBD: Identity By Descent. IQR: interquartile range. SNP: single nucleotide Polymorphism		

Imputation

The number and nature of variants after imputation to the reference panels and removal of monomorphic markers in the imputed datasets are detailed in Table 8.3. These are the variants used to construct the kinship matrices to be used in association analysis to account for cryptic relatedness.

Table 8.3: Number and nature of polymorphic variants in autosomes and X chromosome after imputation to reference panels used to construct the kinship matrices. *Others refers to a variation of any other type, for example a symbolic allele or a complex substitution.

	NICOLA Batch			
	1		**2**	
Reference Panel	**1KGP3**	**HRC**	**1KGP3**	**HRC**
number of samples	2,567	2,567	402	402
number of SNPs	16,083,011	18,148,478	12,313,319	12,860,489
number of indels	1,747,078	0	1,453,645	0
number of others*	14,804	0	11,124	0
Total	**17,844,893**	**18,148,478**	**13,778,088**	**12,860,489**
Abbreviations: SNP: single nucleotide Polymorphism. 1KGP3: 1000G Phase3 Reference Panel. HRC: Haplotype Reference Consortium				

After applying QC filters for quality of imputation and minor allele count, the total number of SNPs for the 1KGP3 panel are detailed on Table 8.4. These are the variants used for association analyses.

Table 8.4: Number and nature of polymorphic variants in autosomes and X chromosome after filtering by quality of imputation (R2 > 0.3) and minor allele count ≥ 5 to reference panels to be used in association analyses. *Others refers to a variation of any other type, for example a symbolic allele or a complex substitution.

	NICOLA Batch			
	1		2	
Reference Panel	**1KGP3**	**HRC**	**1KGP3**	**HRC**
number of samples	2,567	2,567	402	402
number of SNPs	11,168,920	11,955,381	8,866,368	8,629,476
number of indels	1,325,292	0	1,159,263	0
number of others*	9,348	0	7,265	0
Total	**12,503,560**	**11,955,381**	**10,032,896**	**8,629,476**

Abbreviations: SNP: single nucleotide Polymorphism. 1KGP3: 1000G Phase3 Reference Panel. HRC: Haplotype Reference Consortium

Derived variables

To enhance NICOLA's bioresource, the majority of derived variables have been returned to the main dataset and are available for researchers within our data access agreements.

Imputation dosages

Imputation from directly genotyped SNPs was conducted using the Michigan Imputation Server for the 1000 Genomes Phase3 v5 (1KGP3) and Haplotype Reference Consortium (HRC) r1.1 2016 reference panels.

Kinship Matrices

Four kinship matrices (two sets, one for each reference panel, composed of one kinship for the autosomes and one for the hemizygous region of X chromosome) (Table 8.5) were created to be used to account for relatedness in association analyses (7).

Annotated and Filtered VCF files

Different sets of VCF files, containing the genotyped and imputed SNPs, filtered according to quality parameters (polymorphic variants; R2 > 0.3; minor allele count ≥ 5, see Table 8.5) and annotated with RS information of SNPs and gene information added have been created to be used in subsequent association analysis. Lists of monomorphic variants were also created.

Software specific files

The VCF files containing the genotyped and imputed SNPs have been converted into formats to work with software specific for genome-wide association studies (Table 8.5), such as concatenated VCF files for kinship calculations and generation using rvtests (8), concatenated VCF files with duplicated SNPs removed to be used for clumping with PLINK 1.90 beta and pgen files for input in PLINK 2.00 alpha (2,9).

Summative data

The projects undertaken with the genotyped and imputed biomarkers of NICOLA Batches 1 & 2 have generated summary statistics for their genome-wide association analysis with different phenotypes, detailed in Table 8.5. Summary statistics for genome-wide association meta-analysis for different phenotypes have also been generated (Table 8.5).

Table 8.5: Summary statistics generated for the genome-wide association analysis of the genotyped and imputed biomarkers of NICOLA Wave 1.

