Epigenetics is a hoax.

Dr. Senne Dillen
Promoters:  
Prof. dr. ir. Tim De Meyer  
Prof. dr. ir. Wim Van Criekinge  
Ghent University, Faculty of Bioscience Engineering,  
Department of Mathematical Modelling, Statistics and  
Bioinformatics, Laboratory for Bioinformatics and Compu-  
tational Genomics (BioBix)  
Prof. dr. ir. Sofie Bekaert  
Ghent University Hospital, Centre for Primary Immuno-  
deficiencies, Bimtra Biobank  

Dean:  
Prof. dr. ir. Guido Van Huylenbroeck  

Rector:  
Prof. dr. Anne De Paepe  

Vice-rector:  
Prof. dr. Freddy Mortier
Simon Denil

Epigenetic Profiling of Cardiovascular Ageing

Thesis submitted in fulfillment of the requirements for the degree of Doctor (PhD) of Applied Biological Sciences: Cell and gene biotechnology
Academic year 2014-2015
Epigenetische Profilering van Cardiovasculaire Veroudering

Cover Illustration:
Set against a background consisting of the human telomeric repeat sequence, a heart transitions into a DNA double helix. The helix carries the visage of an ageing face. The heart is textured with a methylated cytosine base. Each of these elements refers to the content of this thesis. Artwork by Christine Chung.

Cite as:

The doctoral candidate and this research were supported by the specialisation grant SB101371 awarded by the Agency for Innovation through Science and Technology (IWT).

The author and the promoters give the authorization to consult and copy parts of this work for personal use only. Any other use is limited by the Laws of Copyright. Permission to reproduce any material contained in this work should be obtained from the author.

The Promoters: The Author:
Prof. dr. ir. Tim De Meyer ir. Simon Denil
Prof. dr. ir. Wim Van Criekinge
Prof. dr. ir. Sofie Bekaert
Members of the examination committee

Prof. dr. ir. Peter Bossier (chair)
Department of Animal Production, Ghent University

Prof. dr. ir. John Van Camp (secretary)
Department of Food Safety and Food Quality, Ghent University

Prof. dr. ir. Sofie Bekaert
Bimetra – Clinical Research Center Ghent, Ghent University Hospital

Prof. dr. Lieven Clement (reading committee)
Department of Applied Mathematics, Computer Science and Statistics, Ghent University
Department of Public Health and Primary Care, KU Leuven

Prof. dr. ir. Tim De Meyer
Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University

Prof. dr. Tim Nawrot (reading committee)
Centre for Environmental Sciences, Hasselt University
Department of Public Health and Primary Care, KU Leuven

Prof. dr. ir. Wim Van Criekinge
Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University
VU Medical Center, Amsterdam, the Netherlands

Prof. dr. Wim Vanden Berghe (reading committee)
Department of Biomedical Sciences, University of Antwerp
Department of Physiology, Ghent University

Dr. ir. Leander Van Neste
MDxHealth
Department of Pathology, Maastricht University, the Netherlands

Prof. dr. Jürgen Veeck (reading committee)
Institute of Pathology, RWTH Aachen University, Germany
Geachte lezer,

Dit doctoraatsproefschrift is enkel tot stand gekomen omdat heel wat mensen mijn levenspad gekruist hebben op het juiste tijdstip. Ik deed mijn best om ze in vloeiende volgorde te vermelden. Indien u uw naam niet terug vindt hieronder en toch meent een bijdrage geleverd te hebben, verontschuldig ik me bij deze. Het zijn immers rijkelijk gevulde jaren geweest.

Wie of wat nu net mijn interesse in de wetenschap gewekt heeft, is een vraag zonder eenduidig antwoord. Mijn ouders, Greet en Johan, hebben me steeds voorzien van een gezonde dosis boeken, encyclopedieën en documentaires. Maar ook de leerkrachten van het college te Mol hebben hierin ongetwijfeld een belangrijke rol gespeeld en meer bepaald Chris De Wilde en Louis Mertens, verdienen een pluim voor hun inzet en didactisch vermogen.

Om nog even in de persoonlijke sfeer te blijven, mag er ook geen erkenning ontbreken voor mijn broers, familie, vrienden en vriendinnen. Jullie hebben kleine en grote bijdragen geleverd tot dit werk. En tenslotte mijn wederhelft, Christine, dank je voor je geduld. Je hebt vaak moeten horen dat er maar een kleine beetje werk meer was dat ik nog snel even moest afwerken. We zijn er nu toch geraakt.

Aan de faculteit Bio-ingenieurswetenschappen hebben vele invloeden mijn keuzes vorm gegeven. Zo ben ik onder andere Wim Soetaert dankbaar om me te helpen inzien dat industriële biotechnologie toch niet mijn ding was. En het waren mijn co-promotoren Sofie Bekaert en Wim Van Criekinge die tijdens hun (gast)lessen mijn nieuwsgierigheid aanwakkerden op de gebieden van telomeerlengte en bio-informatica respectievelijk. Ook tijdens het doctoraat bleven zij een prominente rol spelen, zij het via de Asklepios Studie, zij het met hun schijnbaar onuitputtelijke vermogen om nieuwe onderzoeksprojecten aan te trekken.
Met betrekking tot de Asklepios Studie, die tenslotte het hart van dit onderzoek vormt, ben ik veel dank verschuldigd aan Ernst Rietzschel, Thierry Gilibert, Patrick Segers, Dirk De Bacquer, Caroline Van daele en Marc De Buyzere. De samenwerking was steeds aangenaam en het mag duidelijk zijn dat dit proefschrift nooit tot stand was gekomen zonder hun voortdurende medewerking.

Daarnaast waren er de vele samenwerkingsverbanden met onderzoeksgruppen binnen en buiten de Universiteit Gent die initieel als leerstuk fungeerden maar uit-eindelijk geleid hebben tot een aanzienlijke lijst (toekomstige) co-auteurs. Uit onze voormalige vakgroep, BW14, wil ik vooral Godelieve Gheysen, Tina Kyndt, Annelies Haegeman, Lander Bauters, Winnok De Vos en Tom Sieprath danken. Over de landsgrenzen heen werkten we onder andere met Ed Schuring, Bea Wisman en Martijn Clausen uit Groningen. En in samenwerking met het MUMC waren er de vele medisch getinte projecten die we uitvoerden met Manon Van Engeland, Maurice van Steensel, Hanneke Niessen, Veerle Melotte, Muriel Drath, Marig Vernooij en Tjinta Brinkhuizen. Al deze projecten met zeer uiteenlopende thema’s waarin ik me mocht verdiepen, maakten het werk gevarieerd en uitdagend. En met name Maurice verdient een extra bedankje, hij heeft me de kans gegeven om na dit doctoraat de vleugels nog verder uit te slaan.


Maar al deze mensen, mezelf incluis, hadden er mooi bijgestaan zonder de ondersteuning en expertise van Sarah De Keulenaer, Ellen De Meester, Daisy Flamez, Jean-Pierre Renard en Johan Vandersmissen.

In dit lijstje van collega’s ontbreekt nog één belangrijke naam. Tim De Meyer was niet enkel mijn promotor maar ook mijn mentor. Al sinds mijn master thesis heeft hij me bijgestaan met raad en daad. En vooral over de daad mag uitgeweid worden. Nog nooit ontmoette ik een begeleider die zo plichtsbewust en minutieus het werk van zijn pupil evalueerde. Krappe deadlines, vele projecten, meerdere studenten onder zijn hoede... Tim balanceerde het allemaal en sleepte volledig terecht zijn tenure track in de wacht. Dank u Tim, om mij te vormen tot de doctorandus die ik nu ben.
Tenslotte rest me nog de leden van de jury te bedanken. **Peter Bossier, Lieven Clement, Tim Nawrot, John Van Camp, Wim Vanden Berghe, Leander Van Neste** en **Jürgen Veeck** hebben gezamenlijk gewaakt over de kwaliteit van dit proefschrift en het bijhorende doctoraat. Ook aan hen mijn oprechte dank.

_Gent, februari 2015_  
_Simon Denil_
# Table of Contents

**Dankwoord** i

## 1 Introduction 1-1

- **1.1 Epigenetics** .................................................. 1-1
  - 1.1.1 The Code of Life ........................................... 1-1
  - 1.1.1.1 Genes .................................................. 1-2
  - 1.1.1.2 Central Dogma of Molecular Biology .................... 1-2
  - 1.1.1.3 Epigenetics ............................................ 1-2
  - 1.1.2 Histone Modifications ..................................... 1-3
  - 1.1.3 DNA Methylation .......................................... 1-4
  - 1.1.4 Non-coding RNA ........................................... 1-5
    - 1.1.4.1 microRNA ........................................... 1-5
    - 1.1.4.2 Long Non-coding RNA ................................. 1-5
  - 1.1.5 Telomeres ................................................ 1-6
    - 1.1.5.1 Telomeres as an Epigenetic Feature ................. 1-6
    - 1.1.5.2 Terminal Restriction Fragments ....................... 1-7
    - 1.1.5.3 qPCR ................................................. 1-8
    - 1.1.5.4 Other TL Measurement Techniques .................... 1-9
- **1.2 Cardiovascular Health** .................................... 1-11
  - 1.2.1 Basic Anatomy ........................................... 1-11
  - 1.2.2 Filling Patterns ......................................... 1-11
  - 1.2.3 Atherosclerosis .......................................... 1-13
  - 1.2.4 Epigenetics of Cardiovascular Function, Ageing and Disease 1-15
    - 1.2.4.1 Telomere Length .................................... 1-15
    - 1.2.4.2 DNA Methylation .................................... 1-15
    - 1.2.4.3 Other Epigenetic Traits ............................. 1-16
- **1.3 High-throughput Analysis** ................................ 1-16
  - 1.3.1 Statistical Considerations ............................... 1-17
    - 1.3.1.1 Traditional Framework ............................... 1-17
    - 1.3.1.2 Framework for High-throughput Analysis ............ 1-17
  - 1.3.2 Bioinformatics for Sequencing Data ....................... 1-19
  - 1.3.3 Sequencing Data ......................................... 1-19
  - 1.3.4 RNA Sequencing ......................................... 1-22
1.3.4.1 RNA Preparation ......................... 1-22
1.3.4.2 Normalisation .......................... 1-23
1.3.4.3 Comparing Expression Data ............. 1-24
1.3.5 Methylation Enrichment Sequencing ......... 1-24
1.3.6 Bisulfite Conversion Based Methods .......... 1-26
1.3.7 Limitations of Sequencing Data Analysis .... 1-28
1.3.8 Pathway Analyses .......................... 1-29
  1.3.8.1 Pathway Repositories ................... 1-29
  1.3.8.2 Singular Enrichment Analysis .......... 1-29
  1.3.8.3 Gene Set Enrichment Analysis .......... 1-30
  1.3.8.4 Modular Enrichment Analysis .......... 1-31
  1.3.8.5 Other Considerations ................... 1-31
1.4 Study Rationale ............................ 1-33
  1.4.1 The Asklepios Study on Successful Cardiovascular Ageing .... 1-33
  1.4.2 Asklepios and Epigenetics ................ 1-34
References .................................... 1-35

2 Altered Filling Patterns in Ageing 2-1
  2.1 Introduction .............................. 2-1
  2.2 Methods .................................. 2-2
    2.2.1 Study Population ........................ 2-2
    2.2.2 Biochemical Analyses .................... 2-3
    2.2.3 Telomere Length ........................ 2-3
    2.2.4 Echocardiographic and Vascular Examination ........ 2-3
    2.2.5 Data Analysis .......................... 2-4
  2.3 Results .................................. 2-4
    2.3.1 Main Correlates of Telomere Length .......... 2-6
    2.3.2 Diastolic Function ........................ 2-6
    2.3.3 Systolic Function and LV Structure .......... 2-8
    2.3.4 Arterial Stiffness ........................ 2-8
  2.4 Discussion ............................... 2-10
    2.4.1 LV Filling and Diastolic Function .......... 2-10
    2.4.2 LV Systolic Function, Structure and Vascular Stiffness .... 2-11
    2.4.3 Socio-economic Status  .................... 2-12
    2.4.4 Limitations ............................ 2-12
    2.4.5 Mechanistic Insights  .................... 2-12
  2.5 Conclusion ............................... 2-13
  2.6 Citation .................................. 2-14
References .................................... 2-15
### 3 Peripheral Blood Leukocyte DNA Methylome Reflects Atherosclerosis Progression

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>3-1</td>
</tr>
<tr>
<td>3.2</td>
<td>Methods</td>
<td>3-3</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Study Population Selection</td>
<td>3-3</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Biochemical Analyses</td>
<td>3-3</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Echocardiographic and Vascular Examination</td>
<td>3-3</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Experimental Design</td>
<td>3-4</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Methylated CpG Enrichment Sequencing</td>
<td>3-4</td>
</tr>
<tr>
<td>3.2.5.1</td>
<td>Experimental Procedures</td>
<td>3-4</td>
</tr>
<tr>
<td>3.2.5.2</td>
<td>Analysis Strategy</td>
<td>3-6</td>
</tr>
<tr>
<td>3.2.6</td>
<td>Targeted Bisulfite Deep Sequencing</td>
<td>3-7</td>
</tr>
<tr>
<td>3.2.7</td>
<td>Pathway Analyses</td>
<td>3-9</td>
</tr>
<tr>
<td>3.3</td>
<td>Results</td>
<td>3-10</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Differential Methylation Profiles</td>
<td>3-10</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Validation</td>
<td>3-13</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Pathway Analyses</td>
<td>3-15</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Potential Genomic Deletion</td>
<td>3-16</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>3-18</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Gene Methylation and Atherosclerosis Progression</td>
<td>3-19</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Relation to Biological Processes</td>
<td>3-20</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Relation to Prior Research</td>
<td>3-21</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Predicting Atherosclerosis Progression</td>
<td>3-21</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusions</td>
<td>3-23</td>
</tr>
<tr>
<td>3.6</td>
<td>Acknowledgements</td>
<td>3-23</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>3-24</td>
</tr>
</tbody>
</table>

### 4 A Longitudinal Assessment of Ageing Related Leukocyte DNA Methylation Alterations in the Middle-aged Asklepios Study Population

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>4-1</td>
</tr>
<tr>
<td>4.2</td>
<td>Methods</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Study Population Selection</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Experimental Design</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Methylated CpG Enrichment Sequencing</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2.3.1</td>
<td>Pre-processing</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2.3.2</td>
<td>Analysis Strategy</td>
<td>4-4</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Bisulphite Sequencing</td>
<td>4-5</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Pathway Analyses</td>
<td>4-6</td>
</tr>
<tr>
<td>4.3</td>
<td>Results</td>
<td>4-7</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Differential Methylation Profiles</td>
<td>4-7</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Relation to Prior Research</td>
<td>4-8</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Validation</td>
<td>4-9</td>
</tr>
</tbody>
</table>
4.3.4 Pathway Analyses ............................................. 4-10
4.4 Discussion .................................................... 4-10
4.5 Conclusions .................................................. 4-13
References ......................................................... 4-14

5 Conclusions and Perspectives 5-1
5.1 Introduction .................................................. 5-1
5.2 Telomere Length ................................................. 5-2
5.3 DNA Methylation and Cardiovascular Ageing .......................... 5-3
5.4 Conclusion ..................................................... 5-6
5.5 Perspectives ................................................... 5-7
References ......................................................... 5-10

6 Summary − Samenvatting 6-1

A R-Code A-1
A.1 Chapter 2 ..................................................... A-1
A.2 Chapter 3 & 4 .................................................. A-12
  A.2.1 MethylCap-seq Analyses .................................. A-12
  A.2.2 Permutation Progressor Status ............................. A-19
  A.2.3 Plot Density Permutation Progressors ..................... A-22
  A.2.4 Paired Time Analysis ...................................... A-24
  A.2.5 Paired Time Permutation Approach ....................... A-26
  A.2.6 BS-seq Analyses .......................................... A-29

B Supplementary Figures B-1

C Supplementary Tables C-1

D Curriculum Vitae D-1
# List of Acronyms

**B**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
</tbody>
</table>

**C**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>Coronary Artery Disease</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-Guanine (dinucleotide)</td>
</tr>
<tr>
<td>CRP</td>
<td>C-reactive Protein</td>
</tr>
<tr>
<td>CVD</td>
<td>Cardiovascular Disease</td>
</tr>
</tbody>
</table>

**D**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBP</td>
<td>Diastolic Blood Pressure</td>
</tr>
<tr>
<td>DMR</td>
<td>Differentially Methylated Region</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>dsDNA</td>
<td>double stranded DNA</td>
</tr>
</tbody>
</table>

**E**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF</td>
<td>Ejection Fraction</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FWER</td>
<td>Family-wise Error Rate</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td></td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td><strong>H</strong></td>
<td></td>
</tr>
<tr>
<td>HDL</td>
<td>High-density Lipoprotein</td>
</tr>
<tr>
<td>HF</td>
<td>Heart Failure</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>Interleukine</td>
</tr>
<tr>
<td>IQR</td>
<td>Inter Quartile Range</td>
</tr>
<tr>
<td>IVRT</td>
<td>Isovolumic Relaxation Time</td>
</tr>
<tr>
<td><strong>K</strong></td>
<td></td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td><strong>L</strong></td>
<td></td>
</tr>
<tr>
<td>lncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>LDL</td>
<td>Low-density Lipoprotein</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>LVM</td>
<td>Left Ventricular Mass</td>
</tr>
<tr>
<td>MBD</td>
<td>Methyl Binding Domain</td>
</tr>
<tr>
<td>MC</td>
<td>Methylation Core</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NB</td>
<td>Negative Binomial (distribution)</td>
</tr>
<tr>
<td>ox-LDL</td>
<td>Oxidised LDL</td>
</tr>
<tr>
<td>PBL</td>
<td>Peripheral Blood Leukocyte</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PP</td>
<td>Pulse Pressure</td>
</tr>
<tr>
<td>PWV</td>
<td>Pulse Wave Velocity</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative PCR</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SBP</td>
<td>Systolic Blood Pressure</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single stranded DNA</td>
</tr>
<tr>
<td>TL</td>
<td>Telomere Length</td>
</tr>
<tr>
<td>TRF</td>
<td>Terminal Restriction Fragment</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
1.1 Epigenetics

1.1.1 The Code of Life

Life, as we know it, depends critically on the multiplication of self-replicating organic entities. With the exception of protists, eukaryotes are multicellular organisms in which individual cells take on specialised roles to the benefit of the organism as a whole. Each somatic cell is a small semi-autonomous unit which, although highly specialised in its functionality, shares the same strands of deoxyribonucleic acid (DNA) at its core. Individual organisms carry small variations in the DNA code of their species, but by and large they are readily distinguishable as members of their species. The morphological and structural differences between the different cell types within an organism are far more complex. Throughout the development of an organism all cell types differentiate from a single zygote with one unique genome. The timed activation and deactivation of genes at different stages of development is crucial to the cellular differentiation process [1].
1.1.1.1 Genes

The abstract information encoded in the genome is usually discretised in terms of genes. The definition of a gene is a subject of continuous debate as there are many instances of positional and/or functional overlap between genes. Two commonly cited definitions are "A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions." [2] or "A union of genomic sequences encoding a coherent set of potentially overlapping functional products" [3]. Genes encoded in DNA are a relatively stable, heritable form of genetic information.

1.1.1.2 Central Dogma of Molecular Biology

DNA, and thus genes, consists of a long string of nucleotides which are linked by a sugar and phosphate backbone. These nucleotides are adenine (A), cytosine (C), guanine (G) and thymine (T). DNA in the cell core is present in a condensed form called chromatin. This molecule is particularly stable because two strands are present which are complementary to each other. Adenosine forms non-covalent bonds with thymine and cytosine with guanine.

The central dogma of molecular biology represents the information flow in a cell [4]. In order for genes to exert an effect in the cell they need to be translated to ribonucleic acid (RNA). To generate RNA the chromatin needs to be decompressed so that the translation complex can access it. The messenger RNA (mRNA) is subsequently translated to proteins in the ribosomes and the proteins fulfil their function inside or outside the cell. It is now generally understood that the central dogma is not entirely accurate as genes may manifest themselves in other forms than proteins (see Section 1.1.4).

1.1.1.3 Epigenetics

As debated as the definition of a gene is, the precise definition of epigenetics is possibly even more contested. Despite a substantial increase in publications on the subject (see Figure 1.1), multiple definitions are used to date. Much of the debate stems from the hereditary nature of certain traits [5].
For the purpose of this dissertation the working definition of the NIH Epigenetics Roadmap project is adopted: "Epigenetics refers to both heritable changes in gene activity and expression (in the progeny of cells or of individuals) and also stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable. While epigenetics refers to the study of single genes or sets of genes, epigenomics refers to more global analyses of epigenetic changes across the entire genome." [6]

The following sections will highlight different biological mechanisms of regulation that are considered epigenetic traits.

1.1.2 Histone Modifications

To achieve the compacted state which allows DNA to be stored within the nucleus, double stranded DNA (dsDNA) is wrapped twice around an octamer of histone proteins. This structure, called a nucleosome, stores approximately 147 bp of DNA. The individual nucleosomes are interspersed with "linker DNA". The histone proteins contain extensions (histone tails), which are the target of post-translational modifications. These modifications influence the degree of compaction of the DNA into lower
and higher order structures and therefore influence transcription through accessibility and the recruitment of transcription factors. The concept of a "histone code" was introduced which states that a specific combination of modifications is required to obtain a particular effect on gene expression [7, 8]. Transcriptionally active loci for instance, are featured by histone lysine acetylation [9].

Next to histone modifications, nucleosome positioning (or phasing) is also involved in the epigenetic regulation of gene expression by chromatin remodeling. Positioning of the nucleosomes depends on several components and their interactions. As a result, the core enhancer, promoter, and terminator regions of genes are typically depleted of nucleosomes, whereas most of the genomic DNA is occupied [10].

1.1.3 DNA Methylation

Cytosine bases can be covalently modified by the addition of a methyl group to the C5 carbon residue (5mC) by DNA methyltransferases [11]. In adult human beings the majority of methylated cytosines occur in a CpG dinucleotide context (note that this context is symmetrical with regard to the opposite strand) [12]. DNA methylation in the promoter region of a gene is typically associated with transcriptional silencing. The function of gene body methylation is less straightforward but important biological functions, such as regulation of splicing and silencing of transposable elements, have been assigned to methylation in these loci [13, 14].

Cytosine may also be modified with a hydroxymethyl group. Hydroxymethylcytosine (5hmC) has recently garnered attention as a novel epigenetic marker. It is hypothesised that 5hmC is an intermediary in active or passive 5mC demethylation. However, hydroxymethylation is most likely more than an intermediary state. One potential role lies in brain development where it specifically localises to gene bodies of genes involved in neuronal differentiation [15]. Recent findings also suggest that hydroxymethylation is critical for the regulation of a range of previously unidentified erythroid transcription factors, which likely have an affinity for 5hmC specifically. Hydroxymethylation thus plays a critical role in lineage commitment of hematopoietic pluripotent cells [16].
1.1.4 Non-coding RNA

In the central dogma of molecular biology RNA was considered an intermediary carrier of information between DNA and proteins. Although it was recognised that information encoded in RNA may be transferred to the DNA under specific conditions, it has since been discovered that RNA exerts effects on the cells function through a host of mechanisms other than strictly coding for proteins [17]. Although a full review of different RNA species is beyond the scope of this introduction, a short selection relevant to the interpretation of the experimental results will be highlighted.

1.1.4.1 microRNA

microRNAs (miRNAs) are short RNA fragments about 22 bp in length, which are complementary to specific mRNAs and can thus interfere with normal transcription of said mRNA. In animals most miRNAs are imprecisely complementary to their mRNA targets and they inhibit protein synthesis through mRNA degradation or preventing mRNA from being translated. miR16 for example, contains a sequence complementary to the AU-rich element found in the 3’ untranslated region of many unstable mRNAs, such as TNF-α or GM-CSF [18]. In case of complete complementarity between the miRNA and the target mRNA sequence, AGO2 can cleave the mRNA and lead to direct mRNA degradation. If complementarity is imperfect, the silencing is achieved by preventing translation [19].

Most miRNA genes seem to be solitary, and are expressed under the control of their own promoters and regulatory sequences. Other miRNA genes are arranged in clusters, and may be co-regulated with other members of the cluster. Vertebrate genomes contain about 250 miRNA genes, as shown by complementary DNA cloning and computational predictions [20].

1.1.4.2 Long Non-coding RNA

Long non-coding RNAs (lncRNAs) are somewhat arbitrarily defined as non-protein coding transcripts longer than 200 nucleotides [21]. This class of RNAs is estimated to have tens of thousands of members based on current research [22]. Although the
In this context, the majority of these have no known function, those that are well annotated show very diverse mechanisms of action.

For example, the inactivation of one X-chromosome in female mammals is directed by Xist, one of the earliest and best characterised examples of long ncRNAs. XIST expression is followed by irreversible layers of chromatin modifications. The expression of XIST from one X-chromosome leads to the inactivation of said X-chromosome. This process occurs in early embryonic stem cell differentiation [23]. In Drosophila lncRNAs induce the expression of the gene UBX by recruiting and directing the chromatin modifying functions of the trithorax protein ASH1 to HOX regulatory elements [24]. As a last example consider ZEB2, this gene’s mRNA has a long 5’ untranslated region (UTR) and requires the retention of a 5’UTR intron that contains an internal ribosome entry site for efficient translation. However, retention of the intron is dependent on the expression of an antisense transcript (lncRNA) that complements the intronic 5’ splice site [25].

1.1.5 Telomeres

1.1.5.1 Telomeres as an Epigenetic Feature

When cells divide through a process known as mitosis, the DNA in the cell’s core is duplicated and passes on to the new cell. This process is semi-conservative in the sense that each daughter cell receives dsDNA which is made up of an ”old” strand and a complementary strand newly synthesised before the cell division. In theory both sets of dsDNA are identical and several proof-reading mechanisms exist to ensure a faithful copy. However, due to a phenomenon known as the end-replication problem, by which the very end of each chromosome can not be copied, both daughter cells have slightly shorter chromosomes (see Figure 1.2). To prevent loss of genes after multiple divisions, the chromosomes are capped with a repetitive buffer sequence which may shorten throughout an organism’s lifespan without loosing

Figure 1.2: Illustration of the end-replication problem [26].
vital genetic information. These buffers present on both ends of each chromosome are known as the telomeres. In humans and other vertebrates the telomeric repeat sequence is TTAGGG. For additional protection the telomere ends are wrapped up in a protein complex known as shelterin [27].

Although telomere biology developed as a field unto its own, it can be considered as an epigenetic trait as defined in section 1.1.1.3. Telomeres after all do not display changes in the DNA code but do alter cell functions depending on the repeat length. Further it has been demonstrated that telomere length (TL) changes in response to external and internal stimuli and depends on parental age (and therefore TL) at birth.

In human subjects TL is often measured in whole blood samples where estimates typically vary between 4 and 11 kbp [28]. Due to the repetitive nature of the telomeric sequence, it is difficult to estimate the exact length with currently available techniques. Additionally, the estimates of these techniques do not always closely agree [29].

1.1.5.2 Terminal Restriction Fragments

The oldest method for measuring TL was first described in 1990 [30]. This method is quite robust and reproducible. To date, it is still considered the gold standard to which all alternatives should be compared.

In brief, the genomic DNA is treated with two endonucleases with a short recognition site (for instance rsa1 and hinf1). Most of the DNA will be broken up into relatively short pieces except for the telomeres which do not match either recognition site. In the next step, the terminal restriction fragments (TRF) are loaded on an agarose gel to separate them by length through electrophoresis. When migration is complete, the gel is transferred to a membrane by Southern blotting. Subsequently the DNA is allowed to hybridise with radioactively labelled telomere specific probes. The radiation allows the TRF length distribution to be visualised (see Figure 1.3 for experimental configuration and resulting blot). With the help of length markers the average telomere length can be determined as follows:

\[
\sum \frac{OD_i}{L_i} \sum \frac{OD_i}{L_i} (1.1)
\]
In formula 1.1 $L_i$ represents the length of DNA and $OD_i$ the measured optical density at position $i$. This formula is applied to a certain range of $i$, for instance from 3 to 17 kbp.

Below 3 kbp the quantification is not reliable. Other limitations of the technique are the relatively large amounts of DNA required ($\geq 2 \mu g$), the specialised equipment needed for handling radioactive materials and the fact that it cannot distinguish telomeres from subtelomeric regions (regions with a similar sequence to telomeres that do not possess sufficient restriction sites) [31].

### 1.1.5.3 qPCR

The quantitative polymerase chain reaction (qPCR), sometimes referred to as real time PCR, is a modified version of the original PCR which allows quantification by measuring fluorescent light emission as a proxy for the levels of synthesised dsDNA between reaction cycles [33]. In theory the amount of dsDNA doubles in every reaction cycle. In practice however the emitted light signal displays a logistic curve with an initial lag phase where the increase in light intensity is imperceptible and a plateau phase due to the exhaustion of reagents and/or the saturation of the light intensity.

To quantify the amount of DNA, a fluorescence threshold is selected and the cycle in which a given sample crosses this threshold is called the quantification cycle ($C_q$).
\[ C_q \text{ values are linearly related to the logarithm of the dsDNA concentration. By comparing } C_q \text{ values with those of a dilution series or some other known reference, the amount of DNA can be determined in absolute or relative terms. }

PCR primers that are perfectly complementary to the telomere sequence are also perfectly self-complementary and would therefore be unsuited for (q)PCR. Further, the primers can bind along the entire telomere which precludes accurate length estimations by gel electrophoresis. However, the amount of dsDNA synthesised is still representative of the initial amount of telomeric sequences. Using 3’ degenerate primers for the telomeres and a second set of primers for a known single copy gene, the amount of telomeric DNA can be estimated by normalising the telomere \( C_q \) to the single copy \( C_q \), this is the T/S ratio [31, 35].

The T/S ratio does not provide absolute TL measurements but does allow samples to be ranked by relative telomere length and many samples can be processed in parallel using relatively small quantities of DNA. Reasonably good correlations with TRF measurements have been obtained [29].

1.1.5.4 Other TL Measurement Techniques

Other techniques to measure TL have been used each with their own advantages and limitations. Fluorescent in situ hybridisation (FISH) can be performed both on immobilised cells [36] and liquid suspended cells (through flow cytometry) [37]. In both cases the amount of fluorescent signal from telomere specific probes is used as a proxy for TL.

It is also possible to measure the TL of individual chromosomes by a technique known as single telomere length analysis (STELA). This technique relies on amplifying full length telomeres with primers designed for the subtelomeric regions specific to the respective chromosomes [38].

Lastly, it has been proposed that fairly accurate TL estimates can be obtained as an unforeseen by-product of genomic sequencing data [39]. Although this technique is still biased by the PCR steps involved in second generation sequencing (see Section 1.3.2), further refinement in sequencing techniques could make this a highly accurate and cost-efficient technique.
Figure 1.4: Simplified diagram of the human circulatory system in anterior view. Reproduced from WikiMedia under the public domain licence.
1.2 Cardiovascular Health

1.2.1 Basic Anatomy

The circulatory system, also known as the cardiovascular system, pervades all other organs of the human body. It is responsible for the circulation of nutrients, metabolites, signalling molecules, oxygen and many other components throughout the body. It consist of the heart which pumps blood into the arteries, the arterial system which distributes blood throughout the body and the venous system which guides blood back to the heart (see Figure 1.4 for a schematic representation). Blood cells, among which leukocytes, are replenished by the bone marrow.

A normally developed human heart consists of four chambers: the left atrium, left ventricle, right atrium and right ventricle. The left atrium collects blood from the superior vena cava and inferior vena cava. From the right atrium blood passes to the right ventricle from where it is pumped into the pulmonary artery to continue on to the lungs to exchange CO$_2$ for O$_2$. Oxygen rich blood then returns to the left atrium through the pulmonary veins. Finally blood is passed to the left ventricle which pumps blood into the aorta to be circulated in the rest of the body [40].

The chambers of the heart are separated from each other and the arterial system by a series of valves which impose unidirectional blood flow (see Figure 1.5). The valves between the atria and ventricles are called atrioventricular valves. The semilunar valves sit at the openings that lead to the pulmonary trunk and aorta. These latter valves are termed the pulmonary and aortic valve respectively. The valve located between the right atrium and the right ventricle is the right atrioventricular valve, or tricuspid valve. In the opening between the left atrium and left ventricle sits the mitral valve, also called the bicuspid valve or the left atrioventricular valve.

1.2.2 Filling Patterns

The human heart beats in a two phase cycle, these phases are called systole and diastole. Systole is the phase of contraction and diastole the phase of relaxation. These phase can be defined both from the ventricular or atrial perspective. At the start of ventricular systole both ventricles are filled with blood. Through a powerful contraction of smooth muscle cells in the ventricle walls, blood is pumped into the aorta
and the pulmonary trunk. By the end of this contraction roughly 60% of the blood volume is forced out of the heart, this proportion is known as the ejection fraction (EF), a measure of systolic function [41]. After the contraction the ventricle walls start relaxing but without a change in ventricle volume, this period is the isovolumic relaxation time (IVRT).

The start of the IVRT also marks the end of ventricular systole and the start of ventricular diastole. During ventricular systole and IVRT the atria have been filled with blood from the veins, during ventricular diastole blood flows from the atria to the ventricles. At first the ventricles spontaneously draw in blood from the atria by returning to their fully relaxed state, this is early filling. Shortly thereafter the atria actively pump blood into the ventricles to complete loading, this is late filling [40].

The ratio of early over late filling is an indicator of diastolic function. A healthy heart draws mostly on early, spontaneous filling whereas an aged heart is less compliant and requires more active filling. Multiple measures are used to quantify diastolic function. One of the best known is E/A or the ratio of peak blood flow velocity during early (E) and late (A) filling as measured by Doppler echography. Another measure
is e which represents the movement speed of the cusp of the mitral valve during early (e) filling as measured by speckle tracking echography [42].

1.2.3 Atherosclerosis

A healthy vascular system is characterised by smooth arteries, elastic or muscular depending on the location in the body. Arteries consist of a trilaminar structure: tunica intima, tunica media and tunica externa. The endothelial cells in contact with the blood in the arterial lumen, rest upon a basement membrane [40].

Atherogenesis is a slow process that starts as early as birth. Insults to the arterial walls occur primarily at sites of non-laminar or turbulent blood flow (e.g. at branching sites [43]) due to classic risk factors (such as blood lipids, hypertension, diabetes, smoking, reactive oxygen species ... [44]), eventually leading to vulnerable sites. These vulnerable sites accumulate lipids and attract inflammatory cells. This leads to the formation of a lipid-rich core, causing the artery to enlarge outward (i.e. blood flow continues mostly unimpeded). This first stage is called fatty streak and many streaks recede or remain stable throughout an individual’s lifespan. Some fatty streaks however, are marked by elevated cytokine signalling which causes an influx of monocytes. Monocytes differentiate into macrophages which attempt to ingest the lipids but often fail to process them and become bloated foam cells, signalling for even more leukocytes to migrate to the lipid core. Unless blood lipid levels (particularly LDL cholesterol) and/or inflammatory signalling are remediated, this triggers a vicious cycle transforming the fatty streak into an atheroma [45].

In the center of an atheroma, foam cells and extracellular lipid droplets form a core region, which is surrounded by a fibrous cap of smooth-muscle cells and a collagen-rich matrix. In a growing lipid core proteinases secreted by the activated leukocytes can degrade the extracellular matrix while pro-inflammatory cytokines such as interferon-\(\gamma\) (IFN-\(\gamma\)) can limit the synthesis of new collagen. These changes can thin the fibrous cap and render it friable and susceptible to rupture.

When the plaque ruptures, blood that comes in contact with the tissue factor in the plaque coagulates, instigating thrombus formation. A thrombus can settle locally or in a remote artery causing restricted or blocked blood flow to the downstream organ. If the thrombus occludes the arteries which feed blood to the heart muscle or brain this can pose an acute threat to the individual’s life known as acute myocardial infarction or stroke respectively.
Figure 1.6: Gradual development of endothelial dysfunction into a ruptured atheroma.
Source: Wikimedia Commons.

A non-occlusive thrombus may eventually resorb as a result of endogenous or therapeutic thrombolysis. However, a wound healing response triggered by thrombin generated during blood coagulation can stimulate smooth muscle proliferation in the ruptured atheroma. Platelet-derived growth factor (PDGF) released from activated platelets stimulates smooth muscle cell migration. Transforming growth factor–β (TGF–β), also released from activated platelets, stimulates interstitial collagen production. This increased migration, proliferation and extracellular matrix synthesis by smooth muscle cells, thickens the fibrous cap and causes further expansion of the intima, often now in an inward direction, yielding constriction of the arterial lumen. Stenotic lesions produced by the lumenal encroachment of the fibrous plaque in arteries may restrict flow. Particularly under situations of increased cardiac demand this may lead to ischaemia, commonly provoking symptoms such as angina pectoris if the coronary arteries are affected. Advanced stenotic plaques, being more fibrous, may prove less susceptible to rupture and renewed thrombosis [46].
1.2.4 Epigenetics of Cardiovascular Function, Ageing and Disease

Epigenetic signalling has been firmly implicated in several diseases [47]. In colon cancer for instance, several studies have reported that transcriptional silencing of the gene *MGMT* is associated with hypermethylation of the promotor region [48, 49]. Yet the function of epigenetic signatures in cardiovascular disease is still largely unexplored. The study of epigenetic markers is gaining importance as it may aid in a deeper understanding of molecular mechanisms underlying the modulation of gene expression in the pathways linked to atherosclerosis, angiogenesis, ischemia-reperfusion damage, and the cardiovascular response to hypoxia and shear stress, among many others [50, 51].

1.2.4.1 Telomere Length

One of the first *in vivo* demonstrations of the association between systemic TL and CVD was a study showing shorter telomeres in peripheral blood leukocytes of subjects with severe coronary artery disease [52]. Later, similar associations were found in subjects with myocardial infarction [53, 54], chronic heart failure [55], stroke [53] and degenerative aortic valve stenosis [56].

There is also evidence that replicative senescence (and hence TL) of vascular endothelial cells is an integral part of atherosclerosis development and progression [57, 58]. In the tunica intima, an age dependent telomere attrition was observed, which was higher in arterial regions characterised by higher haemodynamic stress and increased cell turnover (the same regions that are prone to atherosclerotic lesion formation) [59, 60]. Systemic telomere length, as measured in the peripheral blood leukocytes, is also associated with cardiovascular health parameters such as pulse pressure [61], oxidised low density lipoprotein (ox-LDL) [62] and filling patterns [63].

1.2.4.2 DNA Methylation

Hiltunen et al. first reported that atherosclerotic lesions are characterised by global hypomethylation of DNA [64]. Using LINE-1 repeat element methylation as a proxy for global methylation, revealed lower methylation levels in patients affected by
CVD. At follow up, individuals with lower LINE-1 methylation had an increased risk of ischemic heart disease and stroke [65]. Despite advances such as these it has not been proven conclusively whether global hypomethylation is a consequence or cause of the proliferation of vascular cells in atherogenesis.

Besides global methylation levels, there are also numerous methylation differences in specific genomic locations which have been linked to CVD. There are numerous reviews providing a good overview of loci associated with atherosclerosis [66], coronary artery disease, hypertension, ... [50]. Additionally, a very recent study explored the methylation profile of vascular lesions in a semi-genome-wide manner [67].

Similar to TL involvement in CVD, it appears that DNA methylation is not only associated with strong outcomes such as the ones listed above. Changes of DNA methylation induced by nothing more than the patterns of blood flow have been shown both in vitro and in vivo [68].

1.2.4.3 Other Epigenetic Traits

Other epigenetic traits linked to cardiovascular function include, among others, histone modifications in cardiac growth [69], ncRNAs in atherosclerosis [70], miRNAs controlling cardiovascular differentiation [71] and miRNAs implicated in hypertension [72].

1.3 High-throughput Analysis

The research presented in this thesis relies heavily on the acquisition and processing of large quantities of data. This data tracks changes in specific parameters chosen to be monitored. These parameters however are only a representation of an underlying biological phenomenon. In order to draw meaningful conclusions from large amounts of data it is crucial to a) use appropriate statistics, b) utilise efficient algorithms and c) have a good understanding of how the data is related to the biological entity of interest. Each of these aspects will be addressed in this section.
1.3.1 Statistical Considerations

Statistics has long been applied to biology and medicine. In turn, specific requirements for biological research have driven further development of new statistical methods. This link was personified by Sir Ronald Fisher (1890–1962), both an accomplished statistician and geneticist he is perhaps best known as a proponent of the frequentist school of statistics. The statistical considerations highlighted in the following sections are exemplified by inter group difference but they are just as relevant to other types of associations such as regression analysis.

1.3.1.1 Traditional Framework

The traditional framework for analysing biological research centres around experiments with a limited number of parameters to be measured and ample replicates to estimate the contribution of the respective parameters on an outcome of interest. This type of data can be formalised as a matrix of dimensions $m \times n$ where $m \leq n$ and $m$, the number of rows, represents the number of variables measured and $n$, the number of columns, describes the sample size.

In the most simple of cases $m = 2$, i.e. one factor splitting $n$ into two groups and one variable of interest. In this case well known statistical tests for group comparisons such as the Student’s or Welch’s t-test may be applied if their specific conditions are met. For $m \geq 3$ methods such as MANOVA and multivariable linear models may be applied. Due to measurement errors the observed effect is likely different from the true effect. Increased $n$ will increase confidence in the estimates of coefficients by reducing the uncertainty of our observations.

1.3.1.2 Framework for High-throughput Analysis

Similar to the traditional framework, high throughput analyses can be described as a $m \times n$ matrix. However, due to technological advances we are now able to measure many different parameters in parallel leading to the situation where $m >> n$.

In the traditional framework, one test is performed at a given significance level $\alpha$ (usually 5% is chosen). In this case there is only a 5% chance of rejecting the null-hypothesis if the null-hypothesis is true. Multiple testing arises when an experimenter
sets out to compare two groups with respect to more than one outcome measure. Consider for instance a list of 100 gene expression values. If in reality all null-hypotheses are true (expression is identical in all of the 100 genes), the expected number of incorrect rejections is 5. If the tests are independent, the probability of at least one incorrect rejection is $99.4\% = 1 - (1 - 0.05)^{100}$. These errors are called false positives or Type I errors.

Several methods have been developed to keep the number of false positives in check for high-throughput analyses. One way of achieving this is to control the family-wise error rate (FWER). The FWER may be formulated as the probability of making at least one Type I error, or the probability of making at least one false positive decision. One example is the Bonferroni correction which simply sets a new significance threshold by dividing the cut-off $\alpha$ by the number of tests performed [73]. This is however a very conservative correction and in practice it is more useful to retain a few false positives rather than missing results due to false negatives.

One technique addressing this concern is frequently used throughout this thesis: the Benjamini-Hochberg false discovery rate (FDR) [74]. The concept behind FDR is to control the number of false positives within those variables which are declared statistically significant. Similar to the 5% threshold for Type I errors used in the traditional frequentist approach, we wish to limit the number of false positives to a threshold value $\alpha$. This procedure is formalised as follows:

$$P_k \leq \frac{k}{m} \alpha \quad (1.2)$$

Assume an experiment where $m$ null-hypotheses were tested and a list of $m$ P-values (P) in ascending order was obtained. For each $k$ from $m$ to 1 we check whether the inequality in 1.2 holds true. All $k$ null-hypotheses starting from the first one where this is the case, are rejected and the difference is considered significant. The Benjamini-Hochberg procedure assumes independence of all $m$ hypotheses but also works in several cases of dependence [75]. More importantly it is scalable with $m$ (even for millions of hypotheses) and does not depend on the true number of null-hypotheses that should be rejected. For example, assume 1000 features are compared between two groups and 20 (100) measures are returned significant by our statistical test. If the FDR is set at 5%, we expect on average 1 (5) false positive(s) in a list of 20(100) significant features. Note that in practice the extent to which truly differential features are recovered, depends on the magnitude of the difference, experimental noise and sample size.
1.3.2 Bioinformatics for Sequencing Data

Even the most simple tasks involved in analysing biological data require significant amounts of time if performed manually (the reader is invited to align a 25 bp sequence to a random 1000 bp reference or calculate the FDRs for 100 paired t-test in a two group comparison). In practice the human genome is approximately $3 \times 10^9$ bp in length, the number of sequences to be aligned are often expressed in millions, a gene expression comparison covers thousand of actively transcribed genes and large population studies typically involve hundreds if not thousands of subjects. These kinds of problems are not tractable by hand or even with basic computational techniques.

Like most scientific disciplines there is no comprehensive definition of what constitutes bioinformatics. In broad terms bioinformatics is a discipline that uses computer science, mathematical, statistical and engineering principles to investigate biological systems. Many sub-fields of bioinformatics exist but it is beyond the scope of this introduction to list all of them [76]. Most relevant to the subsequent chapters is the rise of sequence data analysis as a major discipline within bioinformatics.

1.3.3 Sequencing Data

The ability to sequence DNA fragments constituted a major breakthrough for biological research, the importance of which was acknowledged with the Nobel prize in Chemistry in 1980. The Sanger method of sequencing more specifically, would dominate sequencing efforts throughout the 1980’s, 1990’s and early 2000’s. Although the Sanger method and its subsequent improvements revolutionised sequencing, it is always limited in scope. Every reaction can determine only a single purified sequence.

For applications like expression profiling and metagenomics a complex mix of different (c)DNA species must be examined. Although hybridisation probe arrays offer one possible solution to this problem, they do not offer the same combination of flexibility and granular detail.

The first next generation sequencers, the Roche 454 and Illumina Genome Analyst, enabled the parallel sequencing of an arbitrarily complex mix of (c)DNA albeit with their own specific trade-offs. 454-sequencing offered $\sim 500000$ reads per
run with a maximum length of approximately 500 bp [77]. The first generation of Genome Analysers sacrificed read length (20 to 55 bases initially) for greatly increased throughput (2 to 10 million fragments per lane with 8 lanes on a plate) [78]. Both platforms have known multiple iterations since their introduction improving both on hardware and software allowing for faster sequencing with more accuracy and throughput. The research presented in this thesis only uses data from Illumina type sequencing.

Illumina sequencing (see Figure 1.7) relies on photosensitive terminators that allow each base to be read by a photo detector upon incorporation, one known base at a time. To produce a sufficiently strong light signal, each fragment to be sequenced is hybridised to a "spot" on the plate using universal adapters that were ligated to the fragment of interest. These adapters allow for the amplification of the target fragment by bridge amplification. In this process each fragment is amplified locally by PCR but they remain contained to the respective spots because the primers are attached to the plate. Once the bridge amplification is complete, all complementary strands are washed away and each spot carries multiple single stranded DNA copies of the original fragment. In theory all copies incorporate the same nucleotide in each sequencing cycle and their combined light signal is strong enough to detect.

Figure 1.7: Overview of the Illumina sequencing process. Adapted from [79].
The light signals are then converted to a base call which determines the identity of a base in a given position on a given fragment. The call is made with a certain certitude of correctness which is reported along with the base’s identity in a flat text file format known as "fastq" (see Figure 1.8). A single read from an Illumina machine consists of four lines: 1) a unique identifier line which usually contains the machine id and the position on the plate, 2) a line with each individual base call, 3) a line beginning with a "+" character optionally followed with more information and 4) a line of single characters which encode the base call confidence of each base and is therefore of equal length as the base call line [80]. In paired-end mode, two fastq files are generated, the first containing reads from one end of each fragment and the second from the other end. The reads in both files appear in the same order, a requirement for many applications.

Figure 1.8: An example of 6 reads in the typical 4-line fastq format.

There are many ways to process fastq files depending on the application. In human experiments typical steps include: quality control, trimming adaptors at the 5’ end and low quality bases at the 3’ end, mapping reads to a reference genome or
transcriptome and finally summarising the results. For each of these steps several tools are available depending on the application [81–83], whether they be published open-source software, licensed software or self-scripted solutions.

In this thesis three main types of sequencing experiments were performed: 1) RNA sequencing, 2) genome wide methylation enrichment sequencing and 3) targeted bisulphite sequencing. The general workflow and specific data properties of each will be highlighted below.

1.3.4 RNA Sequencing

The general purpose of RNA sequencing is to obtain a snapshot of the transcription and splicing activity of cells. Several techniques have been devised to extract RNA from cells and enrich the fragments of interest.

1.3.4.1 RNA Preparation

The first hurdle is the overwhelming abundance of ribosomal RNA (rRNA). rRNA is an integral component of the ribosomes that constitute the translation machinery. Estimates vary between cell types but the proportion of rRNA is usually in excess of 80% [84]. Although there are good applications for rRNA profiling, usually in the realms of evolutionary [85] and microbial community research [86], for many other applications this represents a situation where less than 20% of the sequences is actually relevant to the research question.

Several techniques exist to enrich the non-ribosomal fraction of RNA: capturing based on the polyadenylated tail of mRNA [87], rRNA removal by probes which rely on conservation of the ribosomal sequences [88] or removal of the most abundant cDNA species (obtained from RNA) by duplex-specific nuclease [89]. Platform specific RNA sequencing protocols are usually available and the sequencing is relatively straightforward as the RNA is converted to cDNA before sequencing. These protocols vary by input material (RNA input amount, fresh frozen, formalin-fixed and paraffin-embedded, ...) and output (paired vs single end, directional, coverage, ...).
1.3.4.2 Normalisation

Once sequencing is complete and data pre-processing has been handled, the sequences need to be mapped with a splicing-aware alignment algorithm like Bowtie2 [90] or STAR [91]. The mapping results are then summarised by genomic features. These features are usually genes, transcripts or exons. After converting mapped sequences to feature coverage, the first step of analysis is typically normalisation. Normalisation is an essential step in sequence data analysis. If we are looking for expression differences between two samples but the number of total fragments sequenced for sample 2 is twice the number for the first sample, it would appear that all genes have roughly double the expression of the first sample even if the samples are identical.

Several methods have been used for normalisation. The simplest is the introduction of uniform correction factors per sample to adjust for the mean coverage [92]. More intricate correction procedures include accounting for feature length and total coverage (Fragments Per Kilobase of exon per Million fragments mapped: FPKM) [87], log transformation of the data after correcting for feature length and application of micro-array analysis techniques [93] or adjusting for the coverage distribution of each sample when calculating scaling factors [94].

An example of this last class of methods is known as trimmed mean of M-values (TMM) normalisation. The underlying assumption is that even if a sample is characterised by extreme expression values in a few genes, the majority of genes will have ”moderate” expression values. Because sequencing experiments have a limited number of ”slots” (i.e. reads) available, a highly expressed gene will not only produce a lot of reads, it will also suppress other genes ability to be picked up in reads. TMM normalisation counteracts this by first calculating log transformed ratio’s of expression (a.k.a. fold change). Subsequently it discards an arbitrarily chosen percentage of genes with the highest expression or highest absolute fold changes (30% by default) and calculates scaling factors based on two assumptions. The first being that the majority of genes most likely have little or no change in expression and secondly that more highly expressed genes are more stable and should be attributed additional weight in normalisation.
1.3.4.3 Comparing Expression Data

After normalisation the sequence data can be compared between groups. Summed sequence reads as a proxy for gene expression values are prone to substantial variation whether it be due to technical or biological variation. This behaviour can be modelled quite well by a negative binomial (NB) distribution [95]. But the facts that many genes are tested at once and that the number of replicates is usually limited by the experimental cost, makes test moderation, i.e. avoiding unrealistic small variance estimates, crucial in order to avoid spurious results [96]. One way of achieving this was implemented in the R package "edgeR" [97]. In this approach, three sources of information regarding the dispersion of reads are combined: an estimate of dispersion across all genes in a sample, an estimate of variance in function of the expression level and lastly an estimate of dispersion per gene (across samples). Each of these dispersions needs to be estimated. Robinson et al. initially proposed a relatively simple quantile adjusted conditional maximum likelihood method [95] but have since expanded the estimation methods to better describe the nature of RNA-seq data and allow more complex experimental designs. Regardless, the end result is a dispersion parameter $\varphi$ which can be plugged into a NB distribution describing the reads. Finally, in the case of a two-group comparison, an exact P-value can be calculated analogues to the Fisher exact test but replacing the hypergeometric distribution by a NB distribution for P-value calculation. For more complex designs generalised linear models are used with slight modifications to estimate the coefficients [98].