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
Total cholesterol	mg/dL	raw	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 Sex (only in 'All') 4 first PCs	GLGC	AJM	MCG LJS
		inverse normal		1		1KGP3 HRC			GLGC	AJM	MCG LJS
HDL-cholesterol	mg/dL	raw	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 Sex (only in 'All') 4 first PCs	GLGC	AJM	MCG LJS
		inverse normal		1		1KGP3 HRC			GLGC	AJM	MCG LJS
LDL-cholesterol	mg/dL	raw	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 Sex (only in 'All') 4 first PCs	GLGC	AJM	MCG LJS
		inverse normal		1		1KGP3 HRC			GLGC	AJM	MCG LJS
Non-HDL cholesterol	mg/dL	raw	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 Sex (only in 'All') 4 first PCs	GLGC	AJM	MCG LJS
		inverse normal		1		1KGP3 HRC			GLGC	AJM	MCG LJS
Triglycerides	mg/dL	natural log	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 Sex (only in 'All') 4 first PCs	GLGC	AJM	MCG LJS
		natural log inverse normal		1		1KGP3 HRC			GLGC	AJM	MCG LJS

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
Height	cm	raw	All Men	1	chr1-chr22	1KGP3 HRC	Linear regression	Age 4 first PCs	GIANT	AJM	MCG LJS
Body mass Index	kg/m2	raw inverse normal	Women	1	chrX	1KGP3 HRC		Age Age^2 4 first PCs	GIANT	AJM	MCG LJS
Waist-to-hip ratio		raw unadjusted inverse normal	All Men Women	1	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Age^2 4 first PCs	GIANT	AJM	MCG LJS
		adjusted by BMI inverse normal		1		1KGP3 HRC		Age Age^2 4 first PCs BMI	GIANT	AJM	MCG LJS
Early age-related macular degeneration	Yes / No	-	All	1	chr1-chr22	1KGP3	Logistic Regressio	Age 2 first PCs	IAMDGC	RH/ AJM	MCG
eGFR	mL/min/1.73m2	natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Sex 10 first PCs		AJM	MCG LJS
Serum Creatinine	mg/dL	natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Linear regression	Age Sex 10 first PCs		AJM	MCG LJS

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
Chronic Kidney Diseasn	Yes / No	-	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Sex 10 first PCs		AJM	MCG LJS
Serum Urea	mmol/L	raw	All	1/2	Set of 2,527 unique autosomal genes chrMT	1KGP3 HRC	Linear regression	Age Sex		AJM	RC
eGFR	mL/min/1.73m2	natural log inverse normal	All	1/2	Set of 2,527 unique autosomal genes chrMT	1KGP3 HRC	Linear regression	Age Sex		AJM	RC
eGFR	mL/min/1.73m2	natural log quantile-normalized	All	1/2	Set of 2,527 unique autosomal genes chrMT	1KGP3 HRC	Linear regression	Age Sex 10 first PCs		AJM	MCG
Chronic Kidney Disease	Yes / No	-	All	1	Set of 2,527 unique autosomal genes chrMT	1KGP3 HRC	Logistic Regressio	Age Sex 10 first PCs		AJM	MCG

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
Serum Creatinine	mg/dL	natural log quantile-normalized	All	1	Set of 2,527 unique autosomal genes chrMT	1KGP3 HRC	Linear regression	Age Sex 10 first PCs		AJM	MCG
eGFR	mL/min/1.73m2	natural log quantile-normalized	All	1	mitochondrial haplotypes	-	Linear regression	Age Sex 10 first PCs		AJM	MCG
Serum Creatinine	mg/dL	natural log quantile-normalized	All	1	mitochondrial haplotypes	-	Linear regression	Age Sex 10 first PCs		AJM	MCG
Chronic Kidney Disease	Yes / No	-	All	1	mitochondrial haplotypes	-	Logistic Regressio	Age Sex 10 first PCs		AJM	MCG
eGFR	mL/min/1.73m2	natural log inverse normal	Men	1/2	chrY	-	Linear regression	Age Sex 10 first PCs		AJM	KA/ MCG
Serum Creatinine	mg/dL	natural log inverse normal	Men	1/2	chrY	-	Logistic Regressio	Age 10 first PCs		AJM	KA/ MCG
Chronic Kidney Disease	Yes / No	-	Men	1/2	mitochondrial haplotypes	-	Logistic Regressio	Age 10 first PCs		AJM	MCG