A substantial range of other tools is available, each with their own approach to handling variance estimation, significance testing, multiple testing correction and reporting [99–101]. These analysis steps are also required for sequencing data obtained from other sources than RNA but one should keep in mind that some applications, such as methylation enrichment sequencing, have very different read distribution properties.

1.3.5 Methylation Enrichment Sequencing

To query the methylation status of cytosine residues over the entire length of the genome a relatively simple and cost effective technique was developed. Instead of inspecting the entire genome, the relevant sections can be isolated by affinity capturing of $5m$CpG [102]. When the enrichment step is followed by sequencing this method is
known by several names: MethylCap-seq [103], MBD-seq [104], and MBD-isolated genomic sequencing (MIGS) [105]. For clarity the term MethylCap-seq will be used throughout this thesis.

The conceptual basis for MethylCap-seq is depicted in Figure 1.9. In the first step, genomic DNA is fragmented usually by sonication but alternatives such as nebulisation [106] are also suitable. In the second step, sonicated DNA fragments are captured by naturally occurring or specifically modified Methyl Binding Domain (MBD) containing proteins such as MBD2 or MCEP2 which are affixed to beads or elution columns. Under low salt concentrations these proteins will preferentially bind methylated CpGs in a non-covalent way. The last step is to elute the bound DNA with increasing salt concentrations. Low salt concentrations will release fragments that are slightly methylated while high concentrations will also elute fragments with multiple methylated CpGs.

![Figure 1.9: Schematic outline of the MethylCap procedure. Genomic DNA is first fragmented. Glutathione S-Transferase—MBD fusion protein coupled to magnetic beads is then used in an (automated) capture process. Methylated DNA fragments from the genomic DNA pool is bound to the immobilised GST—MBD under low salt conditions. A salt gradient is subsequently used to wash and elute DNA fragments from the immobilised GST—MBD. FT: Flow Through [103].](image-url)
the reference genome. This results in coverage maps that reflect the methylation status in the sample. It should be noted that MethylCap-seq actually yields no information about the methylation degree of any specific CpG dinucleotide and the resolution is limited to the fragment sizes. Most studies use paired-end sequencing. But even though contemporary Illumina type sequencers easily obtain reads 100 bp long, which means a 200 bp fragment is fully covered in a paired-end run, the sequencer cannot distinguish between unmethylated CpG and $5m$CpG (or $5hm$CpG for that matter).

To analyse this type of data it is assumed that the methylation degree in a given region is proportionate to the sequencing coverage. It has been demonstrated that CpG density and CG-content are sources of bias which can make certain parts of the genome unnameable to analysis by MethylCap-seq [107, 108]. In practice sequence coverage is not necessarily comparable between regions in the same sample but fairly similar to the same region in other samples. Lastly, coverage of a region can be expressed in terms of the number or reads within a region but can also be quantified as the maximum coverage in a region. For the analyses in this thesis we rely on maximum coverage as this measure is independent of the length of the region.

As far as data analysis is concerned, methods developed for differential expression detection in RNA-seq are compatible with MethylCap-seq data as long as one accounts for the fact that the coverage distribution and number of variables are quite different. Exploratory analyses have shown that TMM for instance, often does not offer any improvements over simple library size scaling (data not shown).

1.3.6 Bisulfite Conversion Based Methods

Although MethylCap-Seq offers a good trade-off between coverage and cost, it only provides a very rough estimate of the methylation degree in each region. The *de facto* standard for DNA methylation is bisulphite sequencing.

Bisulfite sequencing measures methylation by using a chemical conversion which it is named for. By treating DNA with (sodium) bisulphite, cytosine bases are converted to uracil (see Figure 1.10). Uracil has the same base pairing behaviour as thymine. If the converted DNA is amplified by polymerases the C-G base pair will be converted to T-A. However, when a cytosine is methylated it is protected from conversion and will remain a C-T base pair [109].
Figure 1.10: Bisulfite-mediated conversion of cytosin to uracil. Adapted from [110]

The bisulfite treated DNA can be analysed in several ways but whole genome sequencing offers unparalleled coverage and detail. This strategy is very costly as high coverage is needed to obtain a reliable methylation throughout most of the genome [111]. Additionally, CpG dinucleotides are considerably under-represented in the human genome [112] and thus much of the sequencing capacity is wasted. To increase coverage in the relevant sections of the genome one possible solution is reduced representation bisulphite sequencing (RRBS) [113]. This technique relies on restriction enzymes which are indifferent to methylation such as MspI which has a 3’-CCGG-5’ restriction site. By amplifying only the regions surrounding these breakpoints (typically 500 bp) only ~ 1% of the genome is sequenced. Other options exist but will not be discussed here [114].

A different way to use bisulfite treated DNA is to perform (q)PCR with two primer pairs, one for the case of unconverted, originally methylated, cytosine and another for the converted thymine (unmethylated cytosine) [115, 116]. These techniques, collectively referred to as methylation specific PCR (MSP), offer considerable cost reduction when dealing with large sample sizes but even though tools are available for primer design (such as MethPrimer [117]), in practice it can be difficult if not impossible to design good primers for a specific region.

An intermediate in this range of techniques is targeted bisulphite sequencing. In this approach primers are designed which target a predefined region of interest. These primers typically contain no or few CpG dinucleotides and are thus able to amplify the target regardless of methylation status. One or several regions per sample may be amplified and the amplicons are then sequenced, typically at very high coverage resulting in very precise methylation estimates [118]. Despite these advantages, the accuracy of these estimates is still not optimal because sequence differences as a result of the bisulphite conversion tend to affect the efficiency of the PCR amplification [119, 120].
The analysis of sequencing reads from bisulphite treated DNA is fairly straightforward. For each cytosine (or group of neighbouring cytosines) a methylation percentage is obtained. These percentages can simply be summarised or compared between treatment groups. One analysis package implements both logistic regression by modelling log odds of the treatment groups (i.e. logarithm of the percentage methylated divided by the percentage unmethylated) while accounting for the coverage depth and, a Fisher’s exact test in case no replicates are present [121].

All of these methods share a set of caveats that should always be considered when interpreting results [111]. Getting the reaction circumstances for bisulphite conversion right is a difficult process. If the exposure time is too long or the bisulphite concentration too high, a substantial amount of DNA will be destroyed by the caustic nature of the reaction. If the time is too short or insufficient bisulphite is used, the conversion of cytosine may be incomplete and yield false positive methylation results. Lastly, bisulphite converted DNA typically shows strand breaks which reduce the average fragment length considerably. Therefore primers should always be designed to amplify a relatively short amplicon (typically less than 500 bp). Lastly, some regions may no longer be mappable to the genome as the conversion of cytosine to thymine can greatly reduce the sequence complexity.

1.3.7 Limitations of Sequencing Data Analysis

As mentioned above, sequencing data is best modelled by a NB distribution characterised by a mean and a dispersion parameter. Although the above mentioned ”edgeR” package (and therefore also the ”methylKit” package which uses ”edgeR” code for the actual differential methylation analysis) was developed and optimised with NB distributions in mind, there are still unresolved questions about the behaviour of such data in more complex experimental design.

In chapters 3 and 4, we will introduce an experimental design which includes repeated measurements of the same subjects for both within and between subject comparisons. For within subject comparisons, power can be gained by using what are in effect repeated measurements. This can be taken into account by introducing subject identity as a random effect using mixed models or by incorporating it using fixed effects. The former has the advantage that both within and between subject effects can be assessed, but, imposes additional assumptions.
Random effects are assumed to be normally distributed and orthogonal on the fixed effects. The use of subject specific fixed effects on the other hand, requires fewer assumptions but these models are less suited for estimating effects of variables that do not change over time. For NB data, the use of subject specific effects is not well studied yet, and they might lead to liberal P-values [122]. These limitations will be discussed in more detail in sections 3.4 and 4.4.

1.3.8 Pathway Analyses

Sequencing experiments are often used to generated a ranked list of genes. This is not a true endpoint for biological research. More formally a ranked gene list can be seen as a basis on which to generate new research hypotheses. To derive biologically meaningful results, pathway analysis seeks to place observed gene expression or methylation differences in a functional context. For instance, if 10% of the significantly differentially expressed genes are kinases, as opposed to 1% of the genes in the human genome which are kinases, this is a strong indication that kinases in general are relevant to the research questions (without putting too much emphasis on any particular kinase coding gene). Pathway analysis methods fall in three broad categories: 1) singular enrichment analysis (SEA), 2) gene set enrichment analysis (GSEA) and 3) modular enrichment analysis (MEA) [123].

1.3.8.1 Pathway Repositories

Regardless which pathway analysis method one chooses, a reference database of known pathways and their constituent genes is required. Several repositories are available each varying in focus, content curation or accessibility such as KEGG, Reactome, WikiPathways ... [124].

1.3.8.2 Singular Enrichment Analysis

The most traditional strategy for enrichment analysis is to take the users preselected genes of interest (e.g. differentially expressed genes between experimental versus control samples) and then iteratively test the enrichment of each annotation term one-by-one in a linear mode. Thereafter, the individual, enriched annotation terms
passing the enrichment P-value threshold are reported in a tabular format ordered by the enrichment probability (enrichment P-value). The enrichment P-value calculation, i.e. number of genes in the list that hit a given biology class as compared to pure random chance.

A common weakness of tools in this class is that the linear output of terms can be overwhelming (from hundreds to thousands). Therefore, the interrelatedness of relevant terms can be diluted. For example, relevant GO terms like apoptosis, programmed cell death, induction of apoptosis etc. may be spread out across this long list. In addition, the quality of pre-selected gene lists could largely impact the enrichment analysis, which makes singular enrichment analysis (SEA) analyses unstable to a certain degree when using different statistical methods or cut-off thresholds.

1.3.8.3 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) uses the core principles of SEA, but with a distinct algorithm to calculate enrichment P-values as compared to SEA [125]. The core concept of GSEA is the elimination of the need for a cut-off. Besides reducing the need for arbitrary factors in the gene selection step that could impact SEA analyses, this strategy offers the added advantage of deriving information from all genes to a varying extent by using a weighing factor. The maximum enrichment score (MES) is calculated from the rank order of all gene members in the annotation category. Thereafter, enrichment P-values can be calculated by comparing the MES to a MES distribution obtained by randomly shuffling the samples (i.e. permutation).

However, tools in the GSEA class are also characterised by certain limitations. First, the GSEA method requires a summarised biological value (e.g. fold change) for each of the genes as input. Sometimes, it is a difficult task to summarise many biological aspects of a gene into one meaningful value when, for instance, one considers different splice forms of a gene. Additionally, the underlying assumption is made that genes with large differences in expression (e.g. fold changes), which drive the enrichment P-values, are contributing more to the biology. In practice, small changes in signal transduction genes can result in larger downstream biological consequences. Depending on the questions that the researcher is asking, the mildly changed signal transduction genes may be more relevant than genes with large expression changes. One last concern is that GSEA is underpowered in cases where the number of replicates is low because this limits the number of potential permutations.
1.3.8.4 Modular Enrichment Analysis

Modular enrichment analysis (MEA) inherits the basic enrichment calculation found in SEA and incorporates extra network discovery algorithms by considering term-term relationships. MEA tools organise and condense a wide range of heterogeneous annotation content, such as GO terms, protein domains, pathways and so on, into term or gene classes. These methods take into account the redundant and networked nature of biological annotation content in order to concentrate on building the larger biological picture rather than focusing on an individual term or gene. Such data-mining logic seems closer to the nature of biology in that a biological process works in a network manner.

However, the limitation of MEA is that “orphan” terms or genes (without strong relationships to neighbouring terms/genes) could be left out from the analysis. Thus, it is important to examine those terms or genes that are left out during MEA analyses [126]. In addition, the quality of the pre-selected gene list impacts the analytic results, just as it does in SEA analysis.

1.3.8.5 Other Considerations

The classifications listed so far offer a useful insight into the general nature of pathway analysis tools. However, there are other ways in which tools within a category may vary and certain properties may be shared between categories [127].

The first property is the definition of the null hypothesis. In pathway analysis there are two ways in which the null hypothesis is defined. The first states that the genes in the gene set (pathway) of interest are no more enriched in the significantly differentially expressed genes than they are in the rest of the gene list. The second states that no single gene in the gene set is differentially expressed. Pathway analysis tools based on these hypotheses are called competitive or self-contained respectively. The interpretation of the results differ for both and they have different properties as self-contained tests are typically more easily rejected.

A second property relates to the calculation of P-values. Most current methods use gene permutation to determine whether genes are enriched in the significant set i.e. how likely is it to obtain a comparable enrichment when all gene rankings are
shuffled at random and the same number of genes at the top of the list are considered significant. This approach operates under the assumptions that there is no correlation between genes in the ranking. In practice, this is often not the case which can lead to a substantial overestimation of the number of significant pathways. An alternative method is to shuffle the samples several times and determine the ranking for each permutation in order to estimate the probability of a certain enrichment. This latter method preserves any correlation that may be present in the gene expression levels and thus arrives at much more conservative estimates.
1.4 Study Rationale

1.4.1 The Asklepios Study on Successful Cardiovascular Ageing

The Asklepios Study is a longitudinal population study focusing on the interplay between ageing, cardiovascular haemodynamics and inflammation in (preclinical) cardiovascular diseases [128]. The 2524 participants (1301 women) are a representative cohort of 35-55-year-old individuals, free from overt cardiovascular disease at study initiation, randomly sampled from the twinned Belgian communities of Erpe-Mere and Nieuwerkerken.

Baseline examinations conducted in 2002-2004 include: questionnaires, conventional risk factors and biochemistry. Additional phenotypes under study include a) vascular structure and function and b) cardiac structure and function. A novel aspect of the study is "integrated" non-invasive biomechanical assessment of cardiac, arterial and ventriculovascular function. This integrated haemodynamic phenotype is being tested in detection, prediction and prevention of clinical cardiovascular pathology (atherosclerosis progression, atherothrombosis, development of heart failure etc.). A second aspect is the systematic evaluation of peripheral blood leukocyte TL as a marker for biological ageing.

Previous studies of the Asklepios population have revealed, among other findings, several aspects of telomere dynamics in cardiovascular ageing. In this population women had longer telomeres than men, inflammation and oxidative stress markers were correlated with TL but classical CVD risk factors were not [129]. In addition preclinical atherosclerosis was also not correlated with TL [130]. Lastly, paternal age at birth was an independent predictor of adult TL [131].

Concurrently, the Asklepios study has also resulted in approximately 80 non telomere related publications ranging from dietary studies [132], over biometric normalisation strategies [133] to a novel risk stratification strategy [134].

Currently the first follow-up round (2012-present) is being conducted after an average follow-up of 10 years. Baseline examinations are repeated and the incidence of (cardiovascular) events is recorded. This follow-up round also introduced the collection of whole blood RNA and the investigators intend to include a small cohort of new participants to study intergenerational differences.
1.4.2 Asklepios and Epigenetics

Drawing on various elements introduced in this chapter, the remainder of this thesis seeks to expand our current understanding of cardiovascular ageing. More specifically high-throughput data analysis will be applied to link epigenetic signals to other parameters of cardiovascular health.

Chapter 2 introduces the Asklepios study population which provided the patient data and samples at the core of this thesis. After a summary of the current understanding of how telomere length and cardiovascular health parameters are related, this chapter focusses on the associations between telomere length on one hand and a series of functional parameters such as left ventricular filling and vascular stiffness on the other.

The study of DNA methylation changes is highlighted in Chapter 3. A small pilot group of subjects was selected, half of which remained in good cardiovascular health, the other half showed a marked increase in atherosclerosis. Initial findings were validated in a second group of subjects and novel potential markers of atherosclerosis were identified.

In Chapter 4 the methylation data are revisited but this time with a particular focus on the epigenetic changes that occur in the course of ageing. Through the additional power conferred by a paired study design, this study found distinct differences between ageing as a linear function of age (cross-sectional analysis) and ageing at the individual level (longitudinal analysis).
REFERENCES


2.1 Introduction

The progression of acquired cardiovascular diseases (CVD) throughout the human life can typically be tracked by a series of gradual changes in physical, chemical and biological parameters. Levels of systolic blood pressure (SBP), cholesterol, C-reactive protein (CRP), smoking status and sex have all been linked with an increased likelihood of adverse cardiovascular events [1–3]. Telomere length (TL), although not used in clinical practice, is one such parameter that has repeatedly been linked with cardiovascular health and disease development [4].

Telomeres are the nucleotide-protein complexes that shield the chromosomal ends from erosion caused by the end-replication problem during cell division and distinguishes them from double-stranded breaks to prevent chromosomal fusion [5]. Throughout the replicative lifespan of cells, their TL will decrease until a critical threshold is reached. Critically short telomeres will typically lead to a cell crisis resulting in senescence, apoptosis or immortalisation [6]. TL is of particular interest because it potentially provides a cumulative measurement of stresses throughout life representing "biological age" [4].
Although there is still uncertainty about the mechanism(s) by which telomere biology and CVD pathogenesis affect each other, results from both molecular biology and epidemiology have repeatedly shown significant associations [7–11]. The same is true for cardiovascular risk factors such as insulin resistance, hypertension [12], smoking status [13], oxidative stress and inflammation [14].

Systemic TL has also been linked to left ventricular (LV) structure and function but mostly in smaller, patient-specific settings and not in a general population [15–22]. Shorter TL can be found in heart failure (HF) patients [23, 24] and patients suffering from chronic HF have an increased morbidity if their telomeres are shorter [25]. However, reports on the association between TL and indicators of diastolic dysfunction show conflicting results [18, 19]. Similarly, a positive correlation has been described between LVM and PBL TL [20–22], but other studies did not detect a significant association between TL and LVM index or LV hypertrophy [18, 26, 27].

The population-based Asklepios Study offers the advantages of a large sample size and the measurement of numerous potential confounders of TL and CVD. We therefore investigated the relations between systemic TL and proven prognostic parameters [28–32] of vascular stiffness, cardiac stiffness, systolic function, diastolic function and ventricular mass, to shed light on the baseline state of these correlations.

### 2.2 Methods

#### 2.2.1 Study Population

All data presented in this paper were collected during the first round of the Asklepios study on successful (cardiovascular) ageing. The study comprises 2524 subjects approximately 35 to 55 years of age, free from overt cardiovascular disease or other significant pathologies at baseline. The full description of the study design, inclusion criteria, detailed methodology and population baseline characteristics have been published previously [33]. The study was conducted in concordance with the principles of the Declaration of Helsinki. All subjects gave written informed consent and the study was approved by the Ghent University Ethical Committee. For the analyses reported here, we used the subset of 2509 subjects for which reliable TL and all major TL confounder measurements (age, sex, paternal age at birth) were available (cf. De Meyer et al. [34]).
2.2.2 Biochemical Analyses

All subjects were fasting, had refrained from smoking for at least 6 hours and were screened for active infection/inflammation before blood sampling. Conventional serum parameters were measured using commercial reagents according to the manufacturer’s recommendations on a Modular P automated system (Roche Diagnostics, Mannheim, Germany), in an ISO 9002 certified reference laboratory [33]. Coefficient of variation of all tests was < 3.0%. These parameters included Interleukin-6 (IL-6), C-reactive protein (CRP), oxidised low-density lipoprotein (ox-LDL), serum uric acid concentrations and brain natriuretic peptide precursor [33].

2.2.3 Telomere Length

For TL-analyses, whole blood was collected in EDTA tubes cooled to 4°C. DNA isolation was performed within 3 days of collection using the Puregene Genomic Purification Kit (Gentra Systems, Minnesota, USA). The DNA was long-term stored at -80°C before TL measurement in duplicate. 5 μg was digested with 5U RsalI and 10U HinfI followed by gel electrophoresis, Southern blotting, radioactive hybridisation of the telomeric fragments and weight markers, phospho-imaging and quantification (expressed as kbp: kilo base pairs) [14].

2.2.4 Echocardiographic and Vascular Examination

Blood pressure was recorded using bilateral triplicate measurements (1 minute intervals) on a rested, sitting subject using a validated oscillometric Omron HEM device (Omron Healthcare Co. Ltd., Kyoto, Japan). Blood pressure values of these six readings were averaged and the mean value of systolic blood pressure (SBP) is used throughout this study. The subjects underwent a resting echocardiographic examination and a scan of the left and right carotid and femoral arteries (VIVID 7, GE Vingmed Ultrasound, Horten, Norway). Left ventricular (LV) internal dimensions were measured at end-diastole (LVEDD) with the area-length method. Sphericity was defined as LV width divided by LV length and is expressed as a percentage. Standard 2-D volumetric methods were used to calculate ejection fraction (EF) from end-diastolic and end-systolic LV volumes and to calculate LV mass (LVM). The LVM was scaled allometrically following the recommendations of Chirinos et al. [35]
as $LVM/(Height)^{1.7}$ to account for the effects of both obesity and blood pressure on LVM. We also scaled LVM to the body surface area ($g/m^2$) [36].

Other cardiac and arterial measurements included the following: systolic ($s'$), and early ($e'$) and late ($a'$) diastolic septal mitral annulus pulsed wave tissue Doppler (TDI) velocities, pulsed wave Doppler early (E) and late (A) diastolic transmitral flow velocities, E-wave propagation velocity (Vpe) and carotid-femoral pulse wave velocity (PWV). PWV was calculated as follows:

$$PWV = \frac{\Delta L_{S-F} - \Delta L_{S-C}}{\Delta T_{Q-F} - \Delta T_{Q-C}}$$ (2.1)

In formula 2.1 $\Delta L_{S-F}$ and $\Delta L_{S-C}$ are the distances measured from sternal notch to femoral and carotid measuring sites respectively, $\Delta T_{Q-F}$ and $\Delta T_{Q-C}$ are the time delays between the start of the QRS complex and the onset of systolic flow in the femoral and carotid artery measured by pulse wave Doppler imaging (full methodology described in the online supplements of Rietzschel et al. [33]). CW Doppler recordings were used to measure isovolumic relaxation time (IVRT) as the interval from the closure spike of the aortic valve to onset of mitral flow.

### 2.2.5 Data Analysis

Statistical analyses were performed in R 2.15.2. Continuous variables are reported as the mean value with standard deviation. Means of groups were compared with Student (homoscedasticity) or Welch t-test (heteroscedasticity) as appropriate. To evaluate the contribution of the different confounders to the response variables under study we applied general linear models as implemented in the glm’ function. We report both P-values and the estimated unstandardised effect sizes (b) for TL in these models.

### 2.3 Results

Baseline characteristics of the population are presented in Table 2.1.
Table 2.1: Baseline characteristics of the Asklepios study population.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Women (n = 1291)</th>
<th>Men (n = 1218)</th>
<th>P-value^a</th>
<th>Population (n = 2509)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>45.9 ± 6.0</td>
<td>46.1 ± 5.9</td>
<td>0.316</td>
<td>46.0 ± 6.0</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>66.7 ± 12.7</td>
<td>82.0 ± 12.4</td>
<td>&lt;2.2E-16</td>
<td>74.1 ± 14.7</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>163 ± 6</td>
<td>176 ± 7</td>
<td>&lt;2.2E-16^b</td>
<td>169 ± 9</td>
</tr>
<tr>
<td>Body Mass Index (kg/m(^2))</td>
<td>25.1 ± 4.6</td>
<td>26.5 ± 3.7</td>
<td>&lt;2.2E-16^b</td>
<td>25.8 ± 4.3</td>
</tr>
<tr>
<td>Systolic Blood Pressure (mmHg)</td>
<td>123 ± 14</td>
<td>131 ± 13</td>
<td>&lt;2.2E-16^b</td>
<td>127 ± 14</td>
</tr>
<tr>
<td>Pulse Pressure (mmHg)</td>
<td>45.5 ± 9.1</td>
<td>48.3 ± 7.4</td>
<td>&lt;2.2E-16^b</td>
<td>46.9 ± 8.4</td>
</tr>
<tr>
<td>Pulse Wave Velocity (m/s)</td>
<td>6.60 ± 1.45</td>
<td>6.65 ± 1.46</td>
<td>0.397</td>
<td>6.62 ± 1.45</td>
</tr>
<tr>
<td>Heart Rate (min(^{-1}))</td>
<td>67.2 ± 9.5</td>
<td>64.0 ± 10.7</td>
<td>6.01E-15^b</td>
<td>65.6 ± 10.2</td>
</tr>
<tr>
<td>Used Antihypertensive Drugs</td>
<td>145 (11.2%)</td>
<td>118 (9.69%)</td>
<td>0.284^c</td>
<td>263 (10.5%)</td>
</tr>
<tr>
<td>PBL TL (kbp)</td>
<td>7.96 ± 0.73</td>
<td>7.78 ± 0.71</td>
<td>3.26E-9</td>
<td>7.87 ± 0.73</td>
</tr>
<tr>
<td>E (cm/s)</td>
<td>78.9 ± 14.2</td>
<td>70.6 ± 13.0</td>
<td>&lt;2.2E-16^b</td>
<td>74.9 ± 14.2</td>
</tr>
<tr>
<td>A (cm/s)</td>
<td>63.5 ± 11.9</td>
<td>59.6 ± 10.9</td>
<td>&lt;2.2E-16^b</td>
<td>61.6 ± 11.6</td>
</tr>
<tr>
<td>e' (cm/s)</td>
<td>9.41 ± 2.13</td>
<td>8.63 ± 1.83</td>
<td>&lt;2.2E-16^b</td>
<td>9.03 ± 2.03</td>
</tr>
<tr>
<td>a' (cm/s)</td>
<td>8.67 ± 1.55</td>
<td>9.23 ± 1.46</td>
<td>&lt;2.2E-16^b</td>
<td>8.94 ± 1.53</td>
</tr>
<tr>
<td>E/A</td>
<td>1.29 ± 0.32</td>
<td>1.22 ± 0.29</td>
<td>2.65E-7^b</td>
<td>1.25 ± 0.31</td>
</tr>
<tr>
<td>c'/a'</td>
<td>1.13 ± 0.36</td>
<td>0.970 ± 0.29</td>
<td>&lt;2.2E-16^b</td>
<td>1.05 ± 0.34</td>
</tr>
<tr>
<td>E/e'</td>
<td>8.68 ± 1.99</td>
<td>8.41 ± 1.78</td>
<td>3.36E-4^b</td>
<td>8.55 ± 1.90</td>
</tr>
<tr>
<td>s' (cm/s)</td>
<td>7.91 ± 1.13</td>
<td>7.93 ± 1.20</td>
<td>0.756</td>
<td>7.92 ± 1.16</td>
</tr>
<tr>
<td>Vpe (m/s)</td>
<td>79.4 ± 22.1</td>
<td>71.9 ± 19.5</td>
<td>&lt;2.2E-16\d</td>
<td>75.7 ± 21.2</td>
</tr>
<tr>
<td>DT (ms)</td>
<td>167 ± 29</td>
<td>170 ± 30</td>
<td>9.29E-4</td>
<td>168 ± 30</td>
</tr>
<tr>
<td>Isovolumic Relaxation Time (ms)</td>
<td>81.8 ± 13.8</td>
<td>90.2 ± 12.9</td>
<td>&lt;2.2E-16</td>
<td>85.9 ± 14.0</td>
</tr>
<tr>
<td>LV Ejection Fraction (%)</td>
<td>64.8 ± 6.7</td>
<td>62.5 ± 6.4</td>
<td>&lt;2.2E-16</td>
<td>63.7 ± 6.7</td>
</tr>
<tr>
<td>LVEDD (mm)</td>
<td>44.9 ± 3.9</td>
<td>49.4 ± 4.4</td>
<td>&lt;2.2E-16^b</td>
<td>47.1 ± 4.7</td>
</tr>
<tr>
<td>Sphericity (%)</td>
<td>56.9 ± 6.4</td>
<td>57.2 ± 6.5</td>
<td>0.246</td>
<td>57.0 ± 6.4</td>
</tr>
<tr>
<td>Left Ventricular Mass (g)</td>
<td>124 ± 31</td>
<td>179 ± 40</td>
<td>&lt;2.2E-16^b</td>
<td>151 ± 45</td>
</tr>
<tr>
<td>LVM index (g/m(^2)(^1/7))</td>
<td>54.2 ± 13</td>
<td>69.0 ± 15.0</td>
<td>&lt;2.2E-16^b</td>
<td>61.3 ± 15.9</td>
</tr>
<tr>
<td>NTproBNP (pg/ml)</td>
<td>81.3 ± 64.1</td>
<td>36.4 ± 39.2</td>
<td>&lt;2.2E-16^b</td>
<td>59.5 ± 58.0</td>
</tr>
</tbody>
</table>

\(\text{P-value}\): P-value for comparison between sexes using independent t-test (\^{b}unequal variance) or chi-square test (\^{c}).