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
SDD - reticular pseudodrusen	Yes / No	-	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Age2 4 first PCs		AJM	MCG
SDD - reticular colour	Yes / No	-	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Age2 4 first PCs		AJM	MCG
Macular pigment: peak height		natural log quantile-normalized	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Age2 4 first PCs		RH/ AJM	MCG
Macular pigment: peak volume		natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Age2 4 first PCs		RH/ AJM	MCG
Naevi	Yes / No	-	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Age2 4 first PCs		RH/ AJM	MCG
Arterial calibre		raw	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Sex MABP		GMK/ AJM	MCG
Venular calibre		raw	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Sex MABP		GMK/ AJM	MCG
Arteriovenous ratio		raw	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Age Sex 10 first PCs	Age Sex MABP		GMK/ AJM	MCG
Arteriolar fractal dimension		natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Age Sex 10 first PCs	Age Sex MABP		GMK/ AJM	MCG

Trait	Units	Transformation	Sub-cohort	Genotyping Batch	Chr	Reference Panel	Analysis	Covariates	Lead Consortia	NICOLA Contact	Analyst
Venular fractal dimension		natural log inverse normal	All	1/2	chr1-chr22 chrX	-1KGP3 HRC	Logistic Regressio	Age Sex MABP		GMK/ AJM	MCG
Arteriolar tortuosity		natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Sex MABP		GMK/ AJM	MCG
Venular tortuosity		natural log inverse normal	All	1/2	chr1-chr22 chrX	1KGP3 HRC	Logistic Regressio	Age Sex MABP		GMK/ AJM	MCG
Clinically relevant variants		Based on ACMG and pharmacoge-netically active SNPs	All	1/2	chr1-chr22 chrX	1KGP3 HRC	No association performed	-		AJM	CB
		Lists of dosage information on SNPs provided for the analysis of candidate gene polymorphisms or genetic risk scores									
Age-related macular degeneration	-	-	All	1/2	Set of 55 SNPs	-	-			RH/ AJM	MCG
Arsenic Levels	-	-	All	1/2	Set of 64 SNPs	-	-			JW/ AJM	MCG

Abbreviations: 1KGP3: 1000 Genomes Project Phase3 Reference Panel; AJM: Amy Jayne McKnight; CB: Caitlin Bailie; chr: chromosome; eGFR: estimated glomerular filtration ratio; HDL: High-density lipoproteins; HRC: Haplotype Reference Consortium; KA: Kerry Anderson; GMK, Gareth McKay; JW, Jayne Woodside; LDL: Low-density lipoproteins; LJS: Laura Smyth; MCG: Marisa Cañadas-Garre; MT: mitochondrial; PCs: Principal components; RC: Ruaidhri Cappa; RH: Ruth Hogg; SDD: subretinal drusenoid deposits; SNPs: single nucleotide polymorphisms.

8.3 Epigenetic-based biomarkers

This section describes the epigenetic analysis applied to the DNA methylation data generated from Wave 1 NICOLA and the derived variables. This section is relevant to researchers accessing and using the epigenetic data available in the NICOLA resource.

Methods

Blood samples were collected from participants in EDTA tubes and processed within Belfast City Hospital to separate plasma and buffy coats. DNA was extracted from buffy coats with DNA quantified using PicoGreen and normalised to 200 ng/uL aliquots with 0.1 TE in 2D and readable barcode tubes. DNA was stored in multiple aliquots at -80°C to minimise freeze/thaw cycles and maximise the utility of high molecular weight DNA for molecular studies. This is particularly important for epigenetic analysis where the DNA storage methods often have a major impact on DNA methylation levels.