\(\text{Data transformed}\): Data was log transformed before statistical testing.

\(\text{LVM index} \): allometrically scaled Left Ventricular Mass index, PBL TL: Periferal Blood Leukocyte Telomere Length, E & A: peak transmitial flow velocities during early (E) and late (A) diastolic filling, e' & a': peak movement speed of mitral annulus during early (e') and late (a') diastolic filling, DT: transmittal Deceleration Time, NTproBNP: N-terminal prohormone of brain natriuretic peptide, LVEDD: Left Ventricular End-Diastolic Diameter, s': peak systolic mitral annulus movement speed, Vpe: pulse propagation velocity in early diastole.
2.3.1 Main Correlates of Telomere Length

Previous studies of the Asklepios population showed that TL length correlates with age and gender. Mean TRF was shown to be inversely correlated with age ($P < 1E-25$). The yearly telomere attrition calculated from these cross-sectional baseline data was estimated at 26 base pairs (bp). Mean TRF was found to be dependent on gender ($P < 1E-8$) with men having on average 172 bp shorter telomeres than women. After age adjustment, this difference was 166 bp ($P < 1E-8$). The cross-sectional data showed that telomere attrition proceeded faster in men ($30.0 \text{ bp per year, } R^2 = 0.062$) compared to women ($20.3 \text{ bp per year, } R^2 = 0.028$) [14]. Further, paternal age showed a substantial positive linear correlation with TL ($R^2 = 0.127, P < 1E-29$) [34].

It has been proposed that socio-economic status (SES) captures a large fraction of TL variability [37]. Using higher education (college or university) as a proxy for SES, we indeed discovered a correlation with TL ($b = 0.06827, P = 0.024$) but it did not remain upon addition of the two main confounders age and gender ($b = 0.0462, P = 0.119$).

2.3.2 Diastolic Function

Unadjusted models yielded positive linear associations between TL and E/A, $e'/a'$ (Figure 2.1) and E/e' (Table 2.2, Model 1). In successive (general linear) models, known major confounders of diastolic function were added, i.e. age and sex in Model 2, additionally heart rate (HR), systolic blood pressure (SBP) including use of anti-hypertensive drugs and body mass index (BMI) in Model 3. We did not remove non-significant independent variables from the proposed models in Table 2.2 for the individual response variables as removal of the non-significant terms did not alter the significance of the TL component. Addition of further potential confounders: LV sphericity, oxidative stress (oxidised-LDL cholesterol, serum uric acid) or inflammatory markers (high-sensitive CRP, IL-6), did not significantly alter the TL - diastolic dysfunction relationships when added to Model 3 as an independent variable (data not shown).
In all models the positive association between E/A and TL remained significant (see Table 2.2, $P \leq 0.002$) with Model 3 accounting for approximately 43.2% of E/A variability (2.64% of the total variability could be attributed to TL). Examining the data separately by sex, we found that, upon adjustment for confounders (Model 3), the association between E/A and TL was significant in both women ($b = 0.030, P = 0.001$) and men ($b = 0.023, P = 0.014$). A similar approach was adopted for the $e'/a'$ ratio. Results of the linear models were included in Table 2.2 and demonstrated significance of the adjusted associations for $e'/a'$ (Model 3: $b = 0.018, P = 0.012$). Looking at both sexes separately, TL was a significant independent variable in women ($b = 0.022, P = 0.032$), but only borderline in men ($b = 0.016, P = 0.098$). Both for E/A and $e'/a'$ $b$ is larger in women and the P-value smaller. To formally test for this difference we added an interaction term for gender and TL length to Model 3. The interaction term proved not to be significant ($P = 0.293$ for E/A, $P = 0.136$ for $e'/a'$).

We further examined the components of these ratios (E/A and $e'/a'$) separately to determine whether the associations were attributable to one of both components or whether the ratios contained additional information beyond the terms they consist of. Surprisingly, the results indicated that neither E ($b = 0.226, P = 0.535$) nor $e'$

Figure 2.1: TL $\sim$ E/A. Scatterplot showing the unadjusted correlation between telomere length (TL) and the ratio of early (E) over late (A) mitral annulus movement speed for women (red) and men (blue) with the regression line for both combined (black).
(b = 0.043, P = 0.341) were correlated with TL (Model 3). There was however a correlation of TL with A (b = −0.905, P < 0.001) and a’ (b = −0.079, P = 0.040).

Accordingly, E/e’ was not correlated upon adequate adjustment (Model 3: b = 0.014, P = 0.775), neither were other indices of diastolic function: flow propagation velocity of the E-wave (Vpe; b = 0.212, P = 0.718), mitral inflow deceleration time (b = −0.302, P = 0.712) or duration of the atrial contraction (b = −0.028, P = 0.938). Only IVRT was inversely correlated with PBL TL (Figure 2.2, Model 3: b = −0.900, P = 0.011). IVRT was also inversely correlated with PBL TL after age, sex and paternal-age adjustment (PBL TL as dependent variable; b = −2.92E − 3, P = 0.010). Additionally, we found no significant association with brain natriuretic peptide (log(NT-proBNP): b = −0.005, P = 0.617) upon adjustment.

2.3.3 Systolic Function and LV Structure

We could not document significant associations between TL and systolic function after adequate adjustment (Model 3, EF: b = 0.016, P = 0.933 and s’: b = −0.038, P = 0.238). Similarly no associations were found with LV structure assessed by LV end-diastolic diameter (LVEDD: b = −0.124, P = 0.261) and LV mass (b = −0.214, P = 0.801), body surface area adjusted LV mass (b = 0.010, P = 0.981) or allometrically height-adjusted LV mass (b = 0.023, P = 0.943).

2.3.4 Arterial Stiffness

No significant partial correlation was found between PWV and age-adjusted TL (cf. supra), a result that remained unchanged after additional adjustments (Model 3, b = −0.001, P = 0.971). The same is true for pulse pressure (PP) in the Asklepios population (Model 3, b = 255E − 6, P = 0.999).
### Table 2.2: The association of TL (kbp) with different parameters of cardiovascular function using general linear models.

<table>
<thead>
<tr>
<th>Response variable (RV)</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E/A</td>
<td>0.0696, 2.83E-16</td>
<td>0.0249, 1.25E-03</td>
<td>0.0264, 8.09E-05</td>
</tr>
<tr>
<td>P</td>
<td>&lt;2.2E-16</td>
<td>0.0178</td>
<td>0.0182</td>
</tr>
<tr>
<td>E/(cm/s)</td>
<td>1.91, 1.09E-06</td>
<td>0.600</td>
<td>0.535</td>
</tr>
<tr>
<td>P</td>
<td>-1.60, 5.37E-07</td>
<td>-9.05</td>
<td>4.51E-04</td>
</tr>
<tr>
<td>e’/(cm/s)</td>
<td>0.397, 8.93E-13</td>
<td>0.050, 0.304</td>
<td>0.431</td>
</tr>
<tr>
<td>P</td>
<td>-0.211, 5.77E-07</td>
<td>-0.068</td>
<td>0.0403</td>
</tr>
<tr>
<td>E/e’</td>
<td>-0.158, 2.62E-3</td>
<td>-0.0214</td>
<td>-0.0124</td>
</tr>
<tr>
<td>P</td>
<td>-1.86, 2.24E-02</td>
<td>-0.123</td>
<td>-0.302</td>
</tr>
<tr>
<td>IVRT (s)</td>
<td>-2.70, 2.32E-12</td>
<td>-0.941</td>
<td>-0.900</td>
</tr>
<tr>
<td>log(NTproBNP (pg/ml))</td>
<td>0.017, 0.119</td>
<td>-0.004</td>
<td>-0.005</td>
</tr>
<tr>
<td>LVM index (g/m$^{1.7}$)</td>
<td>-2.37, 5.41E-08</td>
<td>-0.194</td>
<td>0.0228</td>
</tr>
<tr>
<td>LVEDD (mm)</td>
<td>-0.497, 1.11E-04</td>
<td>-0.164</td>
<td>-0.124</td>
</tr>
<tr>
<td>EF (%)</td>
<td>0.0216, 0.906</td>
<td>-0.0183</td>
<td>0.0156</td>
</tr>
<tr>
<td>s’ (cm/s)</td>
<td>0.018, 0.579</td>
<td>-0.033</td>
<td>-0.038</td>
</tr>
</tbody>
</table>

Model 1: RV $\sim$ TL
Model 2: RV $\sim$ TL + Age + Sex
Model 3: Model 2 + Systolic BP + Heart Rate + BMI + Used Antihypertensive Drugs

E & A: peak transmitral flow velocities during early (E) and late (A) diastolic filling, e’ & a’: peak movement speed of mitral annulus during early (e’) and late (a’) diastolic filling, DT: transmitral Deceleration Time, IVRT: Isovolumic Relaxation Time, NTproBNP: N-terminal prohormone of brain natriuretic peptide, LVM: allometrically scaled Left Ventricular Mass index, LVEDD: Left Ventricular End-Diastolic Diameter, EF: Ejected Fraction of end-diastolic volume, s’: peak systolic mitral annulus movement speed

b: effect size (unstandardised), P: P-value of TL component
2.4 Discussion

Our main finding is that we can extend a number of previously described associations between PBL TL and cardiovascular structure and function - usually detected in smaller, diseased cohorts - towards a middle-aged, apparently healthy population. After adjustment for confounders, we could not document an association with systolic function, cardiac structure or vascular stiffness. We do however document an intriguing association with certain parameters of LV filling.

2.4.1 LV Filling and Diastolic Function

As mentioned in the introduction, previous reports on the association between TL and indicators of diastolic dysfunction have led to apparently conflicting results. Quartiles of telomere length were shown to correlate with diastolic dysfunction in CAD patients (evaluated by E/A and pulmonary vein flow) [18]. In contrast both E/A-ratio and diastolic dysfunction were not correlated with PBL TL in a study of elderly subjects (> 85 years) [19]. This can potentially be explained by the very different nature of the populations under study. Our data showed a significant association between PBL TL and both the E/A-ratio and the e’/a’-ratio and with IVRT [38], but not with the E/e’-ratio (see Table 2.2 for details). E/e’ is mainly used as an indicator of elevated filling pressures, reflecting more advanced diastolic dysfunction not likely to be present in healthy subjects [39]. Concurrently, NT-proBNP, a biochemical marker reflecting elevated filling pressure and diastolic dysfunction [40], was not associated with PBL TL after correction for confounders (Model 3).

Analysis of the terms constituting the E/A and e’/a’ ratios in this population suggests that TL was associated with atrial contraction (A and a’), rather than with parameters that could reflect myocardial relaxation or filling pressure, such as E and e’, or reflect myocardial stiffness, such as shortened mitral deceleration time [41]. As previous publications only described correlations with E/A but not with the individual components, we cannot tell whether this was the case in other study populations.

Relaxation and stiffness induce opposite effects on mitral E and A and hence on E/A [41]. The present data do not provide sufficient evidence for an independent association between PBL TL and LV relaxation or LV stiffness in the general middle-aged population. Although there is an association with longer IVRT after
correction for heart rate and blood pressure, it is difficult to attribute this to myocardial relaxation without a persisting association with the best validated determinants e’ and Vpe [42]. The persisting associations with atrial contraction flow velocity (A) and with the simultaneously occurring annular velocity (a’) then most likely affects compound measures such as E/A and e’/a’. One could speculate that increased IVRT and enhanced atrial contraction represent an early and subtle delay of myocardial relaxation, which is not yet apparent in other measurements. We therefore would describe the findings as correlation between PBL TL and altered filling pattern without sufficient evidence for diastolic dysfunction.

2.4.2 LV Systolic Function, Structure and Vascular Stiffness

In this population without overt cardiac disease, we report that PBL TL was not associated with minor changes in systolic parameters such as EF and tissue Doppler movement speed of the septal mitral annulus (s’). Increased LV mass correlates with increased all-cause and cardiovascular mortality [43] and a positive correlation has been described between LVM and PBL TL [20–22]. However, our findings do not support an association between PBL TL and LVM, body surface area scaled LVM or height-scaled LVM in the context of a relatively young population. Two other studies also failed to detect an association between TL and LVM index or LV hypertrophy in two older populations (∼ 65 years) [18, 26]. These findings and the fact that the former studies do not agree on whether normotensive or hypertensive patients show a TL - LVM correlation, lead us to the conclusion that there is likely an as of yet unidentified confounder at work. A third study found no cross-sectional or longitudinal associations between PBL LTL and cardiac measurements including LVM (adjusted for body surface area) [27].

With respect to vascular stiffness, no significant associations between PWV (or PP) and TL were found in either sex. Previous studies have reported this association to be significant in men [15, 44]. The discordance may be attributable to age and health characteristics of the respective populations. Indeed these populations were featured by a higher mean age (∼ 10 years) and a higher mean PWV (60% higher) compared to the Asklepios study albeit with a different measurement protocol for PWV.
2.4.3 Socio-economic Status

Socio-economic status (SES) has been correlated with TL and SES itself is significantly correlated with variables that reflect poorer health status [37]. Limited access to health services, poor dietary options, increased stress etc. can affect TL. Having a higher education (college or university) as a proxy for SES was correlated with TL in the Asklepios population but not so after correction for the confounders of age and gender. Addition of higher education to our model did not affect the correlations with E/A or e’/a’ (data not shown).

2.4.4 Limitations

The Asklepios data set does not yet include any longitudinal information thus limiting it to all the drawbacks associated with cross-sectional study designs. Particularly claims of causality can not be made with cross-sectional data alone. Despite the extensive characterisation of the Asklepios study, there are some descriptors of CV function which were not measured (e.g. pulmonary venous flow). We cannot exclude the involvement of these factors.

2.4.5 Mechanistic Insights

There are two general models in which TL is tied to cardiovascular health. The first states that telomere shortening is a primary driver of (cardiovascular) ageing. In support of this model our findings indicate that the associations between some parameters of LV filling and telomere length, are not limited to (chronic) HF patients, but may already be present in a young to middle-aged, apparently healthy population and are more pronounced in women. It is tempting to speculate that telomere biology could be mechanistically involved in the early pathogenesis of diastolic dysfunction and possibly HF by extension. As a matter of fact, in mice, knock-out of the telomere elongating enzyme telomerase, resulted in shortened telomere length over several generations which was associated with the development of overt chronic HF [45]. However, telomere biology in mice cannot be easily transposed to humans and additional experiments would be absolutely necessary to pinpoint the exact mechanisms.
In our data, TL was clearly correlated with E/A, e'/a’ and IVRT but not with other indices of diastolic function. We further explored the relationship between PBL TL and other cardiac and hemodynamic parameters such as sphericity of the ventricle, Vpe, duration of the A-wave and deceleration time (data not shown). No significant associations were found that could help provide clues as to the mechanism by which PBL TL and diastolic function might be linked.

The second model states that TL is merely an epiphenomenon, an indicator influenced by conditions in the ageing body. Accelerated telomere attrition in subjects with mildly impaired diastolic function could also be caused by oxidative stress and inflammation, two factors that are known to affect diastolic function as well as telomere length [14, 23, 46–48]. In this case it might be expected that the addition of oxidative stress and inflammation would cause reduced significance for TL. However, additional markers (CRP, oxidised LDL, IL-6 and serum uric acid) did not substantially alter the significance of the PBL TL component relative to Model 3 in Table 2.2 (data not shown). It should be noted though that these markers only reflect point measurements of oxidative stress and inflammation, which are variable by nature, whereas telomere length has been hypothesised to reflect their cumulated effects (reviewed in De Meyer et al. [4]).

Our data, at present, are insufficient to determine the more likely model. Some of the non-replicated associations might still become apparent with ageing, assuming that a certain threshold of telomere attrition needs to be reached before it has a measurable effect on cardiovascular stiffness or visa versa.

2.5 Conclusion

Our results show that several parameters of cardiovascular structure and function which have been associated with TL, fail to replicate in a well-phenotyped middle-
aged population sample. However, PBL TL is associated with subtle changes in certain parameters of LV filling in this population. Further investigation of the underlying biological mechanisms is warranted to provide insights into the relationship between PBL TL, diastolic function and cardiovascular health.

### 2.6 Citation

This chapter is based on an article accepted in PLoS ONE for publication on the 15th of December 2014.

**On cross-sectional associations of leukocyte telomere length with cardiac systolic, diastolic and vascular function. The Asklepios Study**


*: These authors contributed equally to this work.
References


Peripheral Blood Leukocyte DNA Methylome Reflects Atherosclerosis Progression

3.1 Introduction

The dynamics of atherosclerosis, the gradual thickening of vascular endothelial walls by accumulation of saturated fatty cells, are influenced by a host of factors [1]. Some are considered intrinsic, others are deemed environmental influences and interactions between them are abundant [2]. Several genomic loci are associated with an increased risk or severity of atherosclerotic plaque formation [3]. Since an individual’s genetic code is set at conception, it is considered a non-modifiable risk factor. However, genetic factors explain only a limited portion of atherosclerotic risk [4]. Some of this missing variability might be attributable to epigenetics.

Epigenetics is commonly defined as the continuum of heritable alterations that affect gene expression without being reflected in the genetic code itself [5]. Epigenetic effectors include histone modifications, DNA modifications,
non-coding RNAs and telomeres. Epigenetic changes are induced by both internal and environmental cues and are generally reversible [6].

Recent advances in technologies for measuring epigenetic changes have led to a steady increase in research assessing the role of epigenetics in the pathogenesis of atherosclerosis [7]. One marker particularly, 5-methylcytosine in genomic DNA, has garnered much attention. The majority of studies have established DNA methylation profiles in vasculature affected by atherosclerosis [8, 9]. Additionally, it was demonstrated in APOE-null mice that the appearance of vascular lesions was preceded by DNA methylation alterations in aortic tissue but also in peripheral blood mononuclear cells [10].

Indeed, atherosclerosis is considered an inflammatory disease [11] and our current understanding of atherosclerosis hinges on the notion that macrophages are recruited from the bloodstream to sites of increased endothelial stress [12]. Yet few studies address atherosclerosis related epigenetic changes in circulating leukocytes. Despite the fact that blood contains a complex and changing mix of cell types [13], the accessibility of peripheral blood makes it an interesting research subject as it may lead to straightforward clinical applications. In cancer for instance, blood DNA methylation markers are being evaluated for this purpose [14]. Previous research targeting blood elucidated atherosclerosis related changes in miRNA expression [15], methyl metabolism [16], CpG methylation levels in repeat elements [17, 18] and gene specific changes in CpG methylation [19–23]. This study aims to identify novel genomic locations in leukocytes, that exhibit DNA methylation changes during atherosclerosis progression through genome-wide screening.

To this end, we selected a subset of female participants from the Asklepios longitudinal study on successful ageing. Using data from both study rounds, the subjects were classified in two groups: progressors and non-progressors. At study initiation (first round) only volunteers without overt cardiovascular disease were included. We defined progressors as those subjects who developed more pronounced — but still subclinical — atherosclerosis in the decade between both study rounds. Non-progressors were those who remained stable in terms of the extent of atherosclerosis. Only females were evaluated to avoid sex-related epigenomic variation and significantly fewer members of the male subpopulation were available in the non-progressor arm, thus it was more difficult to match according to the stricter inclusion criteria imposed for this research.
3.2 Methods

3.2.1 Study Population Selection

All data presented in this paper, were collected from participants of the Asklepios study on successful (cardiovascular) ageing. This study comprises 2524 subjects approximately 35 to 55 years of age, free from overt cardiovascular disease or other significant pathologies at enrolment in 2002-2004. The participants returned for follow up examination in 2013-2015. The full description of the study design, inclusion criteria, detailed methodology and population baseline characteristics were published previously [24]. All subjects gave written informed consent, the study complies with the declaration of Helsinki and was approved by the Ghent University Ethical Committee. For the analyses reported here, we used two non-overlapping subsets of 20 and 46 subjects.

3.2.2 Biochemical Analyses

All subjects were fasting and were screened for active infection/inflammation before blood sampling. Conventional serum parameters were measured using commercial reagents according to the manufacturer’s recommendations on a Modular P automated system (Roche Diagnostics, Mannheim, Germany), in an ISO 9002 certified reference laboratory. Coefficient of variation of all tests was < 3.0%. These parameters included Interleukin-6 (IL-6), C-reactive protein (CRP), oxidised low-density lipoprotein (ox-LDL), total cholesterol, high-density lipoprotein (HDL) and low-density lipoprotein (LDL) concentrations among others.

3.2.3 Echocardiographic and Vascular Examination

Blood pressure was recorded using bilateral triplicate measurements (1 minute intervals) on a rested, sitting subject using a validated oscillometric Omron HEM device (Omron Healthcare Co. Ltd., Kyoto, Japan). Blood pressure values of these six readings were averaged and the mean value of systolic blood pressure (SBP) is used throughout this study. The subjects underwent a resting echocardiographic examination and a scan of the left and right carotid and femoral arteries (VIVID 7, GE Vingmed Ultrasound, Horten, Norway).
3.2.4 Experimental Design

Asklepios Study participants were classified in two categories. The first group consisted of those who showed good cardiovascular health in the second round, i.e. they did not develop significant atherosclerosis since their initial enrolment during the first round as evaluated by echography (from here on called "non-progressors"). The second group consisted of those that where healthy in the first round but developed substantial atherosclerosis since (from here on called "progressors"). The extent of atherosclerosis was assessed as the augmentation of plaque size and thickness near vascular branching sites. Other parameters of atherosclerosis progressions such as intima media thickness or oxLDL concentration were not considered.

To minimise the effect of confounding variables that might alter the epigenetic profile and to exclude participants with obvious indications of poor cardiovascular health, we retained only subjects that met these criteria. More specifically, we required that the subject was female, had a systolic blood pressure below 140 mmHg and less than 90 mmHg diastolic blood pressure in both study rounds, had a BMI smaller than 31 in both study rounds, did not use any statins prior to or during the study and was a self-reported non-smoker. We further balanced both groups as best as possible with respect to the following criteria: age, BMI, BMI change between study rounds and blood pressure.

3.2.5 Methylated CpG Enrichment Sequencing

3.2.5.1 Experimental Procedures

We used stored DNA samples collected for the Asklepios study obtained from whole blood samples. We selected 10 Progressors and 10 Non-Progressors and analysed their DNA from both the first and second round. To assess reproducibility we included a technical replicate in each round, parallel processing from sampling of stock DNA onward, for a total of 42 samples.

MethylCap-Seq was performed using the MethylCap kit (Diagenode, Liège, Belgium) according to the manufacturer’s instructions. We opted for high salt concentration elution. The enriched fragments were prepared as a paired-end library with the NEBNext Ultra DNA Library Prep (New England Biolabs,
Massachusetts, USA). Library quality was assessed by High Sensitivity DNA chip on a 2100 Bioanalyzer (Agilent Technologies, California, USA). The libraries were sequenced 2x50bp on a HiSeq 2000 (Illumina, California, USA). With the exception of two samples, we pooled 4 samples equimolarly per lane based on qPCR measured concentration (Kapa Biosystems, Massachusetts, USA) while ensuring that the technical replicate samples were placed in 4 different lanes. With exception of the technical replicates each lane (sample pool) contained both the first and second round samples of one progressor and one non-progressor. Sequencing was performed in 2 separate runs with sample pools randomly distributed between runs. Sequencing results for all samples are listed in Supplementary Table C.3.