Samples were processed by Eurofins Scientific who extracted the DNA from the buffy coats and quantitated each sample using PicoGreen. The-derived DNA for all individuals was quantitated before a concentration of 800 ng per sample was bisulphite treated (BST) using the EZ DNA Methylation™ Kit (Zymo Research, USA) using the manufacturer's instructions. All samples were analysed together, in the same laboratory.

To assess the methylation status of the CpG sites, the Infinium MethylationEPIC BeadChip array (Illumina, USA) was used following the manufacturer's instructions. This array quantitatively targets 862,927 CpG sites across the genome. Participant samples were randomly distributed across each array. This high throughput platform evaluated individual methylation levels (▯ values) for each CpG site, ranging from 0 for unmethylated to 1 for complete methylation. ▯ values provide a more intuitive biological interpretation. M values are also generated which are more statistically valid for conducting differential methylation analyses (10). For replication purposes, DNA is available for a further independent ~1,500 individuals within the NICOLA cohort. Several significant EWAS associations from microarray data have been validated in-house using bisulfite treated targeted next generation sequencing or Sequenom EpiTYPER.

Quality Control

Raw methylation data was assessed for dye bias and quantile normalised as previously reported (11). Quality control (QC) included evaluation of the bisulphite treatment conversion efficiency, dye specificity, hybridisation, and staining. This was assessed using GenomeStudio v2011 and BeadArray Controls Reporter software platforms (both Illumina). Data was extracted from GenomeStudio using the Partek plugin and analysed using both beta and M values. Alternatively, .idat files were analysed

directly in R packages, paying careful attention to pre-processing and quality control measures if using this approach.

Proportional white cell counts (WCCs) were estimated following the Houseman method (12) using the raw .idat files output from the iScan machine. The minfi Bioconductor (v3.10) package was utilised. Estimation of six WCCs, CD8+ T, CD4+ T and CD19+ B lymphocytes, CD56+ natural killer cells, CD14+ monocytes and CD15+ granulocytes was performed using the estimateCellCounts function.

MethylationEPIC analysis was performed using Partek® Genomics Suite® v7.19.1018 and R (13). Partek® Genomics Suite® was employed to complete Gene Ontology (GO) analysis and pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

Results

In all, 1938 NICOLA participants who consented to DNA methylation analysis passed the QC threshold and were included in downstream analyses (Table 8.6).

Table 8.6: Summary statistics for NICOLA population included in the epigenetic analysis (DNA methylation)

Variable	n (%)
Gender	
Males	952 (49.1)
Females	986 (50.9)
Age group (yrs)	
50 - 59	668 (34.5)
60 - 69	716 (36.9)
70 - 79	423 (21.8)
80 +	131 (6.8)

Proportional WCCs were determined for the whole population (Table 8.7).

Table 8.7: Average proportional white cell counts for the NICOLA population and both males and females individually

Average	CD8T	CD4T	NK	BCELL	Mono	Gran
Males	0.0246	0.1867	0.0773	0.0457	0.0858	0.5771
Females	0.0460	0.2040	0.0707	0.0474	0.0677	0.5591
Total Population	0.0355	0.1955	0.0739	0.0466	0.0766	0.5679
Abbreviations: BCELL: CD19+ B lymphocytes; CD4T: CD4+ T cells; CD8T: CD8+ T cells; GRAN: CD15+ granulocytes; MONO: CD14+ monocytes; NK: CD56+ natural killer cells.						