The sequenced reads were mapped to the human genome using bowtie (v0.12.7) [25]. The bowtie parameters were set to 0 mismatches in the seed (first 28 nucleotides). Only unique paired reads were retained and both fragments were required to locate within 400bp of each other on the human reference genome (GRCh37/hg19). For each sample duplicate paired reads with the exact same location were discarded, as these are most likely amplified from the same DNA fragment. Quality control was performed with FastQC (v0.10.1). All samples were of high quality (Phred Score, duplication levels, elevated CpG content) and did not require resequencing according to these quality metrics or the CpG-enrichment plots (Supplementary Figure B.1) [26]. The mapped reads were summarised to the ’Map of the Human Methylome’ Build 3 (http://www.biobix.be/map-of-the-human-methylome/). This map contains 3618706 unique genomic locations that are susceptible to DNA CpG methylation called ’methylation cores’ (MCs). For each MC, the maximum coverage within the MC was determined per sample.

The resulting count table was used for differential methylation analyses in R (v3.0.1). First we performed additional QC by estimating the correlation coefficients between the technical replicates after normalising the counts by simple library size scaling based on the number of unique mapped pairs, adding 1 to each count and log transforming the data (Supplementary Figure B.2). Although the replicates show reasonably good correlation (round 1 $R = 0.862$, $P < 2.2E-16$, round 2 $R = 0.971$, $P < 2.2E-16$), the correlation is better for the second round suggesting modest deterioration of reproducibility due to storage at -80°C over time. When we performed unsupervised clustering on the 2000 MCs with the highest overall variance, the cluster dendrogram (Supplementary Figure B.3) shows a clean pairing of first and second round samples with the exception of one subject (AS1.07 and AS2.28). We
additionally looked for the presence of homozygous or heterozygous SNPs in all known SNPs with a coverage higher than 10. Pairing of samples by an independent, blinded investigator recovered the correct sample pairings (triplets for technical replicates). We therefore concluded that no samples were swapped and that MethylCap-Seq was successful.

Prior to analyses, we selected the technical replicates with the highest coverage (AS1.20 and AS2.47) reducing the number of samples to 40. The MCs were annotated based on their location relative to Ensembl features (release 72).

### 3.2.5.2 Analysis Strategy

In this chapter there were three research questions of interest. First, what MCs are differentially methylated between subjects with atherosclerosis progression and without in the second study round? Second, what MCs differentiated between progressors and non-progressors in the first study round and might therefore be predictive. Third, were there MCs that consistently differed between progressors and non-progressors irrespective of age (i.e. in both study rounds). We used the edgeR package (v3.4.2) to detect differential methylation [27].

In this study design, there was a potential interaction between progressor status and study round. We modelled the main effects and the interaction in a fixed effects model by eliminating the intercept term and a fixed effect with four levels: ”NonProgressor-Round1”, ”Progressor-Round1”, ”NonProgressor-Round2” and ”Progressor-Round2”. The contrasts corresponding to the three research questions are listed in Table 3.1.

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonProg-Round1</td>
<td>-1</td>
<td>0</td>
<td>-0.5</td>
</tr>
<tr>
<td>Prog-Round1</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>NonProg-Round2</td>
<td>0</td>
<td>-1</td>
<td>-0.5</td>
</tr>
<tr>
<td>Prog-Round2</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note that this model does not include a variable for the individual subjects. For the comparison in each round separately, this model is appropriate as there
are no repeated measurements of the individual and the addition of a fixed effect for subject identity would mask the atherosclerosis related effects. For the global comparison however, not taking the paired nature of the samples into account would ignore correlations between the repeated measurements of the same study subjects in both rounds.

For NB data, the use of fixed effect models that account for subject specific effects is not well studied yet. Fixed effect model can be expected to give unbiased point estimates for within subject comparisons in NB distributed data, but there might be issues with the estimation of the standard errors that are provided by standard software [28]. We opted not to include a subject specific effect. This means that our differential methylation analysis comparing all progressor samples to all non-progressor samples, could be overconfident because the correlation between first and second round samples of the same subject was not taken into account. The following strategy was implemented to evaluate the P-value estimates of the edgeR package.

We performed a single permutation in which the samples progressor status for half of the progressors and non-progressors were switched (in both rounds identically). This represented a divergent permutation that eliminated potential atherosclerosis progression effects while preserving the within subject correlation. An enrichment of low P-values in the permuted dataset would indicate that the P-values were too liberal and the inference on the real data has to be interpreted with care.

3.2.6 Targeted Bisulfite Deep Sequencing

Based on the results obtained in the MethylCap experiment we selected 16 locations to validate the epigenetic differences observed. We used an independent set of 23 non-progressors and 23 progressors selected according to the criteria described above. Sufficient round 1 DNA was available for all subjects. In round 2, DNA was unavailable for 6 subjects. For 1 subject no blood could be drawn. For the remaining 5 subjects, yield of the extraction was insufficient. We also included a methylated and unmethylated control in the form of in vitro methylated DNA (S7821, Millipore, Massachusetts, USA) and DNA from a methyltransferase double knock-out cell line (HCT116) for a total of 88 samples and one no-DNA blank (negative for all primers, not included in sequencing).
Bisulfite conversion was performed with the EZ DNA Methylation — Lightning kit (Zymo Research, California, USA) according to the manufacturer’s instructions. After bisulphite treatment we amplified the regions of interest with the primers listed in Supplementary Table C.4. Primers were designed taking into account that primers may only contain few if any CpG as this might cause the PCR reaction to fail. For the design we used an in-house tool based on Primer3 [29]. If this approach failed to find primers with acceptable thermal parameters we used MethPrimer [30]. For each subject - primer pair combination, digital electrophoresis (LabChip Gx, Perkin Elmer, Massachusetts, USA) was performed to confirm the presence of the desired product. Short DNA fragments, most likely primer dimers, were removed by High Pure PCR Purification Kit (Roche, Basel, Switzerland). The concentration of the remaining product was determined by PicoGreen measurement (Life Technologies, California, USA). The samples were prepared for sequencing with the Truseq DNA PCR-free Kit (Illumina) with the following alterations to the protocol. We had less than 1 μg of total amplicon DNA available per sample. To avoid an additional PCR step, the adapters were diluted ten-fold to reduce library preparation input requirements. The quality of the resulting library was assessed by High Sensitivity DNA chip on a 2100 Bioanalyzer (Agilent Technologies). Sequencing was performed on a single MiSeq lane (Illumina) with all samples loaded as an equimolar mix according to qPCR (Kapa Biosystems, performed on a Roche Lightcycler). All fragments were sequenced for 250 bases in a paired-end fashion. As some target regions were shorter than this length, they were in effect fully covered in duplicate but sequencing does not extend beyond the primer boundaries.

Data preprocessing, including base calling, demultiplexing, trimming and fastq generation was automatically performed in Illumina BaseSpace with standard settings as provided by the manufacturer (Illumina). Additionally we inspected the quality of the reads with FastQC. This inspection showed consistent read quality in both the forward and reverse read. The sequencing data was analysed using the Bismark package (v 0.10.0, available from http://www.bioinformatics.babraham.ac.uk/projects/bismark/) [31]. Default parameters were used for Bismark with the following exceptions: ”-q” to indicate the input was in fastq format, ”–bowtie2” to use Bowtie 2 as the underlying mapping software, ”–no-mixed” and ”–no-discordant” were specified to allow only fragments that map as proper pairs with both reads on the opposing strands and finally, the maximum allowed fragment size was raised to 550 bp with ”–maxins 550” to allow mapping of the longest amplicon.
Differential methylation in the BS-seq data was evaluated with the methylKit package [32]. First we filter out any CpG methylation calls from non-targeted regions and CpGs with low coverage (<100). Next, we used the package’s "calculateDiffMeth" function to calculate differential methylation with the options "weighted.mean = FALSE" and "slim = FALSE". These options respectively specify that we have equal confidence in methylation percentages above our selected cut-off of 100x coverage and that Benjamini-Hochberg correction should be used for multiple testing correction [33]. Besides analysing the individual CpG locations, we also used the "regionCounts" function to aggregate results for each of the amplicons. Differential methylation analysis was performed as described above.

However, BS-seq results provide methylation percentages and per base coverages. These are modelled with logistic regression models by the methylKit package. Again, one could consider adding a subject specific effect for the global comparison. However, inclusion of subject specific fixed effects is known to result in biased point estimates for multinomial data [34, Chapter 13.3]. Hence, more sophisticated methods are needed to address subject specific effects for multinomial data.

Moreover, the user interface of methylKit limited the analyses to simple two group comparisons. Non-progressors and progressors differential methylation was compared for the first round, second round and all samples. The latter comparison did not take potential correlation between first and second round measurements into account in any way. The results should therefore be interpreted with care.

The R-code employed in this chapter may be consulted in Appendix A.2.

3.2.7 Pathway Analyses

Analyses were performed with the Entrez identifiers obtained in the different gene rankings. These identifiers where then used to perform singular enrichment analysis with the online tool WebGestalt [35]. As background we used the "hsapiens_genome" and specified the following non-default options: "Benjamini Hochberg" multiple testing correction and an FDR cut-off of 0.001. WebGestalt offers a choice of reference databases. In this chapter we limit ourselves to the KEGG database.
WebGestalt is a pathway analysis tool of the SEA class with competitive hypothesis testing and a gene permutation strategy for P-value estimation. This implies that the Type I error control is not optimal as the underlying hypergeometric test assumes that methylation levels in the different genes are independent. For gene expression this is usually not the case. For MCs the problem of correlation is harder to quantify as each gene contains several MCs. Although neighbouring MCs are often correlated, wider spaced MCs in the same gene often are not. A ranking of the pathways is useful for interpretation of the data, but the FDR estimates should be interpreted with care.

3.3 Results

3.3.1 Differential Methylation Profiles

We opted to sequence both round 1 and round 2 DNA of 10 progressors and 10 non-progressors after enrichment for CpG methylation (MethylCap-seq) as a cost-effective means to perform genome wide screening. To minimise the effect of potential confounders we carefully matched subjects in both groups (see section 3.2.4) and opted for stringent inclusion criteria: non-smoking, no arterial hypertension, no use of statins, BMI $< 31 \text{ kg/m}^2$ and all subjects were female. See Supplementary Table C.1 for characteristics of both groups.

Summary statistics for the MethylCap-seq results are listed in Supplementary Table C.3. We obtained a median of 38,340,030 ([30,638,088; 53,897,975] IQR) paired-end 50 bp reads per sample. As quality control we verified technical replicates, CpG enrichment and sample identity. Enrichment of CpG containing regions was successful and in line with previously published results for the kit used (see Supplementary Figure B.1) [26]. By using both the methylation profiles (vide infra) for unsupervised clustering and SNP information, we concluded mistaken sample identity was unlikely (see section 3.2.5).

To summarise the MethylCap-seq data we used a predefined set of genomic locations based on aggregated MethylCap-seq samples. We will refer to these locations as methylation cores (MCs, see section 3.2.5). We modelled methylation by coding the progressor status and study round in one variable with four levels: round 1 non-progressor, round 1 progressor, round 2 non-progressor and round 2 progressor.
This design does not include a variable for the individual subjects (see section 4.2.3.2). This means that our differential methylation analysis comparing all progressor samples to all non-progressor samples, could be overconfident because the correlation between first and second round samples of the same subject was not taken into account (see section 3.4). For the comparisons of first and second round samples separately, this should not be a problem.

We inspected the density plots of the P-values for the contrasts of interest to those obtained from a permuted dataset. For both rounds combined and the second round, it was clear that low P-values were more enriched in the actual results. This indicated that many MCs for these two comparisons were likely truly differentially methylated. However, the slight enrichment of lower P-values in the permutation for the two rounds combined, indicated that the FDR might not be properly controlled at its nominal value. For round 1 results, there was clear enrichment of low P-values, comparable to the actual results set. It is therefore expected that many false positives were returned for this comparison. A relatively low number of MCs was called significant in this round and these results should be interpreted with care (see Supplementary Figure B.4).

In differential methylation results obtained for all samples (from both study rounds), 675 MCs were significantly differentially methylated between progressors and non-progressors (FDR < 0.05, see digital Supplementary Table DS1). Of these MCs, 364 were intergenic and 311 were located in 322 Ensembl annotated features. This represents an enrichment of intergenic locations compared to the full set of MCs (Chi-square p = 0.007) but the proportions of genic locations were representative (promoter: 30, exon: 32, intron: 296; Chi-square p = 0.292). Overall there were 524 hypomethylated MCs and 151 hypermethylated MCs in progressors.

Table 3.2: Differentially methylated loci between atherosclerosis progressors and non-progressors in the first study round.

<table>
<thead>
<tr>
<th>Methylation Core</th>
<th>Ensembl Gene Id</th>
<th>HUGO</th>
<th>chr</th>
<th>beg</th>
<th>end</th>
<th>logFC</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>41185718</td>
<td>ENSG00000134317</td>
<td>GRHL1</td>
<td>2</td>
<td>10138660</td>
<td>10138991</td>
<td>2.79</td>
<td>8.08E-09</td>
<td>0.017</td>
</tr>
<tr>
<td>41845833</td>
<td></td>
<td></td>
<td>4</td>
<td>127021827</td>
<td>127022269</td>
<td>2.35</td>
<td>9.70E-09</td>
<td>0.017</td>
</tr>
<tr>
<td>41314995</td>
<td></td>
<td></td>
<td>2</td>
<td>117595094</td>
<td>117595469</td>
<td>4.06</td>
<td>2.94E-08</td>
<td>0.028</td>
</tr>
<tr>
<td>42229427</td>
<td></td>
<td></td>
<td>6</td>
<td>83420498</td>
<td>83420499</td>
<td>-4.32</td>
<td>3.88E-08</td>
<td>0.028</td>
</tr>
<tr>
<td>42948695</td>
<td></td>
<td></td>
<td>X</td>
<td>73097124</td>
<td>73098294</td>
<td>2.06</td>
<td>3.92E-08</td>
<td>0.028</td>
</tr>
<tr>
<td>39655520</td>
<td></td>
<td></td>
<td>11</td>
<td>34048277</td>
<td>34048520</td>
<td>4.68</td>
<td>9.89E-08</td>
<td>0.043</td>
</tr>
<tr>
<td>41689643</td>
<td>ENSG00000213139</td>
<td>CRYGS</td>
<td>3</td>
<td>186263000</td>
<td>186263117</td>
<td>-3.21</td>
<td>1.08E-07</td>
<td>0.043</td>
</tr>
<tr>
<td>41872398</td>
<td>ENSG00000164142</td>
<td>FAM160A1</td>
<td>4</td>
<td>152444053</td>
<td>152444054</td>
<td>-3.21</td>
<td>1.09E-07</td>
<td>0.043</td>
</tr>
<tr>
<td>42315903</td>
<td></td>
<td></td>
<td>6</td>
<td>158699137</td>
<td>158699180</td>
<td>-3.21</td>
<td>1.10E-07</td>
<td>0.043</td>
</tr>
<tr>
<td>40170716</td>
<td>ENSG00000215407</td>
<td>AC126603.1</td>
<td>15</td>
<td>20480571</td>
<td>20481091</td>
<td>-1.53</td>
<td>1.20E-07</td>
<td>0.043</td>
</tr>
<tr>
<td>41247066</td>
<td></td>
<td></td>
<td>2</td>
<td>57560789</td>
<td>57560947</td>
<td>1.54</td>
<td>1.50E-07</td>
<td>0.045</td>
</tr>
<tr>
<td>42425056</td>
<td>ENSG00000196313</td>
<td>POM121</td>
<td>7</td>
<td>72402073</td>
<td>72402605</td>
<td>1.09</td>
<td>1.51E-07</td>
<td>0.045</td>
</tr>
<tr>
<td>40948333</td>
<td></td>
<td></td>
<td>1</td>
<td>222034296</td>
<td>222034672</td>
<td>2.09</td>
<td>1.62E-07</td>
<td>0.045</td>
</tr>
</tbody>
</table>
Examining only first round samples, i.e. to identify loci that independently predict atherosclerosis progression, we found 13 MCs that were differentially methylated (FDR < 0.05, Table 3.2). For the second round samples, thereby assessing the relation between atherosclerosis and the leukocyte methylome, there were 26 significant MCs (Table 3.3). There was no overlap between the first and second round significant MCs. This could be due to the large number of MCs that make multiple testing correction very strict or the fact that a number of false positives are expected in the first round results as indicated by the permutation study.

Table 3.3: Differentially methylated loci between atherosclerosis progressors and non-progressors in the second study round.

<table>
<thead>
<tr>
<th>Methylation Core</th>
<th>Ensembl Gene Id</th>
<th>HUGO</th>
<th>chr</th>
<th>beg</th>
<th>end</th>
<th>logFC</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>40171412</td>
<td>39665408</td>
<td>42509384</td>
<td>39673314</td>
<td>42299298</td>
<td>41365752</td>
<td>42559165</td>
<td>4088885</td>
<td>40014313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>8</td>
<td>11</td>
<td>6</td>
<td>6</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21910675</td>
<td>26609621</td>
<td>46026544</td>
<td>16416305</td>
<td>160704508</td>
<td>19090947</td>
<td>157813530</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21911919</td>
<td>26609685</td>
<td>46026566</td>
<td>164164577</td>
<td>160704633</td>
<td>190909583</td>
<td>157813753</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparing the overlap between the significant hits in one round and the list of MCs meeting the nominal significance cut-off of \( P < 0.05 \) in the other round we find that 2 of 13 round 1 loci were present in the second round list and 3 of 26 round 2 loci were present in the first round list. When we evaluated the overlap by associated gene names at the \( P = 0.05 \) level, 4 out of 5 round 1 genes and 11 out of 15 round 2 genes were significant in the other round as well (note that genes may cover multiple MCs and that several MCs are not located in annotated genes). Moreover, the predominant direction of differential methylation was identical for all genes.
3.3.2 Validation

We validated a subset of our findings by targeted bisulphite deep sequencing. An independent set of 23 progressors and 23 non-progressors was selected with the same criteria and matching procedure (see Supplementary Table C.2 for summary statistics). We opted for a non-overlapping set of volunteers as we already validated the utility of MethylCap-seq as an exploratory tool [36] and given limited resources this choice should result in the most biologically reproducible results. The validation loci were manually chosen from the significant MethylCap-seq results. The amplification primers are listed in Supplementary Table C.4. It should be noted that insufficient DNA was available for 4 progressors and 2 non-progressors in the second round. Additionally, some subject-primer pair combinations yielded no PCR amplification after bisulphite treatment (see digital Supplementary Table DS3 for coverage in each).

A similar definition of the design matrix was not possible because the interface of the methylKit package, used for analysis of the validation data, only allows the user to pass a single factor with two levels (in this case "progressor" or "non-progressor"). This did not pose a problem for the progressor versus non-progressor comparisons of the first or second round data, which we performed separately, as only one factor (progression status) needs to be compared. However, for the comparison of all progressor versus all non-progressor samples, this most likely led to anti-conservative results as the assumption of independent samples was clearly not met.

We successfully converted, amplified and sequenced 15 out of 17 selected loci which showed evidence of differential methylation between progressors and non-progressors in at least one of three hypotheses, i.e. first round differential methylation, second round differential methylation and differential methylation irrespective of time point. Out of these 15 regions, 10 were significantly differentially methylated over the entire region in at least one of the 3 hypotheses and 13 showed differential methylation in at least one hypothesis at the individual CpG level (see Table 3.4 for an overview). Yet at the regional level the direction of differential methylation corresponded with the results obtained from MethylCap-seq data in only 7 loci for one of the hypotheses. At the level of the individual significant CpGs we found that the same 7 regions and ADRA1A contained at least one CpG with corresponding direction in at least one of the hypotheses.
Table 3.4: Overview of validation experiment results at the amplicon level.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Methylation core</th>
<th>Gene</th>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progression 4</td>
<td>146146400</td>
<td>LNC4</td>
<td>6</td>
<td>146146470</td>
<td>146146705</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18888888</td>
<td>ANTXR2</td>
<td>4</td>
<td>87231788</td>
<td>87232178</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18888888</td>
<td>ANTXR2</td>
<td>4</td>
<td>87231788</td>
<td>87232178</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18888888</td>
<td>ANTXR2</td>
<td>4</td>
<td>87231788</td>
<td>87232178</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18888888</td>
<td>ANTXR2</td>
<td>4</td>
<td>87231788</td>
<td>87232178</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18888888</td>
<td>ANTXR2</td>
<td>4</td>
<td>87231788</td>
<td>87232178</td>
<td>Yes</td>
</tr>
<tr>
<td>Progression 4</td>
<td>18045000</td>
<td>MAPK10</td>
<td>4</td>
<td>87231800</td>
<td>87232200</td>
<td>Yes</td>
</tr>
</tbody>
</table>
This rate of confirmation was lower than the FDR set in the screening stage but this was not unexpected considering the fact that FDR calculations are very assumption dependent and typically vary between used statistical methods. Additionally, the actual methylation differences were relatively small (typically between 0.5% and 3%, see Figure 3.1 for a few examples). Differential methylation analysis results of the amplicons and individual CpGs are listed in digital Supplementary Tables DS4 and DS5 respectively.

3.3.3 Pathway Analyses

Potentially affected pathways were identified using the online tool WebGestalt [35]. This is a tool of the SEA class and uses gene permutation for P-value estimation (see section 1.3.8). This means that correlation between features (MCs in this case), is not taken into account which may lead to liberal rejection of the null hypothesis. The results should therefore be interpreted with care.

While a low FDR is crucial for interpretation of the gene-level results, pathway analysis benefits from a larger set of putatively differential loci (i.e.
increased power), even if this means including slightly more false positives. Therefore, we relaxed the gene-wise significance threshold for pathway analyses to 10% FDR but made the requirements for pathway significance stricter (FDR<0.001).

For the first round this yielded no significant results (data not shown). In the second round, only 1 pathway was significant at FDR<0.001 (Table 3.5). The number of differentially methylated MCs for round 1 and round 2 samples combined was much larger. Consequently, no less than 35 pathways were potentially affected (FDR<0.001, digital Supplementary Table DS2).

3.3.4 Potential Genomic Deletion

Based on the MethylCap-seq data we concluded that a genomic deletion in the lncRNA RP11-10O22.1 (ENSG00000241168) of progressors was a plausible explanation for the observed profiles (see Figure 3.2, Fisher exact test $P = 0.005$). At the time of writing nothing is known about this transcript but both losses and gains of this region were reported in individuals from various ethnicities (see Database of Genomic Variants [37]). As expected we saw higher failure rates for this locus in the BS-seq data but, unlike the pilot group, the distribution of these failures alone was not significant in the validation group ($P = 0.5559$, Fisher exact test). However, the methylation percentages that we did obtain in the validation study were significantly different between progressors and non-progressors both for individual CpGs ($\geq 0.27\%$, q-value $\leq 0.0002$) and the amplicon as a whole ($0.27\%$, q-value=9.886E-11) which implies the possibility of similar or combined modulation of gene expression by methylation or deletion.
<table>
<thead>
<tr>
<th>Pathway</th>
<th>Entrez Gene Id</th>
<th>Pathway size</th>
<th>Observed</th>
<th>Expected</th>
<th>Ratio</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vascular smooth muscle contraction</td>
<td>59, 10242, 148, 114</td>
<td>116</td>
<td>4</td>
<td>0.14</td>
<td>28.06</td>
<td>1.31E-05</td>
<td>0.0001</td>
</tr>
<tr>
<td>Calcium signaling pathway</td>
<td>7220, 148, 114</td>
<td>177</td>
<td>3</td>
<td>3.00</td>
<td>14.79</td>
<td>0.001</td>
<td>0.0063</td>
</tr>
<tr>
<td>Progesterone-mediated oocyte maturation</td>
<td>5602, 114</td>
<td>86</td>
<td>2</td>
<td>0.11</td>
<td>18.92</td>
<td>0.005</td>
<td>0.0103</td>
</tr>
<tr>
<td>Salivary secretion</td>
<td>148, 114</td>
<td>89</td>
<td>2</td>
<td>0.11</td>
<td>18.29</td>
<td>0.005</td>
<td>0.0103</td>
</tr>
<tr>
<td>GnRH signaling pathway</td>
<td>5602, 114</td>
<td>101</td>
<td>2</td>
<td>0.12</td>
<td>16.11</td>
<td>0.007</td>
<td>0.0103</td>
</tr>
<tr>
<td>Pancreatic secretion</td>
<td>7220, 114</td>
<td>101</td>
<td>2</td>
<td>0.12</td>
<td>16.11</td>
<td>0.007</td>
<td>0.0103</td>
</tr>
<tr>
<td>Protein processing in endoplasmic reticulum</td>
<td>5602, 7324</td>
<td>165</td>
<td>2</td>
<td>0.20</td>
<td>9.86</td>
<td>0.018</td>
<td>0.0198</td>
</tr>
<tr>
<td>Ubiquitin mediated proteolysis</td>
<td>4734, 7324</td>
<td>135</td>
<td>2</td>
<td>0.17</td>
<td>12.06</td>
<td>0.121</td>
<td>0.0156</td>
</tr>
</tbody>
</table>

*Table 3.5: KEGG pathways differentially methylated between atherosclerosis progressors and non-progressors in second round.*
3.4 Discussion

In this study we demonstrated for the first time in a human population that atherosclerosis progression is reflected in consistent but subtle modifications in the peripheral blood leukocyte DNA methylation profile, independent of other known major risk factors. Importantly, several loci were independently associated with atherosclerosis progression in apparently healthy subjects, thereby providing atherosclerosis susceptibility biomarkers. However, differences were typically subtle (< 5% difference in methylation degree), which may be related to the study design but also the validation methodology.

Through limitations arising from the need to perform genome-wide experiments in a cost-efficient manner, the challenges involved in interpreting DNA methylation profiles from whole blood are considerable. Firstly, under ideal circumstances separate analyses of DNA obtained from the various leukocyte subtypes would have yielded more detailed information than whole blood. It has been demonstrated that the complex mix of cell types that make up the circulating leukocytes, can confound epigenetic profiles obtained in the context of ageing as was demonstrated using Illumina’s Infinium HumanMethylation 450k bead array technology. These effects are fairly consistent though, which allowed for the development of algorithms that estimate the proportions of the major cell types [38, 39]. Of the 49950 MCs that matched cell type associated probes, only 2 were found to be differentially methylated in the present study, suggesting that most of our results do not solely reflect changes in cell type composition of the subjects’ blood. Even if the latter would be the case, our data provide a comprehensive basis to elucidate associations of leukocyte fractions with atherosclerosis progression in the future.

The second drawback to this mix of cell types is that highly significant changes in one type might be diluted by a steady or opposite methylation levels in the other types. Whereas in a single cell type the theoretical methylation values should approach 0, 50 or 100%, our validation experiment found a range of intermediary methylation values (mostly larger than 50%). Further, the absolute differences were smaller than one might expect based on the MethylCap-seq data, although this can at least partially be attributed to the fact that the MethylCap-seq assay requires a threshold degree of methylation (i.e. multiple neighbouring methylated CpGs) before a locus can be picked up consistently [26].

Third, there are still considerable limitations to the methods available for the analysis of both the MethylCap-seq and BS-data. Even edgeR, despite
being one of the most flexible tools to date, has limitations. The NB nature of sequencing data complicates the analysis of a complex design such as ours. More specifically the presence of paired samples can be seen as repeated measurements of the same subject for the comparison of progressors and non-progressors at both time points combined (between subject comparison). This might have led to some overoptimism in rejecting null hypotheses for this comparison as measurements are likely correlated between rounds. An analysis of a permuted version of the data indicated this methodological limitation may have had some impact on the comparison of all samples and a significant influence in the first round samples. Further, the analysis tool edgeR does not allow the addition of random effects to model subject identity. If we had added subject identity as a fixed effect this would absorb any variation between progressors and non-progressors. More research into the consequences of adding subject specific fixed effects to models of NB data is needed [28].

Lastly, the methylKit package used for the validation data uses edgeR for the underlying differential methylation calculations. However it is far more limited as it only allows the user to pass a single treatment factor with two levels. Again, this did not pose a problem for comparisons in the separate rounds, but likely led to some overoptimism in calling genes differentially methylated when we compared all progressors to all non-progressors because the correlation between first round and second round measurements of the same subject, were not taken into account. A rewriting of the code was beyond the scope of this thesis but would be essential to use the full potential of our experimental design.

Given these limitations the reproducibility of the loci we identified could be questioned. However, the fact that many results were still consistent upon validation, in two independent population samples and with two different techniques even in such a complex background, suggests that future research focussing on the identification of the cell types driving the observed methylation differences may lead to biomarkers with clinical utility.

### 3.4.1 Gene Methylation and Atherosclerosis Progression

Several identified genes fulfil roles that can be linked with atherosclerosis. To underscore the relevance of our results we will highlight a few. For example, *GRHL1*, the most significant result for round 1, is one of the genes activated
by \textit{PPAR}_\alpha, which is an important regulator of lipid metabolism in atherosclerosis [40]. We also found several genes directly related to atherosclerosis and inflammatory processes such as \textit{LY75} (also known as CD205), a modulator of oxLDL metabolism of dendritic cells [41] and the apoptosis inhibitor \textit{CD5L} (a.k.a. AIM or API6) [42]. A potential role for other genes is less straightforward, but the link to atherosclerosis is plausible. \textit{CHGN2} expression for instance, was increased by a pro-atherosclerotic diet in mice [43]. Another gene, \textit{MAPK10} (a.k.a. JNK3), is a member of the janus kinases which are mostly associated with atherosclerosis in vascular cells [44]. Yet \textit{MAPK10} was proposed as a key gene with regard to focal adhesion in coronary artery disease [45]. The differentially methylated gene adenylate cyclase 8 (\textit{ADCY8} a.k.a. AC8 or ADCY3), encodes a membrane bound protein which is involved in the vascular and muscular contraction, Ca\textsuperscript{2+} signalling and several secretory pathways. Adenylate cyclases in general are involved in transendothelial migration of leukocytes [46]. Lastly, mutations in \textit{ACTA2}, one of the round 2 candidate genes, were associated with a range of vascular diseases [47].

A potential deletion in lncRNA RP11-10O22.1 was strongly associated with atherosclerosis progression in our genome wide screening but not in the validation population. This result is not entirely unexpected as a deletion would constitute a heritable, non-modifiable change and would most likely have been detected in previous research. None the less, this locus showed differential methylation in our validation population which might obscure or mimic the effects of a deletion. Further research into the association between this locus and atherosclerosis is warranted.

### 3.4.2 Relation to Biological Processes

Among the pathway level results we found several biological processes that might be related to atherosclerosis progression. Pathways such as ”Calcium signaling pathway”, ”Regulation of actin cytoskeleton” and ”Leukocyte transendothelial migration” might reflect macrophage mobilisation to sites of atherosclerotic plaque development. A less expected result was the ”Neuroactive ligand-receptor interaction” pathway. Closer inspection of the latter pathway revealed differential methylation in several receptor encoding genes that are associated with atherosclerosis or related CVD before such as glucagon-like peptide receptors [48], growth hormone receptor [49] and leptine receptor [50].
Pathway analyses are inherently limited to those pathways which have already been described. But besides the possibility of yet unidentified pathways, nearly half of the differentially methylated MCs we identified, stem from unannotated loci (not within 2000 bp upstream or the 3’ end of an Ensembl feature) and were validated at a comparable rate as the annotated loci. At this point it is too early to speculate how these loci might be involved but several mechanisms of action are possible (unidentified lncRNA [51], regulation of nearby genes [52], structural conformation of DNA [53]...).

### 3.4.3 Relation to Prior Research

Previous research [9] identified 1858 Illumina 450k probes with differential methylation between paired atherosclerotic lesions and non-atherosclerotic tissue in veins from the same subjects. Although we examined leukocytes and the commonalities of our results should be limited, we did attempt a comparison. Indeed none of the 16 of our differentially methylated MCs that overlap an Illumina 450k probe (progressors vs non-progressors both rounds, FDR < 0.05), was differentially methylated in vascular lesions (data not shown).

### 3.4.4 Predicting Atherosclerosis Progression

There is a long standing interest in predicting the progression of atherosclerosis and subsequent cardiovascular events [54]. Our study population was relatively young and selected to be a population representative sample without clinically overt CVD at baseline and can thus be expected to have good cardiovascular health at enrolment. Due to the younger age and the relative health of the population, it is too early to have significant outcome data. Therefore we currently do not dispose over data indicating a heightened risk of events. However, we did identify 13 genomic locations which show a strong correlation with atherosclerosis progression in the subsequent decade. Most of the sites are not located within a known gene and the potentially affected genes are currently not known in an atherosclerosis context (*GRHL1, POM121*) or have no confirmed function at all (*FAM160A1*). None the less, in our validation population we confirmed 2 of the 4 selected loci.

The performance of biomarkers is indifferent to the true biological underpinnings and this study showed that leukocyte DNA methylation has potential for early classification of patients from a simple blood sample. The next step in evaluating these markers is to test their performance in a larger cohort.
Figure 3.2: Methylation profile at a site of differential methylation obtained by MethylCap-seq. The first 10 samples are non-progressors, the last 10 are progressors. This region encompasses a known copy number variation site.
3.5 Conclusions

DNA methylation constitutes a measurable, modifiable risk factor for atherosclerosis. We found that atherosclerosis progression is reflected in an epigenetic profile throughout the genome of peripheral blood leukocytes. These methylation differences are likely driven by a subset of cell types. We further demonstrated that healthy subjects prone to atherosclerosis development in the subsequent decade, were characterised by differential methylation in a small set of CpG locations. Validation experiments at a larger scale are required to evaluate their potential as biomarkers.

3.6 Acknowledgements

Besides the acknowledgements from chapter 2, we would also like to thank Sarah De Keulenaer, Ellen De Meester, Johan Vandersmissen and Jean-Pierre Renard for their indispensable mastery of the wet-lab side of this study.
References


A Longitudinal Assessment of Ageing Related Leukocyte DNA Methylation Alterations in the Middle-aged Asklepios Study Population

4.1 Introduction

Ageing is widely recognised as a complex phenomenon that is influenced by several genes [1], disease, environmental factors and random chance [2].

Currently several studies have found DNA methylation markers of chronological age, particularly in whole blood leukocytes but also in saliva [3, 4]. However, it is difficult to isolate epigenetic effects related to ageing from the host of other factors that influence an individual’s epigenetic profile such as the proportions of leukocyte subtypes and generational effects.

Several studies partially remediate these problems by careful analysis adjusting for cellular heterogeneity using a computational approach [5] or the
use of twins in an attempt to reduce background noise [6–8]. A very recent cross-sectional study analysed age-related DNA methylation changes in leukocyte sub-populations (monocytes and T-cells). This study demonstrated that T-cell specific epigenetic alterations are also found in monocytes. However, when limiting these results to loci that are also functionally relevant to the expression level of corresponding gene, the overlap is absent. Upon further evaluation of the latter loci, hypomethylation is predominantly found in older age, located in predicted enhancers and often linked to antigen processing and presentation genes [9]. This study somewhat contrasts a previous investigation of a leukocyte DNA methylation based ageing model that also appears to be functional for different tissues and is reflected in the transcriptome, though here no formal discrimination was made between expression and non-expression linked ageing associated DNA methylation changes [4].

Based on these and other studies, e.g. [10], it is clear that a fraction of whole blood leukocyte ageing results will reflect cellular heterogeneity, but also that a large fraction — at least for DNA methylation — is shared between leukocyte subsets, making whole blood leukocytes an interesting topic of research. Generational effects may arise with respect to telomere length, another potential problem ignored in said studies [11, 12]. For example, air pollution is linked with leukocyte DNA methylation alterations [13], has an age dependent impact [14] and is also subject to evolution over time [15]. In other words, putative age dependent alterations may as well reflect changing external factors. To the best of our knowledge no studies yet have used the most straightforward method to account for such effects: to follow up epigenetic changes in individuals over the course of several years and to use the baseline measurements to control for inter-individual and inter-generational methylation differences.

We used methylation enrichment sequencing which offers a broader scope than most studies mentioned so far, which rely primarily on methylation probe arrays. One previous study used this methodology for a large sample of subjects with a broad age-span [16]. In this chapter we present the first methylome-wide methylation data obtained from whole blood in the same subjects at two distinct time-points.
4.2 Methods

4.2.1 Study Population Selection

For the analyses reported here, we used the same two non-overlapping subsets of 20 and 46 Asklepios study volunteers as described in Chapter 3. The methodology is very similar and only the differences will be highlighted here.

4.2.2 Experimental Design

For the purpose of this ageing study we used atherosclerosis progressor status as a potential confounding factor. A general population sample would most likely contain a significant number of (subclinical) atherosclerosis cases and ageing differences between progressors and non-progressors are of secondary interest. However, we accounted for a potential interaction between progressor status and ageing in order to use an appropriate model for each methylation core (see section 4.2.3.2). Although study design was the same, the primary difference regarding the analyses was that accounting for the correlations between the paired samples could boost the power to detect ageing related methylation differences.

4.2.3 Methylated CpG Enrichment Sequencing

4.2.3.1 Pre-processing

Data generation and pre-processing for this experiment was described in section 3.2.5. Briefly, we generated MethylCap-seq reads, mapped them, generated coverage profiles for the methylation cores (MCs) in the Map of the Human Methylome and analysed differential methylation in R (v3.0.1). The MCs were annotated based on their location relative to Ensembl features (release 72). Prior to analyses, the technical replicates with the highest coverage (AS1.20 and AS2.47) were selected, reducing the number of samples to 40. Library size scaling normalisation was applied.
4.2.3.2 Analysis Strategy

In this chapter three main research questions are of interest. First, what MCs changed in methylation degree between study rounds? Second, were there age-associated patterns in non-progressors? And third, were there age-associated differences in the progressors? The paired nature of the samples offers an additional advantage in the sense that baseline measurements can be adjusted for individually in each subject. Conversely, not taking the paired nature of the samples into account, ignores correlations between the measurements in both rounds and leads to a loss in power since we assessed within-subject effect in this chapter.

One way to pose these questions would be to define a model with three factors: study round, progressor status and subject identity. A classical approach for data analysis consists of a mixed effects model with two fixed effects for round and progressor status, and a random effect for subject identity. However to the best of our knowledge, there were no analysis packages to date that were able to model random effects while borrowing strength from the large number of features in a sequencing experiment while offering the same usability as edgeR.

Another approach is to incorporate subject identity as a fixed effect in the model. This has the advantage of fewer assumptions being imposed than in a traditional mixed model (i.e. normality of the random effects and orthogonality of random and fixed effects). This model can be implemented in standard software for analysing sequencing experiments. The inclusion of subject specific fixed effects leads to proper inference for within subject effects in many distributional families. However, it only allows for conditional inference.

For NB data, the use of fixed effect models that account for subject specific effects are not well studied yet. The fixed effect model can be expected to give unbiased point estimates for the within subject effects for the NB distribution but there might be issues with the estimation of the standard errors that are provided by standard software [17].

We adopted a subject specific fixed effect model in edgeR. However, incorporating a fixed effect for subject identity, study round, progressor status and their interaction would lead to an overspecified model, i.e. the difference between progressors and non-progressors at base-line would be aliased with the subject specific effects. Therefore an alternative parametrization was proposed with a subject specific fixed effect and a factor RoundProg with three
levels (Round1, NonProgRound2 and ProgRound2). Note, the Round1 level was encoded as the reference class and allowed to immediately estimate differential methylation between both rounds in non-progressors and progressors by the parameters for NonProgRound2 and ProgRound2 respectively. The global differential methylation between round 1 and round 2 could then be estimated as the average over the NonProgRound2 and ProgRound2 effects. The corresponding contrasts are listed in Table 4.1.

Table 4.1: Contrasts for differences between round 1 and round 2 with progressor status and subject specific fixed effects, \( i \) ranges from 1 to 20.

<table>
<thead>
<tr>
<th>Subject, ( i )</th>
<th>Non-progressors</th>
<th>Progressors</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 2 NonProg</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Round 2 Prog</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Since the use of models with subject specific fixed effects is not well studied, we evaluated the P-values using a single permutation in which the samples’ study round was switched for half of the non-progressor subjects and half of the progressor subjects. This represented a divergent permutation that eliminated potential round effects while preserving the within subject correlation. An enrichment of low P-values in the permuted dataset would indicate that the P-values were too liberal and inference on the real data has to interpreted with care.

4.2.4 Bisulphite Sequencing

Based on the results obtained in the MethylCap experiment we selected 6 locations to validate the epigenetic differences observed. As there was insufficient DNA available for 6 subjects, the corresponding first round samples were excluded for the purpose of paired analyses but not for the group-wise comparison. We also included a methylated and unmethylated control in the form of \textit{in vitro} methylated DNA (S7821, Millipore, Massachusetts, USA) and DNA from a methyltransferase double knock-out cell line (HCT116) for a total of 88 samples and one no-DNA blank (negative for all primers, not included in sequencing).

After bisulphite treatment we amplified the regions of interest with the primers listed in Supplementary Table C.5. All fragments were sequenced for
250 bases in a paired-end fashion. As some target regions were shorter than this length, they were in effect fully covered in duplicate (i.e. once from each side) but sequencing does not extend beyond the primer boundaries.

Data preprocessing, including base calling, demultiplexing, trimming and fastq generation was automatically performed in Illumina BaseSpace with standard settings as provided by the manufacturer (Illumina). The sequencing data was analysed using the Bismark package (v 0.10.0) [18]

Differential methylation in the BS-seq data was evaluated with the methylKit package [19]. Besides analysing the individual CpG locations, we also used the "regionCounts" function to aggregate results for each of the amplicons.

Ideally the design matrix introduced in section 4.2.3.2 should be used for the validation experiment. However, BS-seq results provide methylation percentages and per base coverages. These are modelled with logistic regression models by the methylKit package. Inclusion of subject specific fixed effects, however, is known to result in biased point estimates for multinomial data [20, Chapter 13.3]. Hence, more sophisticated methods are needed to address subject specific effects for multinomial data.

Moreover, the user interface of methylKit limited the analyses to simple two group comparisons. First and second round methylation were compared for non-progressors, progressor and all samples. This means that the latter comparison does not take progressor status into account in any way and the results should therefore be interpreted with care.

The R-code employed in this chapter may be consulted in Appendix A.2.

4.2.5 Pathway Analyses

Analyses were performed with the Entrez identifiers obtained in the MC ranking. These identifiers were then used to perform singular enrichment analysis with the online tool WebGestalt [21]. As background we used the "hsapiens_genome" and specified the following non-default options: "Benjamini Hochberg" multiple testing correction and an FDR cut-off of 0.001. WebGestalt offers a choice of references, in this chapter we limit ourselves to the KEGG database.
WebGestalt is a pathway analysis tool of the SEA class with competitive hypothesis testing and a gene permutation strategy for P-value estimation. This implies that the Type I error control is not optimal as the underlying hypergeometric test assumes that methylation levels in the different genes are independent. For gene expression this is usually not the case. For MCs the problem of correlation is harder to quantify as each gene contains several MCs. Although neighbouring MCs are often correlated, wider spaced MCs in the same gene often are not. A ranking of the pathways is useful for interpretation of the data, but the FDR estimates should be interpreted with care.

4.3 Results

As outlined in the previous chapter we studied two subsets of middle-aged individuals who are part of the Asklepios longitudinal study on cardiovascular ageing. We disposed over whole blood DNA from both the first study round (2002-2004) and the follow-up round which is ongoing at the time of writing. Subjects were called back according to their availability but the average timespan between both rounds for this subset was 9.1±0.58 years. In the first round the age-range of participants was 46±6 years (mean ± standard deviation).

To identify time-associated differences we performed a genome-wide screening experiment. First we analysed the methylation differences between both rounds in a longitudinal manner for the complete dataset but also for the non-progressors and progressors separately in section 4.3.1. We validated a subset of these findings with BS-seq in section 4.3.3. And finally we performed pathway analysis based on the results obtained in the genome-wide screening in section 4.3.4.

4.3.1 Differential Methylation Profiles

To study these questions, we modelled methylation counts using a subject specific fixed effect and ProgRound a factor with three levels (coding for round 1, round 2 non-progressor and round 2 progressor). Overall differential methylation between round 1 and round 2 was assessed in all samples as well as within non-progressors and progressors separately.
As mentioned in section 4.2.3.2 there are limitations to the analysis of paired samples in most analysis tools including edgeR which was used here. More specifically there are concerns about the underestimation of P-values [17]. To evaluate the extent of overoptimism we examined the P-value distribution of three contrasts in the real data set and a permuted data set. These contrasts correspond to the inter-round differences for all samples, only the non-progressors and only the progressors.

Supplementary Figure B.5 shows the density plots of the P-values. For the progressors the density plot shows no enrichment of low P-values compared to the permuted data set. For the non-progressors and all subjects there is substantial enrichment in the lower P-values for the real data but not for the permuted samples. This indicates that the results for these comparisons are probably reliable although there seem to be few ageing-related differences in the progressors.

We identified 14678 DMRs (FDR < 0.05, digital Supplementary Table DS6a) between the first and the second round samples, 7113 of these were not located in the promoter region (-2000 to +300 bp) or gene body of an Ensembl feature. Of the remaining 7565 DMRs, 691 were located in a promoter, 1003 overlapped an exon and 6725 overlapped an intron. This distribution represent a slight enrichment of exons and intergenic regions (Chi-square $P = 0.006$). The overlap with cell type associated probes is modest but significant (n=98, $P < 0.001$).

We identified 19829 DMRs (FDR < 0.05, digital Supplementary Table DS6b) between the first and the second round non-progressor samples, 9582 of these were not located in the promoter region (-2000 to +300 bp) or gene body of an Ensembl feature. Of the remaining 10247 DMRs, 860 were located in a promoter, 1495 overlapped an exon and 9053 overlapped an intron. This distribution represent a slight enrichment of promoters and exons (Chi-square $P = 0.001$). Again, the overlap with cell type associated probes is modest but significant (n=121, $P < 0.001$).

We identified no DMRs (FDR < 0.05, digital Supplementary Table DS6c) between the first and the second round progressor samples.

4.3.2 Relation to Prior Research

We looked for overlap between the regions identified in this study (all samples comparison) and those performed by Horvath et al. [22] (age range 3 to
101 years) and McClay et al. [16] (age range 25 to 92 years) based on the online Supplementary tables available for these publications. We found no overlap between our results (global comparison) and either of the 3 analysis approaches in these two studies for the exception of 1 methylation probe from Horvath et al. supplementary table 4, which overlaps a differentially methylated region from the inter-round analysis.

4.3.3 Validation

Table 4.2: Validation results time differences per amplicon. Amplicon details are listed in Supplementary Table C.5.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERGEN5_1</td>
<td>0.99</td>
<td>-0.34%</td>
<td>0.00</td>
<td>0.09%</td>
<td>0.00</td>
<td>-0.13%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>0.00</td>
<td>0.68%</td>
<td>0.00</td>
<td>-0.12%</td>
<td>0.00</td>
<td>0.31%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>0.00</td>
<td>-0.65%</td>
<td>0.00</td>
<td>-0.35%</td>
<td>0.00</td>
<td>-0.50%</td>
</tr>
<tr>
<td>ADRA1A_1</td>
<td>0.01</td>
<td>-1.41%</td>
<td>0.07</td>
<td>-0.20%</td>
<td>0.53</td>
<td>-0.81%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>0.00</td>
<td>0.18%</td>
<td>0.00</td>
<td>0.30%</td>
<td>0.38</td>
<td>0.23%</td>
</tr>
<tr>
<td>PUS7L_1</td>
<td>0.00</td>
<td>-0.41%</td>
<td>0.31</td>
<td>-0.57%</td>
<td>0.00</td>
<td>-0.47%</td>
</tr>
<tr>
<td>GPR176_1</td>
<td>0.57</td>
<td>-0.22%</td>
<td>0.02</td>
<td>-0.08%</td>
<td>0.07</td>
<td>-0.16%</td>
</tr>
</tbody>
</table>

From the ageing associated MCs we selected 6 for validation in an independent set of subjects. The 7 primer pairs for the targeted amplicons are listed in Supplementary Table C.5. The validation results of non-progressors and progressors separately are listed in Table 4.2. As mentioned in section 3.3.2, the methylKit package used for analysis, only allows for simple two-group comparisons. Because the validation regions were chosen from the average inter-round differences, the inter-round comparison of all samples was also made but the results should be interpreted with care as progressor status could not be taken into account.

Considering the aggregated methylation levels of the amplicons, we were able to confirm the methylation patterns (both significance and consistency) in 4 out of 6 MCs for the non-progressors and 3 out of 6 for the progressors. Considering methylation levels of individual CpGs within the amplicons, we were able to validate 5 out of 6 methylation cores for both non-progressors and progressors (Supplementary Table C.6). This increased validation rate is the consequence of some MCs containing neighbouring CpGs which show opposite methylation changes (Figure 4.1) an aspect which will be examined in more detail in section 4.4.
4.3.4 Pathway Analyses

We identified 49 KEGG pathways (FDR < 0.001, digital Supplementary Table DS7a) associated with ageing differences based on the inter-round differentially methylated regions. The top 3 pathways identified were "Metabolic pathways", "Focal adhesion” and ”Pathways in cancer”.

4.4 Discussion

In this study we reported the first genome-wide screening of DNA methylation in a longitudinal setting. The added power of a paired design allowed us to identify a multitude of loci which showed differential methylation during the ageing process of a middle-aged population, apparently healthy during the first round of the study.

As mentioned in chapter 3, the analysis of paired data is not straightforward both for MethylCap-seq and BS-seq data. More specifically current methods for paired analyses of NB data are overconfident in rejecting null hypotheses (see section 4.3.1). Although a semi-random permutation approach indicated that there were real differences between the study rounds, more conservative estimates of the significance might be appropriate. There were also indications that non-progressors were driving many of the observed associations. A definite analysis of this data requires more in-depth examination of this discrepancy and potentially the development of new methods for random effect modelling in NB data. Further, there were severe limitations with respect to the statistical methods available to model subject specific effects for bisulphite sequencing data. Developing these methods was beyond the scope of this research but would be absolutely necessary to utilize a complex experimental design such as this one.

Ageing is a complex phenomenon that affects different parts of the body in different ways. It is therefore essential to consider which aspects of ageing are universal and which are tissue specific [22]. Additionally, tissue composition can confound epigenome-wide association studies [5]. We looked for overlap between our results and a set of genomic loci associated with cell type composition of blood. The overlap was greater than compared to the atherosclerosis progression associated loci reported in the previous chapter. This suggests that our results partially reflect changing cell type composition with
ageing. Although a method for inferring cell type composition is available for Illumina 450k data, this method can currently not be applied to MethylCap-seq data.

Since there was only limited overlap between studies, further analysis is required to explain our results. The McClay et al. [16] study for instance, used a principle component analysis to ”remove unknown confounders” but this method could potentially eliminate a number of valid ageing-associated regions. This might also be due to the fact that the methylation changes were small (< 1% at the amplicon level) as evident from the validation experiment. Additionally, in this study we considered ageing of peripheral blood leukocytes in a relatively narrow age range. This age range might be one of the reasons that our results had almost no overlap with previous publications which studied correlation between DNA methylation (degree) and calendar age in a wider age range. Regions that gain or lose methylation at a rate relevant to the lifespan of a human being would go unnoticed in the short time frame of this study as we were not sufficiently powered to pick up such small differences.

Despite this lack of overlap we successfully replicated 5 out of 6 selected methylation differences in an independent subset of participants. We also observed that opposite methylation differences within a MC may occur which obscure age-associated differential methylation when considering the MC as a whole. Although we selected the validation loci based on the MethylCap-seq results, it is unclear at this point how such opposing methylation patterns in close proximity affect MethylCap-seq results in general. These consideration will be further evaluated in chapter 5.