Summative data

The projects undertaken with the methylation data from NICOLA was generated from a single batch of data generation. This has also provided summary statistics for methylation association analysis with different phenotypes, summarised in Table 8.8. by Dr Laura Smyth within the QUB Molecular Epidemiology and Public Health research team, funded by the Northern Ireland Kidney Research Fund, the Medical Research Council (MC_PC_15025) and the Public Health Agency R&D Division (STL/4760/13), Science Foundation Ireland (SFI15/US/B3130), NIH R01_DK105154, a Science Foundation Ireland and the Department for the Economy, Northern Ireland US partnership award 15/IA/3152 and the Economic and Social Research Council (ES/L008459/1).

Table 8.8: Summary statistics generated for the epigenetic DNA differential methylation analysis of NICOLA Wave 1.

Trait	Software	Sub-cohort	Lead Consortia	NICOLA Lead	Analyst
Alcohol consumption	PGS and R	1,929 individuals	Lifepath	GMK/AJM	GM/LJS
BMI	PGS and R	1,929 individuals		AJM	LJS
BMI	PGS and R	1,929 individuals	Lifepath	GMK/AJM	GM/LJS
CKD	PGS	1,984 individuals		AJM	RC/LJS
Education level	PGS and R	1,929 individuals		AJM	LJS
Education level	PGS and R	1,929 individuals	Lifepath	GMK/AJM	GM/LJS
Epigenetic clocks	R	19,29 individuals		AJM	AJM/LJS
eGFR	PGS	1,984 individuals		AJM	RC/LJS
eGFR	PGS	1,097 individuals between the ages of 60 and 79		AJM	LJS
Naevi	PGS	1,887		RH/AJM	LJS
Physical activity	PGS and R	1,929 individuals	Lifepath	GMK/AJM	GM/LJS
Risk preference	PGS and R	1,656 individuals		FK/AJM	LJS
Serum Urate	R	1,870 individuals	CKDGen	AJM	LJS/SH
Smoking	PGS and R	1,929 individuals	Lifepath	AJM	GM/LJS
Smoking	PGS and R	1,929 individuals	Lifepath	GMK/AJM	GM/LJS
SDD-reticular pseudodrusen	PGS	1,887 individuals		RH/AJM	LJS
SDD-reticular colour	PGS	1,887 individuals		RH/AJM	LJS
Time preference	PGS	1,648 individuals		FK/AJM	LJS

Abbreviations: AJM: Amy Jayne McKnight; BMI: body-mass index; eGFR: estimated glomerular filtration rate; FK: Frank Kee; GMK: Gareth McKay; LJS: Laura Jane Smyth; PGS: Partek Genomics Suite; RC: Ruaidhri Cappa; RH: Ruth Hogg; SDD: subretinal drusenoid deposits; SH: Sophia Halliday.

8.4 Transcriptomic-based biomarkers

This section describes transcriptomic analysis applied to the RNA generated from Wave 1 NICOLA and the derived variables.

Methods

Blood samples (n ~3,800) were collected from participants in PAXgene® blood RNA tubes. RNA was extracted using a proprietary approach by Eurofins Scientific who quantitated and normalised the RNA. An extraction quality of RIN≥8 as measured by a bioanalyser 2100 was required for further analysis. Analysis was conducted in-house for RNA-Seq on a subset of NICOLA including Ambio® ERCC spike-in controls; further samples are still undergoing RNA-seq. Libraries were prepared using the RiboMinus Eukaryote System v2 for whole transcriptome sequencing on a subset of samples, while the AmpliSeq RNA approach was cost-effectively selected for the majority of samples. Sequencing was conducted using Ion Torrent next generation sequencing on an S5XL or Ion Gene Studio S5 system. Sequence alignment was performed using Torrent Suite and Ion Reporter, Partek Flow and Partek Genome Studio, and R packages.

Data available

Sequence data is available, but is subject to our rare variant limitations within NICOLA. Gene counts and exome-based (from RNA) SNP files have been generated, with further data analysis underway. To date, transcriptome-based data has been used to provide support for EWAS results and as a population control for renal phenotypes – this was conducted by LJS and AJM.

8.5 Conclusion

The availability of rich multiomic data within NICOLA's bioresource provides a powerful landmark resource representing our Northern Ireland population aged 50 years and older. A key motivation of NICOLA to generate genetic-epigenetic-transcriptomic data, linked to biochemical biomarkers and extensive phenotype information, was to facilitate a wide spectrum of research. In the first two years, NICOLA's bioresource has been used to identify multiple biological markers associated with more than 30 different phenotypes. NICOLA has also contributed to developing innovative new approaches for multi-omic analyses, critically highlighting the importance of careful DNA and RNA storage for robust experimental studies. Early detection of declining health, particularly in the asymptomatic stages, is very important to facilitate early interventions that promote health and minimise loss of function; NICOLA is identifying novel biomarkers for cardiovascular, eye, and kidney-related outcomes. Our valuable bioresource facilitates exploration of how health and social experiences may lead to increased biological stress and therefore has potential to identify biomarkers for biomedical and biosocial research. Current studies using NICOLA's bioresource involve multiple phenotypes, with a strong focus on cognitive decline, age-related diseases, and socioeconomics. Working with our international colleagues for global studies of ageing we are working to promote health, support research, and inform policymakers.

Acknowledgements

The generation of molecular biomarkers for NICOLA's Wave 1 was primarily funded by the Economic and Social Research Council, award reference ES/L008459/1. The majority of the derived variables for NICOLA's genotype bioresource, including imputation dosages, kinship matrices and annotated VCF files, and GWAS summary statistics were generated by Dr Marisa Cañadas-Garre, under the guidance of Prof AJ McKnight, within the QUB Molecular Epidemiology and Public Health research team at QUB, funded by the Science Foundation Ireland-Department for the Economy (SFI-DfE) Investigator Program Partnership Award (15/IA/3152) and the Economic and Social Research Council (ES/L008459/1). The majority of derived variables for NICOLA's epigenetic and transcriptomic resource were generated by Dr Laura Smyth, under the guidance of Prof AJ McKnight, within the QUB Molecular Epidemiology and Public Health research team, funded by the Northern Ireland Kidney Research Fund, the Medical Research Council (MC_PC_15025) and the Public Health Agency R&D Division (STL/4760/13), Science Foundation Ireland (SFI15/US/B3130), NIH R01_DK105154, a Science Foundation Ireland and the Department for the Economy, Northern Ireland US partnership award 15/IA/3152 and the Economic and Social Research Council (ES/L008459/1).

References

1. Illumina I. Infinium™ CoreExome-24 v1.3 BeadChip [Internet]. 2018. p. 1–4. Available from: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_core_exome_beadchip.pdf [Accessed 1st March 2021]

2. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):1–16.

3. Slifer SH. PLINK: Key Functions for Data Analysis. Curr Protoc Hum Genet. 2018 ;97(1):e59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30040203.

4. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc 2010;5(9):1564–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21085122

5. Rayner W. HRC or 1000G Imputation preparation and checking [Internet]. 2020. Available from: https://www.well.ox.ac.uk/~wrayner/tools/index.html#Checking. [Accessed 1st March 2021]

6. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48(10):1284–7. Available from: http://www.nature.com/articles/ng.3656

7. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics 2016;32(9):1423–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27153000

8. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics;32(9):1423–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27153000

9. Chang C, GRAIL I, Human Longevity I, Department of Biomedical Data Science S. PLINK 2.00 alpha [Internet]. 2020. Available from: https://www.cog-genomics.org/plink/2.0/ [Accessed 1st March 2021]

10. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010 Nov 30;11(1):1–9.

11. Smyth LJ, McKay GJ, Maxwell AP, McKnight AJ. DNA hypermethylation and DNA hypomethylation is present at different loci in chronic kidney disease. Epigenetics 2014 ;9(3):366–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24253112

12. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1).

13. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: A multi-cohort analysis. Aging 2019;11(7):2045–70.