We also performed pathway analyses. We chose WebGestalt as it is a tool that could be easily combined with the MC rankings we obtained. It is a competitive hypothesis test based on gene permutations and therefore does not take potential correlation between MCs into account. The pathway results may be anti-conservative. Further, these analyses assume that differential methylation affects pathways in a way similar to differential expression. This is another reason why we chose a simple enrichment based pathway analysis tool. Although methylation, in most cases, likely doesn’t correlate linearly with expression, the same is true for gene expression and pathway functionality. A doubling of the expression of a non-rate limiting enzyme in a metabolic pathway may not have an appreciable effect on the pathway as a whole. Further there are genes with pleiotropic effects and pathways such as ”Pancreatic secretion” are likely not truly active in leukocytes whether the
pathway was identified in expression or methylation data. Therefore the interpretation of pathway level results has similar limitations for expression and methylation data, they are indicative, not final.

We observed a near complete overlap between significant age-associated pathways (gene level paired FDR < 0.05, pathway level FDR < 0.001, n=67) and atherosclerosis associated pathways (gene level FDR < 0.05, pathway level FDR < 0.001, n=35). This raises an interesting point of discussion: to what extent is ageing of PBL equivalent to the effect of atherosclerosis progression on PBL? Inspecting the results at the level of the individual methylation cores, the overlap between atherosclerosis (FDR < 0.05, n=675) and time related dif-

**Figure 4.1:** Methylation degree in 7 CpGs which are significantly differentially methylated between study rounds.
ferences (FDR < 0.05, n=14678) is limited to 4 MCs. Repeating this exercise at the level of Ensembl gene identifiers, there were 158 overlaps (FDR < 0.05, atherosclerosis n=323, time n=4845). We further compared the Entrez identifiers between the overlapping pathways where the overlap of 126 was in the same order of magnitude (atherosclerosis MCs, FDR < 0.05; time MCs, FDR < 0.05). We also examined these overlaps for atherosclerosis progression FDR < 0.1 and although the overlap was greater both for MCs and genes, the overlap was relatively small (data not shown). For each shared pathway we calculated the overlap and union of affected genes in either time or atherosclerosis and listed the ratio of the overlap divided by the union (Supplementary Table C.7). This ratio was 33.3% on average, ranging between 16.8% and 47.1%.

To summarise, there is minimal overlap at the level of the MCs, modest overlap at the gene level and high overlap at the pathway level. These results suggest our experimental design was successful at distinguishing atherosclerosis related methylation changes from those associated with ageing. Yet at the functional level, there is considerable similarity between the changes associated with ageing and the changes associated with atherosclerosis progression.

### 4.5 Conclusions

We studied differential methylation patterns associated with ageing across two study rounds spanning roughly a decade. We found a multitude of loci with altered methylation degree between study rounds. Further methodological improvements with respect to the statistical analysis of sequencing count data are required to fully utilise the paired study design. These initial analyses suggest that the distinct methylome changes in leukocytes associated with ageing and atherosclerosis progression, are closely related at the pathway level.
References


Conclusions and Perspectives

5.1 Introduction

The completion of the first human draft genomes around the year 2000 insti-
gated a never seen attempt to unravel the genetic determinants of most human
phenotypes and diseases. Though not completely unsuccessful, at the turn of
the last decade it became clear that the huge budgets spent in, for example
genome-wide association studies, did not deliver a fraction of the expected
genetic causal variants [1].

Though most likely many variants remain to be identified, and the impact
of complex effects, e.g. epistasis, is hard to assess with the existing models,
it has become clear that another facet of biology may be an at least equally
important driver of human variation and disease: the gene-environment inter-
action. If the modifications in gene expression due to a varying environment
can also be inherited meiotically, but most often mitotically, the phenomenon
is typically dubbed ”epigenetics”. Several studies, for example Schuebel et
al. [2], have suggested that the impact of epigenetics is often larger than the
role of genetics. Perhaps this is fortunate as the involvement of epigenetics
also implies that a subject’s future is less predetermined from birth than was
originally anticipated.
The advances made in genetics technology, e.g. second generation sequencing and microarrays, can also be utilised for epigenetics and gene expression studies alike, which has led to a wave of novel studies. Initially, mostly cancer epigenetics studies were performed, with DNA methylation featuring as the most widely analysed epigenetic trait. Though cancer remains the focus of most epigenetics studies, more recent successful efforts also focused on ageing associated diseases such as Alzheimer’s disease [3], osteoporosis [4] and cardiovascular disease [5].

In this doctoral thesis, the link between epigenetics and cardiovascular ageing was further explored, with results regarding both telomere biology and DNA methylation. This research used samples from and is situated within the framework of the large-scale Asklepios Study on successful (cardiovascular) ageing, which tries to build a complete picture of the ageing process and has always envisaged to include epigenetics analyses, such as telomere length and DNA methylation studies.

5.2 Telomere Length

In chapter 2 we explored how peripheral blood leukocyte telomere length relates to a number of diagnostic parameters for vascular stiffness, cardiac stiffness, systolic function and ventricular mass in the Asklepios population. Although each of the parameters has been associated with PBL TL before, this was usually performed separately and in smaller studies of populations with a specific phenotype (i.e. old age or CVD). This was the first study to describe how these parameters were related to TL in a healthy middle-aged population.

Most of the parameters were not significantly associated with PBL TL in this population. This does not preclude the possibility that some of them will correlate more with PBL TL as the Asklepios study ages. Based on our current results it seems less likely that shortening telomeres precede several indicators of cardiovascular ageing and disease such as ejection fraction, pulse wave velocity and left ventricular mass.

However, we did confirm a correlation with E/A and e’/a’. These ratios are measures of diastolic function and we discovered an unexpected correlation between PBL TL and isovolumic relaxation time, another parameter of
diastolic function. We also observed a slight difference in correlation strength between women and men, but this proved non-significant in the current data set. Future Asklepios Study rounds will follow up on this difference as women have an increased occurrence of diastolic heart failure [6].

The question whether TL shortening is a cause, consequence or innocent bystander of CVD and ageing, is an unresolved debate [7]. Altered filling patterns are an early sign of diastolic dysfunction and this leaves room for a causal role for telomeres. But although TL correlates reasonably well between tissues [8], we currently cannot propose a mechanism by which it might be linked to diastolic function. Our results encourage a line of thought where TL is associated early on, passively if not causally, in the pathogenesis of diastolic dysfunction and possibly heart failure by extension. At the time of writing the second study round is incomplete but repeating the analyses from chapter 2 might provide more insight into the evolution of TL and diastolic function.

5.3 DNA Methylation and Cardiovascular Ageing

In chapters 3 and 4 we used a subset of the Asklepios study for which second round examinations were completed and stored DNA was available. With the second round phenotypes available, we were able to design a study that sheds light on the correlations between DNA methylation, atherosclerosis development and ageing.

We first performed a genome-wide screening by methylation enrichment sequencing and selected several loci for validation in an independent subset of participants by targeted bisulphite deep sequencing. This approach led to identification of a multitude of genomic loci which show differential methylation between study rounds or between atherosclerosis progressors and non-progressors. These loci were located both in annotated genes and intergenic regions and were validated at similar rates in both. This underscores the value of performing genome-wide screening rather than limiting oneself to a predefined set of regions as one does with array based technologies.

Although we chose our subjects and experimental design to eliminate as much bias as possible, the currently available techniques for differential count analysis are inadequate to utilise the full extent of the study design.
More specifically, the estimation of P-values for the progressors versus non-progressors comparison from negative binomial count data is likely somewhat optimistic. The main analysis tool (edgeR) cannot account for the paired nature of the samples in our design with a mixed effects model. For the ageing related comparisons this was not a major issue. For the comparisons between progressor and non-progressor, it was because the use of fixed effects models for NB data is not well studied [9]. This may have caused some overoptimism in calling certain regions differentially methylated when combining first and second round samples.

In order to estimate the extent of overoptimism in rejecting null hypotheses for the attempted comparisons, we analysed the results of restricted permutations of the samples. These permutations could not determine to what extent the P-values needed to be adjusted. They did show that, for the majority of the comparisons, the P-values of the true sample ordering were consistently lower. This suggests that the observed effects were truly related to the phenotypes of interest. Alongside the ongoing technological advances in sequencing technology, there is a need to develop appropriate statistical methods that can cope with negative binomial data in complex experimental set-ups including several confounding factors and a subject specific term in the form of random effects.

For the validation experiments, we analysed the methylation percentages with a tool (methylKit) that was straightforward in terms of data (pre-)processing but was limited to simple two-group comparisons. This may have caused overoptimism in confirming differential methylation between progressors and non-progressors over both rounds combined. For the inter-round comparison the use of models with a subject specific term was not possible as these generally lead to biased effect size estimates [10]. Here to, more comprehensive analysis methods and flexible tools are required to model our experimental design in its entirety.

In the course of analysing the validation data, it became apparent that most loci showed minimal differences in methylation degree. For some loci the difference was even significant but opposite to the results of the genome-wide screening (hypomethylation in screening and hypermethylation in validation or visa versa). There are several factors that might account for these observations.

First, the different nature of both technologies used to assess methylation is difficult to account for. Although bisulfite sequencing offers an unambiguous
methylmethylation degree at nucleotide (cytosine) based resolution, the technology is susceptible to PCR bias introduced by the methylation degree it attempts to measure [11]. Methylation enrichment sequencing on the other hand, offers a cost-efficient means of scanning an entire genome but only provides a relative methylation measure biased by CpG and GC content and copy number variation [12]. But contrary to the bisulphite sequencing results, these biases are less relevant when comparing the same locus between (non-tumour) samples, as is the case in this doctoral thesis.

Second, MethylCap-seq affinity for a region is likely influenced by the number of neighbouring methylated CpGs and it is currently unknown how methylation changes in neighbouring CpGs affect binding affinity. For example, a change in methylation status of one or two cytosines in a larger series of adjacent CpGs in a single molecule of DNA implies only a modest quantitative methylation difference (as observed during validation), but may already have a major impact on the (more qualitative) affinity of the methyl-binding domain used. Note that these methyl-binding domains have a biological origin and that their binding affinity may better reflect the biological relevance than quantitative bisulphite sequencing data. In several cases during validation, significant effects were observed opposite to what was predicted. As these effects were present in both directions (i.e. both hyper- and hypomethylation) and were also significant when using other statistical methodologies such as the independent samples t-test (data not shown) and samples were not grouped by phenotype during experimental protocols, it is unlikely that these effects were caused by biases introduced during MethylCap-seq and/or bisulphite sequencing. Moreover, during validation it occurred that neighbouring CpGs were both significantly associated with the phenotype, yet one in the expected direction and one in the opposite direction. However, this observation can be biologically meaningful, as the impact of methylation is very CpG location prone, with effects even depending on the methylation status of single CpGs [13]. As the net-impact of adjacent but opposite methylation events on MethylCap affinity is a priori unclear, additional research on the obtained validation data as well is absolutely necessary to pinpoint the relevance of the individual effects observed and to resolve any remaining ambiguity.

Last, although we chose PBL leukocyte DNA for its accessibility, it is important to remember that it is composed of a complex mix of cell types, each with its own specific methylation profile [14]. It is therefore straightforward to hypothesise that the observed small differences are in fact large differences of an individual cell types, e.g. of monocytes or basophils which typically constitute less than 8% or 1% of all leukocytes respectively (NIH reference values
from http://www.nlm.nih.gov/medlineplus/ency/article/003657.htm) or even monocyte subsets. If this is indeed the case, significant differences will be obscured by the methylation status of the other cell types, thereby introducing extra noise (but not bias) in the measurements. Next to yielding additional biological insight, the identification of the relevant cell type may therefore also lead to more discriminative biomarkers with potential for clinical application. Current innovation in technology may even lead to a scenario where such tests are still applicable on whole blood leukocyte DNA if combined with a computational separation of leukocyte sub-populations (see section 5.5 Perspectives).

These and other technical challenges notwithstanding we were able to identify subtle but consistently differentially methylated loci and individual CpGs. In addition to the loci associated with more advanced atherosclerosis, we also identified 13 loci which were predictive of atherosclerosis progression in the subsequent decade. Although we confirmed two of these loci in the validation arm of this study, the number of samples was too small to perform a rigorous evaluation of the predictive power of these potential biomarkers.

5.4 Conclusion

Cardiovascular ageing is an intricate phenomenon with both pre-determined and modifiable risk factors. Approaching this subject from an epigenetic perspective creates an opportunity to incorporate this complexity. Although we did not dispose over telomere data from the second round, chapter 2 clearly showed that epigenetics are likely involved early on in cardiovascular ageing and that longitudinal studies are required to separate cause and effect.

In the conception of our experimental design we intended to study atherosclerosis progression primarily (chapter 3) and used an orthogonal analysis of the available data to examine ageing-related methylation changes independently (chapter 4). Despite some challenges during validation, we identified two non-overlapping sets of loci for which the methylation state was associated with atherosclerosis or ageing, respectively. However, when we summarised the findings for atherosclerosis progression and ageing to a more functional level, i.e. biological pathways, we discovered a near complete overlap in potentially differentially methylated pathways despite clearly distinct loci being altered in both processes. Our combined findings suggest
that, at least as far as DNA methylation of peripheral blood leukocytes is concerned, there is great functional similarity between ageing and atherosclerosis progression. This hypothesis is very intuitive as having poor cardiovascular health is not unlike being physically older despite having the same calendar age. More research is required to characterise this functional similarity and to determine whether it is reflected in other cellular mechanisms such as gene expression or protein interactions.

5.5 Perspectives

Epigenetics is almost per definition featured by versatility, leading to the notion that any assessment of an epigenetic property will only provide a snapshot of the more complex underlying processes. Longitudinal studies, such as the Asklepios study on successful cardiovascular ageing, therefore provide an ideal setting to study epigenetics. Several current limitations are expected to be remediated during the coming rounds of the study. Furthermore, technological advances will enable us to assess epigenetic features in an increasingly more comprehensive (genome-wide), cheaper, faster and more accurate manner.

Though telomere length has often been independently associated with cardiovascular ageing, and causality has been proposed, only a few studies assessed telomere length in a longitudinal fashion and once the second round is completed, the Asklepios Study will be one of these. Next to the fact that telomere attrition rates may be better linked with cardiovascular ageing than baseline telomere length, a major advantage of the Asklepios Study is the fact that all subjects were apparently healthy at inclusion. For the results from the first round of the study, as presented in chapter 2, this implies that intermediate endpoints at most can be used for cross-sectional analysis. However, as the ageing process in the study population continues, harder endpoints will inevitably occur and will enable us to evaluate to what extent baseline telomere length (and/or telomere attrition) is predictive of clinically relevant systolic, diastolic and vascular (dys)function. This will already be the case for the second round results, which is currently ongoing and will yield an overview of the different cardiovascular parameters for analyses in the near future.

Our current results likely contain a mix of causal, associated and spurious methylation changes and further experiments will be required to determine the
nature of each. One approach to showing causality would be to wait for successive rounds of the Asklepios Study. By observing more volunteers in several rounds, it might be possible to determine which methylation changes occur before the alterations in phenotype. For this type of experiment however, it would be more interesting to follow a much smaller cohort at more frequent intervals. Unrelated methylation changes are expected to be more stochastic in nature, whereas associated or causative epimutations would likely be present more consistently from their initial occurrence onward.

For the atherosclerosis progression markers the strongest proof of causality is an experimental set-up where one can randomly assign near identical subjects to a treatment and a control group. Although ethical considerations would prohibit any experimental design that actively encourages atherosclerosis development in humans, animal studies could be used as a surrogate. Analogously to the very successful CRISPR-Cas technique for genome editing [15], several groups are working on techniques which allow targeted (de)methylation to engineer the desired methylation status [16,17]. Obviously one cannot assign a study group to a non-ageing category with regards to the ageing associated markers but one could study the occurrence ”rejuvenating” effects upon methylation pattern alteration.

Moreover, ongoing technological advances allow us to infer several features of future studies on the topic. More specifically, lower cost of sequencing should will enable population wide DNA methylation studies leading to the epigenetic equivalent of genome-wide association studies [18]. Single molecule real-time sequencing technology entails the promise of direct assessment of not only the DNA sequence and its methylation status (thereby avoiding the often tricky bisulphite conversion step) but also of other nucleic acid based epigenetic features such as DNA hydroxymethylation [19]. Single molecule sequencing technologies typically offer far longer reads (up to tens of kbps), which could offer the possibility to assess the full length telomere directly and also entails a far easier (computational) discrimination of cell types. Indeed, given sufficient read length, phasing of methylation status between different genomic position will be possible [19]. This phasing will allow one to pinpoint epigenetic alterations associated with the phenotype of interest to specific cell type sub-populations.

The data currently available to us, did not enable us to do this but the Asklepios samples remain available for follow-up research. Although lesser quantitative shifts in leukocyte subtypes are expected for atherosclerosis than for ageing as such, it is still of major interest to identify the specific cell
type(s) involved. A first validation indicated that several DNA methylation changes were indeed present, but the next step should be to identify the cell type(s) driving the observed differences.

As CpG level information was provided by the validation study, this enables the development of relatively inexpensive methylation-specific quantitative PCR assays. The latter also require limited input material and can therefore be used on leukocyte sub-populations, generated by e.g. fluorescence-activated cell sorting, in a novel case-control study for atherosclerosis or longitudinal study of ageing. Using this extra knowledge, it can subsequently be decided to perform a large scale validation study on solely the relevant fraction, or — if the other leukocyte sub-populations have no major impact — on whole blood leukocytes as a proxy for the specific fraction involved. This will enable the identification of the determinants of the relevant epigenetic modifications but also to assess to what extent these biomarkers/risk factors are relevant in the general population. Once the optimal methylation markers have been established this methodology could also be translated to a cost-efficient diagnostic test.
References


Cardiovascular ageing is a complex phenomenon that is affected by a host of internal and external factors. These factors are both pre-determined and changeable throughout a person’s life span. Epigenetics is the information layer superimposed on the genetic information encoded in DNA. By definition it reflects phenotypic variation that is changeable over time, adaptive to environmental circumstances and at least partially heritable independent of the underlying DNA sequence. In this thesis we studied the role of epigenetics in cardiovascular ageing.

The study samples were all drawn from the Asklepios longitudinal study on (successful) cardiovascular ageing, which included over 2500 volunteers aged approximately 35 to 55 years and apparently healthy when the study started in 2002–2004. The first follow-up round was initiated in 2012 and is currently still ongoing. We made use of first round and already available second round samples as well as phenotypic information to focus both on hemodynamic parameters and atherosclerosis, the gradual calcification and restriction of arteries that is closely linked with cardiovascular disease.

The length dynamics of peripheral blood leukocyte telomeres, the protective ends of the chromosomes, have been intensely studied both in the
Aklepios and other populations. We found that many parameters of cardiovascular function which had previously been described to be associated with telomere length, were not so in this age range. We observed that telomere length was related to sensitive and clinically relevant indicators of diastolic function, thus suggesting that telomeres, which can also be regarded as epigenetic traits, are involved early on in certain cardiovascular ageing processes. Follow-up research is required to determine how telomere length has changed in the Asklepios Study and what the implications are for the associations with diastolic function.

We also applied second generation sequencing technology which allowed us to profile multiple samples in a genome-wide way with regards to cytosine DNA methylation. Through a carefully considered study design and despite some limitations, we identified distinct methylation patterns associated with atherosclerosis progression and with ageing between study rounds. Summarising these results to the level of biological pathways, we found support for the hypothesis that ageing and atherosclerosis progression are functionally similar from the perspective of leukocytes.

We also found methylation difference which preceded atherosclerosis progression in the following decade. For the time being it is too soon to extrapolate these results to the general population. More advanced study designs, possibly using the fast-approaching single-molecule sequencers, will need to distinguish which leukocyte subtypes are characterised by the methylation changes we observed.
Cardiovasculaire veroudering is een complex fenomeen dat door een reeks intrinsieke en extrinsieke factoren beïnvloed wordt. Sommige van deze zijn bepaald bij de conceptie, andere variëren gedurende het leven. Epigenetica, een verzameling mechanismen die fenotypische informatie coderen bovenop het DNA, is per definitie veranderlijk in de tijd, adaptief aan omstandigheden en ten minste gedeeltelijk erfelijk onafhankelijk van het onderliggende DNA. Dit proefschrift handelt over de rol van epigenetica in cardiovasculaire veroudering.


De gemiddelde lengte van telomeren, de beschermende uiteindes van de chromosomen, werd reeds uitvoerig bestudeerd in de Asklepios populatie maar ook in andere studies. Dit onderzoek heeft uitgewezen dat de meeste parameters van cardiovasculaire functie die in het verleden geassocieerd werden met telomeerlengte, niet significant waren in onze relatief jonge populatie. Wel werd geobserveerd dat telomeerlengte gecorreleerd was met gevoelige, klinisch relevante parameters van diastolische functie. Dit suggereert dat telomeerlengte, welke ook beschouwd kan worden als een epigenetisch mechanisme, reeds vroeg betrokken is bij bepaalde aspecten van cardiovasculaire veroudering. Toekomstig onderzoek moet uitwijzen hoe telomeerlengte verandert in de Asklepios studiepopulatie en wat dit betekent voor de associaties met diastolische functie.

Verder werd tweede-generatie sequencerings technologie gebruikt om verschillende stalen genoomwijd te profileren op het vlak van DNA cytosinemethylatie. Door middel van een weloverwogen studieontwerp gebaseerd op de gedetailleerd gezondheidsinformatie in de Asklepios Studie en ondanks een aantal limitaties, waren we in staat om onafhankelijke wijzigingen in DNA methylatie te detecteren die geassocieerd waren met veroudering tussen de studierondes of de progressie van atherosclerose. Wanneer deze resultaten echter geaggregeerd werden tot het niveau van biologische gennetwerken,
vonden we aanwijzingen dat veroudering en atheroscleroseprogressie functioneel gelijkwaardig zijn vanuit het perspectief van leukocyten.

Bovendien werden in de eerste ronde methylatieverschillen waargenomen die voorafgingen aan atheroscleroseprogressie. Op dit moment kunnen deze resultaten echter nog niet extrapoleerd worden naar de algemene bevolking. Meer geavanceerde studieontwerpen, die mogelijk gebruik maken van de komende generatie technologieën voor de sequentiëring van individuele DNA moleculen, zullen nodig zijn om te bepalen welke leukocytsubtypes bepalend zijn voor de geobserveerde methylatieverschillen.
Appendix
A.1 Chapter 2

### Load data set ###

```r
load("DyastolicDysfunction_BNP.RData")
```

### Scale to SI units ###

```r
DATA$TDE <- DATA$TDE - 0.1
Females$TDE <- Females$TDE - 0.1
Males$TDE <- Males$TDE - 0.1
DATA$TDA <- DATA$TDA - 0.1
Females$TDA <- Females$TDA - 0.1
Males$TDA <- Males$TDA - 0.1
DATA$TDSMAX <- DATA$TDSMAX - 0.1
Females$TDSMAX <- Females$TDSMAX - 0.1
Males$TDSMAX <- Males$TDSMAX - 0.1
```

### Summary measures ###

```r
length(Males$Pulse_pressure_mmHg[is.na(Males$Pulse_pressure_mmHg)==TRUE])
mean(Males$Pulse_pressure_mmHg, na.rm=TRUE)
```
\begin{verbatim}
### Define helper function for 2 group comparison
autoCompare2 <- function (Group1, Group2) {
  require(car)
  Group1 <- as.numeric(Group1)
  Group2 <- as.numeric(Group2)
  parametric = "TRUE"
  eqVar = "TRUE"
  mean1 <- mean(Group1)
  mean2 <- mean(Group2)
  var1 <- var(Group1)
  var2 <- var(Group2)
}
\end{verbatim}
norm1 <- shapiro.test(Group1)$p.value
norm2 <- shapiro.test(Group2)$p.value
if (norm1 < 0.05 | norm2 < 0.05) {
    message("At least one group does not fit the normal distribution.")
    parametric = "FALSE"
}
TEMP <- matrix(0, nrow = (length(Group1) + length(Group2)), ncol = 2)
colnames(TEMP) <- c("measurements", "Factor")
TEMP <- as.data.frame(TEMP)
TEMP[1:length(Group1), 1] <- as.double(Group1)
TEMP[(length(Group1) + 1):dim(TEMP)[1], 1] <- as.double(Group2)
TEMP[1:length(Group1), 2] <- rep("A", length(Group1))
TEMP[(length(Group1) + 1):dim(TEMP)[1], 2] <- rep("B", length(Group2))
TEMP[2, ] <- as.factor(TEMP[, 2])
if (leveneTest(measurements ~ Factor, data = TEMP) [1, 3] < 0.05) {
    eqVar = "FALSE"
}
rm(TEMP)
if (length(Group1) >= 30 & length(Group1) >= 30) {
    parametric = "TRUE"
    message("Sufficient observations for parametric testing.
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
"
```r
if (parametric=="TRUE" & eqvar=="FALSE"){
    message("Performed two-sided Welch t-test on normally distribution, non-homoscedastic.")
    message(cat("Estimated variance ratio = ", var.test(Group1, Group2, ratio = 1, alternative="two.sided"), conf.level = 0.95)$estimate, 
        \n" ,sep="")
    return(t.test(Group1, Group2, alternative = "two.sided", mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95))
}
if (parametric=="FALSE"){
    message("Performed two-sided Wilcoxon rank sum test on non-normally distribution data.")
    return(wilcox.test(Group1, Group2, alternative = "two.sided", mu = 0, paired = FALSE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95))
}
}
autoCompare2(Females$Pulse_pressure_mmHg,Males$Pulse_pressure_mmHg)

autoCompare2(Females$sphericiNDBF2,Males$sphericiNDBF2)
```

#### Export model results as table ######

`# Model 1` ######

```r
A<-rbind(
    t(summary(glm(TMEAR ~ meanTRF, data=DATA)))$coefficients[2,c(1,4)],
    t(summary(glm(TDEAR ~ meanTRF, data=DATA)))$coefficients[2,c(1,4)],
    t(summary(glm(TME ~ meanTRF, data=DATA)))$coefficients[2,c(1,4)],
)```
```r
# Model 1
B <- rbind(
  t(summary(glm(TMA ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDE ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDA ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(EoverEm ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TMDE ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TMDD ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDE ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDA ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(EF ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDSMAX ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]))

##### Model 2 ######
B <- rbind(
  t(summary(glm(TMEAR ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDEAR ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TME ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TMAD ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDE ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDAD ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(EoverEm ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TMDE ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TMDD ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDE ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDA ~ meanTRF + Age_years + Gender, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(EF ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]),
  t(summary(glm(TDSMAX ~ meanTRF, data=DATA))$coefficients[2, c(1, 4)]))
```
$coefficients[2, c(1, 4)]$,
\[ t(\text{summary(glm(IVRT} \sim \text{meanTRF + Age\_years + Gender} \rightarrow, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(log(NTproBNPpgml, 10) \sim \text{meanTRF + Age\_years + Gender}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(alloLVM} \sim \text{meanTRF + Age\_years + Gender, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(LVEDD2D} \sim \text{meanTRF + Age\_years + Gender, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(EF} \sim \text{meanTRF + Age\_years + Gender, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TDSMAX} \sim \text{meanTRF + Age\_years + Gender, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ]

#### Model 3 ####
\[
C<-\text{rbind(}
\[ t(\text{summary(glm(TMEAR} \sim \text{meanTRF + Age\_years + Gender + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TDEAR} \sim \text{meanTRF + Age\_years + Gender + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TME} \sim \text{meanTRF + Age\_years + Gender} \rightarrow + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index} \rightarrow \text{kgm2 + Drugtreated\_hypertension}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TMA} \sim \text{meanTRF + Age\_years + Gender} \rightarrow + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index} \rightarrow \text{kgm2 + Drugtreated\_hypertension}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TDE} \sim \text{meanTRF + Age\_years + Gender} \rightarrow + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index} \rightarrow \text{kgm2 + Drugtreated\_hypertension}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ] ,
\[ t(\text{summary(glm(TDA} \sim \text{meanTRF + Age\_years + Gender} \rightarrow + \text{Diastolic\_BP\_mmHg + TDHR + Body\_Mass\_Index} \rightarrow \text{kgm2 + Drugtreated\_hypertension}, \text{data}=\text{DATA})) \text{coefficients}[2, c(1, 4)]) ]
\]
\( \) $coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(EoverEm \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{TMDT} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{IVRT} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\log(\text{NTproBNPpgml}, 10) \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{alloLVM} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{LVEDD2D} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{EF} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$, 
\( t(\text{summary(glm}(\text{TDSMAX} \sim \text{meanTRF} + \text{Age}\_\text{years} + \text{Gender} + \text{Diastolic}\_\text{BP}\_\text{mmHg} + \text{TDHR} + \text{Body}\_\text{Mass}\_\text{Index}\_\text{kgm2} + \text{Drugtreated}\_\text{hypertension}, \text{data}=\text{DATA}))$coefficients[2, c(1, 4)]$)

AA<-cbind(A[1,1],A[1,2])
for (i in 2:dim(A)[1]) {AA<-cbind(AA,A[i,1],A[i,2])}
BB<-cbind(B[1,1],B[1,2])
for (i in 2:dim(B)[1]) {BB<-cbind(BB,B[i,1],B[i,2])}
CC <- cbind(C[1,1], C[1,2])
for (i in 2:dim(C)[1]) {CC <- cbind(CC, C[i, 1], C[i, 2])}
OUT <- rbind(AA, BB, CC)
rownames(OUT) <- c("Model1", "Model2", "Model3")
#colnames(OUT) <- c("TMMEAR", "TME", "TMA", "TDE", "TDA","
																TMDT", "IVRT", "NTproBNPpgml", "alloLVM",""
																LVEDD2D", "EF", "S \' (TDSET)")
OUT <- t(OUT)
options(digits = 22, scipen = 22)
write.table(OUT, "ImportTableModels.v4.0_SBPandDBP.
																	txt", sep = "\t", dec = ",", row.names = FALSE, col.
																	names = TRUE, quote = FALSE)

### Proportion of variance explained ###
AB <- glm(TMMEAR ~ meanTRF + Age_years + Gender +
																	Systolic_BP_mmHg + TDHR + Body_Mass_Index_+
																	kgm2 + Drugtreated_hypertension, data=DATA)
anova(AB)
summary(AB)
aov(AB)
(240.76 - 136.70) / 240.76
6.364 / 240.76

### Kuznetsova 2010 & Vasan 2011 ###
summary(glm(alloLVM ~ meanTRF + Age_years + Gender +
																	Systolic_BP_mmHg + Height_cm + Weight_kg +
																	Drugtreated_hypertension, data=DATA))
summary(glm(LV_Mass_2D_g ~ meanTRF + Age_years +
																	Gender + Systolic_BP_mmHg + Height_cm +
																	Weight_kg + Drugtreated_hypertension, data=
																	DATA))

summary(glm(alloLVM ~ meanTRF + Age_years + Gender +
																	Systolic_BP_mmHg + TDHR + Body_Mass_Index_+
																	kgm2 + Drugtreated_hypertension, data=DATA))
summary(glm(LV_Mass_2D_g ~ meanTRF + Age_years +
																	Gender + Systolic_BP_mmHg + TDHR + Body_Mass_+
																	Index_kgm2 + Drugtreated_hypertension, data=
																	DATA))
summary(glm(alloLVM ~ meanTRF + Age_years + Gender
\[ \text{TL as dependent variable} \]

```
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + IVRT, data=DATA))
summary(glm(meanTRF ~ Age_years + agefatherbirth +
            IVRT, data=Females))
summary(glm(meanTRF ~ Age_years + agefatherbirth +
            data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + Systolic_BP_mmHg, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + Drugtreated_hypertension, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + alloLVM, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + LV_Mass_2D_g, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + LVEDD2D, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + EF, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + TDSMAX, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + PWVcar_fem, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + Pulse_pressure_mmHg, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + Ee, data=DATA))
summary(glm(meanTRF ~ Gender + Age_years +
            agefatherbirth + crphigh, data=DATA))
```
\[ \text{agefatherbirth} + \text{il6above1.5} \]

**summary**(\text{glm}(\text{meanTRF} \sim \text{Gender} + \text{Age\_years} +
\text{agefatherbirth} + \text{Uric\_acid\_mgdl}, \text{data}=\text{DATA}))

**summary**(\text{glm}(\text{meanTRF} \sim \text{Gender} + \text{Age\_years} +
\text{agefatherbirth} + \text{IVRT} + \text{crphigh}, \text{data}=\text{DATA}))

**summary**(\text{glm}(\text{meanTRF} \sim \text{Gender} + \text{Age\_years} +
\text{agefatherbirth} + \text{IVRT} + \text{il6above1.5}, \text{data}=\text{DATA}))

**summary**(\text{glm}(\text{meanTRF} \sim \text{Gender} + \text{Age\_years} +
\text{agefatherbirth} + \text{IVRT} + \text{Uric\_acid\_mgdl}, \text{data}=\text{DATA}))

### Plot figures ###

**library**(ggplot2)

\[
p \leftarrow \text{ggplot(} \text{data} = \text{DATA, aes(} \text{meanTRF, TMEAR, colour=} \text{Gender)}
\]

\[
p + \text{geom\_point(} \text{alpha=1/2, na.rm=TRUE, size=1.25)} +
\rightarrow \text{geom\_abline(} \text{intercept=0.70647805, slope}
\rightarrow =0.06956889, \text{size=0.5)} + \text{xlab(} \text{"Telomere Length}\text{\_(kbp)}\text{") } + \text{ylab(} \text{"E/A") } + \text{theme(} \text{text} = \text{element}
\rightarrow \text{\_text(} \text{size=8)}, \text{axis\_title.y = element\_text(}
\rightarrow \text{size = rel(1.3), angle = 90)) + \text{theme(} \text{axis\_title.x = element\_text(}
\rightarrow \text{size = rel(1.3), angle}
\rightarrow = 0)) + \text{theme(} \text{axis\_title.x = element\_text(}
\rightarrow \text{vjust =0.25)) + \text{theme(} \text{axis\_title.y = element\_text(}
\rightarrow \text{vjust =0.35)) + \text{theme(} \text{plot\_margin = unit(}
\rightarrow \text{c(rep(1,4)),"mm"}))
\]

**ggsave(} \text{"LTLvsTMEAR\_600dpi\_colourByGender\_tiff"},
\rightarrow \text{width = 4.2677, height = 3.2677, dpi=600,}
\rightarrow \text{compression=\"lzw\")}
\]

\[
p \leftarrow \text{ggplot(} \text{data} = \text{DATA, aes(} \text{meanTRF, IVRT)}
\]

\[
p + \text{geom\_point(} \text{alpha=1/3, colour=\"red", na.rm=TRUE,}
\rightarrow \text{size=1.25)} + \text{geom\_abline(} \text{intercept}
\rightarrow =107.094972, \text{slope=-2.695635, size=0.5)} + \text{xlab(}
\rightarrow \text{\"Telomere Length\_\_omega(kbp)\") } + \text{ylab(} \text{\"IVRT\_\_omega(ms)\")}
\rightarrow + \text{theme(} \text{text} = \text{element\_text(} \text{size=8), axis\_text(}
\rightarrow \text{vjust =0.35))}
\]

\[
p \leftarrow \text{ggplot(} \text{data} = \text{DATA, aes(} \text{meanTRF, IVRT)}
\]

\[
p + \text{geom\_point(} \text{alpha=1/3, colour=\"red", na.rm=TRUE,}
\rightarrow \text{size=1.25)} + \text{geom\_abline(} \text{intercept}
\rightarrow =107.094972, \text{slope=-2.695635, size=0.5)} + \text{xlab(}
\rightarrow \text{\"Telomere Length\_\_omega(kbp)\") } + \text{ylab(} \text{\"IVRT\_\_omega(ms)\")}
\rightarrow + \text{theme(} \text{text} = \text{element\_text(} \text{size=8), axis\_text(}
\rightarrow \text{vjust =0.35))}
\]
title.y = element_text(size = rel(1.3), angle = 90) + theme(axis.ticks.margin = unit(1, "mm")) + theme(axis.title.x = element_text(size = rel(1.3), angle = 0)) + theme(axis.title.y = element_text(vjust = 0.35)) + theme(plot.margin = unit(c(rep(1, 4)), "mm"))
ggsave("PBL_TL_vs_IVRT_600dpi.tiff", width = 3.2677, height = 3.2677, dpi = 600, compression = "lzw")
A.2 Chapter 3 & 4

A.2.1 MethylCap-seq Analyses

##### Differential Methylation Analysis Asklepios

MethylCap study part 1: time and health
without subject identity ######

### Read input data ###

MHMcounts <- read.table("..//MHM/MCap_Athero_MHM_
counts.txt", quote="", comment.char="", skip
=2, header=TRUE, sep="\t", stringsAsFactors=
FALSE)
totcounts <- read.table("..//MHM/MCap_Athero_MHM_
MappingProps.txt", comment.char="", header=TRUE
, row.names=1, sep="\t", stringsAsFactors=
FALSE)
totcounts [1 ,] <- as.numeric(as.character(totcounts
[1 ,]))

### Load analysis tool ###

library (edgeR)

### Retain technical replicates with highest
coverage ###

totcounts [c("AS1.20", "AS1.21", "AS2.29", "AS2.47")]
Sorted <- MHMcounts [ duplicated (MHMcounts$region_ID)==
FALSE, c(13:32, 34:40, 54, 42:53)]
rownames (Sorted) <- MHMcounts$region_ID [ duplicated (MHMcounts$region_ID)==FALSE]
Libsizes <- as.data.frame(totcounts [1, colnames(Sorted
)])
dim (MHMcounts)
dim (Sorted)

### Normalize counts ###

#NormFacts <- calcNormFactors (Sorted, method="TMM",
refColumn = NULL, logratioTrim = 0.3, sumTrim
= 0.05, doWeighting = TRUE, Acutoff = -1e
+
10)
NormFacts<-c(rep(1,40))
names(NormFacts)<-colnames(Sorted)

### edgeR GLM-like approach ###
Sorted<-Sorted[rowSums(Sorted)>10,]
dim(MHMcounts)
dim(Sorted)

ProgStatus<factor(c(rep("np",10),rep("p",10),rep("np",10),rep("p",10)))
ProgStatus<-relevel(ProgStatus,ref="np")
Round<factor(c(rep("t1",10),rep("t1",10),rep("t2","",10),rep("t2",10)))
Round<-relevel(Round,ref="t1")
roundProg<as.factor(paste(Round,ProgStatus,sep="_"))
design<model.matrix(~-1+roundProg)

A<Sys.time()
yNC<DEGList(counts=Sorted,group=c(rep(1,10),rep(2,10),rep(3,10),rep(4,10)),lib.size=unlist(Libsizes),norm.factors=NormFacts)
yNC<estimateGLMCommonDisp(yNC,design)
yNC<estimateGLMTrendedDisp(yNC,design)
yNC<estimateGLMTagwiseDisp(yNC,design)
fitNC<-glmFit(yNC,design)
B<Sys.time()
B-A

save.image("MCapAthero_time_and_health_MHM_LIBsizeNorm_RealLabels.RData")
#load("MCapAthero_time_and_health_MHM_LIBsizeNorm_RealLabels.RData")
#library(edgeR)

### define contrasts ###
contrasts=matrix(0,ncol=6,nrow=4)
rownames(contrasts)=colnames(design)
colnames(contrasts)=c("diff.p.np_t1","diff.p.np_t2","avgDiff.p.np","diff.t2.t1_np","diff.t2.t1_p"
contrasts[1:2,1]=c(-1,1)
contrasts[3:4,2]=c(-1,1)
contrasts[1:4,3]=c(-1,1,-1,1)/2
contrasts[c(1,3),4]=c(-1,1)
contrasts[c(2,4),5]=c(-1,1)
contrasts[1:4,6]=c(-1,-1,1,1)/2

c<−Sys.time()
lrtListProg=apply(contrasts[ ,1:3],2,function(x)
    glmLRT(fitNC,contrast=x))
names(lrtListProg)=colnames(contrasts)[1:3]
lrtListProg$"progSam"=glmLRT(fitNC,contrast=
    contrasts[ ,1:2])
lrtListTime=apply(contrasts[ ,4:6],2,function(x)
    glmLRT(fitNC,contrast=x))
names(lrtListProg)=colnames(contrasts)[4:6]
lrtListTime$"timeSam"=glmLRT(fitNC,contrast=
    contrasts[ ,4:5])
D<−Sys.time()
D−C
save.image("MCapAthero_time_and_health_MHM_
    LibsizeNorm_RealLabels.RData")

sessionInfo()
q(save="no")

load("MCapAthero_time_and_health_MHM_LibsizeNorm_
    RealLabels.RData")

library("edgeR")

names(lrtListProg)
topTags(lrtListProg[[4]])

### Print Group Results (both rounds) ###

TEMP<−topTags(lrtListProg[[3]],n=dim(lrtListProg
    [[3]]))[1]$table

TEMP<−TEMP[order(as.numeric(rownames(TEMP)))]

TEMP<−cbind(rownames(TEMP),TEMP)
colnames(TEMP)[1]<="region_ID"

TEMP<−merge(MHMcounts[,c(1:32,34:40,54,42:53)],TEMP
R-Code

\[
\text{by="region\_ID", all.x=TRUE)}
\]

\text{as.data.frame}(\text{TEMP})

\text{dim}(\text{TEMP[TEMP$FDR}<0.05 \& !\text{is.na}(\text{TEMP$FDR}<0.05),])

\text{options}(scipen=28)
\text{options}(digits=22)
\text{write.table}(\text{TEMP[TEMP$FDR}<0.05 \& !\text{is.na}(\text{TEMP$FDR}<0.05),], "Analysis\_MHM\_groupDiffs\_RealLabels.txt", sep="\t", row.names=FALSE, col.names=TRUE, na="", quote=FALSE)

### Print Time Results (both prog and non-prog unpaired) ###

\text{write.table}(\text{TEMP[TEMP$FDR}<0.10 \& !\text{is.na}(\text{TEMP$FDR}<0.10),], "Analysis\_MHM\_timeDiffs\_RealLabels_FDR0.10.txt", sep="\t", row.names=FALSE, col.names=TRUE, na="", quote=FALSE)

### Print Time Results (both prog and non-prog unpaired) ###

\text{write.table}(\text{TEMP[TEMP$FDR}<0.05 \& !\text{is.na}(\text{TEMP$FDR}<0.05),], "Analysis\_MHM\_groupDiffs\_RealLabels.txt", sep="\t", row.names=FALSE, col.names=TRUE, na="", quote=FALSE)
### Print Results Round 1 differences ###

```r
TEMP <- topTags(lrtListProg[[1]], n = dim(lrtListProg)[[1]][1])$table
TEMP <- TEMP[order(as.numeric(rownames(TEMP)))]
TEMP <- cbind(rownames(TEMP), TEMP)
colnames(TEMP)[1] <- "region_ID"
TEMP <- merge(MHMcounts[, c(1:32, 34:40, 54, 42:53)], TEMP, by = "region_ID", all.x = TRUE)
TEMP <- cbind(TEMP[, -(13:52)], round(t(t(TEMP[, 13:52])), / unlist(Libsizes)) * mean(unlist(Libsizes)))
TEMP <- as.data.frame(TEMP)
dim(TEMP[TEMP$FDR < 0.10 & !is.na(TEMP$FDR)], options(scipen = 28)
options(digits = 22)
write.table(TEMP[TEMP$FDR < 0.10 & !is.na(TEMP$FDR) < 0.10], , "Analysis_MHM_Round1Diffs_RealLabels" _FDR_ 0.10.txt" , sep = \"\t\", row.names = FALSE, col.
names = TRUE, na = \"\", quote = FALSE)
write.table(TEMP[TEMP$FDR < 0.05 & !is.na(TEMP$FDR) < 0.10] & !is.na(TEMP$Entrez.Gene.Id), 4], "InputList_R1_FDR0.10.txt" , sep = \"\t\", row.names = FALSE, col.names = TRUE, na = \"\", quote = FALSE)
```

### Print Results Round 2 differences ###

```r
TEMP <- topTags(lrtListProg[[2]], n = dim(lrtListProg)[[2]][1])$table
TEMP <- TEMP[order(as.numeric(rownames(TEMP)))]
TEMP <- cbind(rownames(TEMP), TEMP)
colnames(TEMP)[1] <- "region_ID"
TEMP <- merge(MHMcounts[, c(1:32, 34:40, 54, 42:53)], TEMP, by = "region_ID", all.x = TRUE)
TEMP <- cbind(TEMP[, -(13:52)], round(t(t(TEMP[, 13:52])), / unlist(Libsizes)) * mean(unlist(Libsizes)))
TEMP <- as.data.frame(TEMP)
dim(TEMP[TEMP$FDR < 0.05 & !is.na(TEMP$FDR)], options(scipen = 28)
options(digits = 22)
dim(TEMP[TEMP$FDR < 0.10 & !is.na(TEMP$FDR)], options(scipen = 28)
options(digits = 22)
```
R-C ODE A-17

```r
options(scipen=28)
options(digits=22)
write.table(TEMP[TEMP$FDR<0.10 & !is.na(TEMP$FDR  ↩<0.10) ,] ,"Analysis_MHM_Round2Diffs_RealLabels  ↩_FDR_0.10.txt",sep="\\t",row.names=FALSE, col.  ↩names=TRUE, na="",quote=FALSE)
write.table(TEMP[TEMP$FDR<0.10 & !is.na(TEMP$FDR  ↩<0.10)& !is.na(TEMP$Entrez.Gene.Id) ,4] ,"  ↩InputList_R2_FDR0.10.txt",sep="\\t",row.names=FALSE, col.names=TRUE, na="",quote=FALSE)

### Print Results Non−Progressors over time ###
TEMP<-topTags(lrtListTime[[1]],n=dim(lrtListTime  ↩[[1]])[1])$table
TEMP<-TEMP[order(as.numeric(rownames(TEMP)) ) ,]
TEMP<-cbind(rownames(TEMP),TEMP)
colnames(TEMP)[1]<="region_ID"
TEMP<-merge(MHMcounts[,c(1:32,34:40,54,42:53) ],TEMP  ↩,by="region_ID", all.x=TRUE)
TEMP<-cbind(TEMP[,-(13:52) ],round(t(t(TEMP[,13:52]  ↩)/unlist(Libsizes)) * mean(unlist(Libsizes))))
TEMP<-as.data.frame(TEMP)
TEMP<-TEMP[order(TEMP$PValue) ,]
dim(TEMP[TEMP$FDR<0.05 & !is.na(TEMP$FDR<0.05) ,]) #  ↩11142
dim(TEMP[TEMP$FDR<0.05 & !is.na(TEMP$FDR<0.05) &  ↩duplicated(TEMP$region_ID)==FALSE ,]) # 10520
options(scipen=28)
options(digits=22)
write.table(TEMP[TEMP$FDR<0.05 & !is.na(TEMP$FDR  ↩<0.05) ,] ,"Analysis_MHM_NonProgressorDiffs_  ↩RealLabels.txt",sep="\\t",row.names=FALSE, col.  ↩names=TRUE, na="",quote=FALSE)

### Print Results Progressors over time ###
TEMP<-topTags(lrtListTime[[2]],n=dim(lrtListTime  ↩[[2]])[1])$table
TEMP<-TEMP[order(as.numeric(rownames(TEMP)) ) ,]
TEMP<-cbind(rownames(TEMP),TEMP)
```
colnames(TMP) <- c("region_ID")

TEMP <- merge(MHMcounts[, c(1:32, 34:40, 54, 42:53)], TEMP, by = "region_ID", all.x = TRUE)

TEMP <- cbind(TEMP[, -(13:52)], round(t(t(TEMP[, 13:52])) / unlist(Libsizes)) * mean(unlist(Libsizes)))

TEMP <- as.data.frame(TEMP)

TEMP <- TEMP[order(TEMP$PValue),]

dim(TEMP[TEMP$FDR < 0.05 & !is.na(TEMP$FDR < 0.05),])

dim(TEMP[TEMP$FDR < 0.10 & !is.na(TEMP$FDR < 0.10),])

options(scipen = 28)

options(digits = 22)

# write.table(TEMP[TEMP$FDR < 0.05 & !is.na(TEMP$FDR < 0.05),], "Analysis_MHM_ProgressorDiffs_RealLabels.txt", sep = "\t", row.names = FALSE, col.names = TRUE, na = "", quote = FALSE)

write.table(TEMP[1:200,], "Analysis_MHM_ProgressorDiffs_RealLabels.txt", sep = "\t", row.names = FALSE, col.names = TRUE, na = "", quote = FALSE)

q(save = "no")
A.2.2 Permutation Progressor Status

### Differential Methylation Analysis Asklepios  ↪ MethylCap study part 1.b: progressor status  ↪ permutation approach ######

### Read input data ###

MHMcounts <- read.table("../MHM/MCap_Athero_MHM_ counts.txt", quote="", comment.char="", skip =2, header=TRUE, sep="\t", stringsAsFactors= FALSE)
totcounts <- read.table("../MHM/MCap_Athero_MHM_ MappingProps.txt", comment.char="", header=TRUE => , row.names=1, sep="\t", stringsAsFactors= FALSE)
totcounts[1,] <- as.numeric(as.character(totcounts[1,]))

### Load analysis tool ###

library(edgeR)

### Retain technical replicates with highest coverage ###

totcounts[c("AS1.20","AS1.21","AS2.29","AS2.47")]

Sorted <- MHMcounts[ duplicated(MHMcounts$region_ID)== FALSE, c(13:32,34:40,54,42:53)]

rownames(Sorted) <- MHMcounts$region_ID[duplicated( -> MHMcounts$region_ID)==FALSE]

Libsizes <- as.data.frame(totcounts[1, colnames(Sorted => )])
dim(MHMcounts)
dim(Sorted)

### Normalize counts ###

#NormFacts <- calcNormFactors(Sorted, method="TMM", refColumn = NULL, logratioTrim = 0.3, sumTrim = 0.05, doWeighting = TRUE, A cutoff = -1e+10)

NormFacts <- c(rep(1,40))
names(NormFacts)<-colnames(Sorted)

### edgeR GLM-like approach ###
Sorted<-Sorted[rowSums(Sorted) >= 10,]
dim(MHMcounts)
dim(Sorted)

ProgStatus <- factor(c(rep("np",10),rep("p",10),rep("np",10),rep("p",10)))
ProgStatus<-relevel(ProgStatus, ref="np")
Round <- factor(c(rep("t1",10),rep("t1",10),rep("t2",10),rep("t2",10)))
Round<-relevel(Round, ref="t1")
roundProg=as.factor(paste(Round,ProgStatus, sep="."
#design <- model.matrix(~ -1+roundProg)

# contrasten van interesse
contrasts=matrix(0, ncol=6, nrow=4)
rownames(contrasts)=colnames(design)
colnames(contrasts)=c("diff.p.np.t1","diff.p.np.t2","avgDiff.p.np","diff.t2.t1_np","diff.t2.t1_p","avgDiff.t2.t1")
contrasts[1:2,1]=c(-1,1)
contrasts[3:4,2]=c(-1,1)
contrasts[1:4,3]=c(-1,-1,-1)/2
contrasts[c(1,3),4]=c(-1,1)
contrasts[c(2,4),5]=c(-1,1)
contrasts[1:4,6]=c(-1,-1,1,1)/2

## Part 2: Permutation group membership once
## keeping half of them constant for maximum
## entropy ##
set.seed(9481547)

C<-Sys.time()
permuteSamps=sample(1:10,5)
permuteSamps
ProgStatusPerm<-ProgStatus
ProgStatusPerm[permuteSamps]<="p"
ProgStatusPerm[permuteSamps+10]<-"np"
ProgStatusPerm[21:40]<-ProgStatusPerm[1:20]
ProgStatusPerm
roundProgPerm=as.factor(paste(Round, ProgStatusPerm, sep="_"))
designProgPerm=model.matrix(~-1+roundProgPerm)
yProgPerm <- DGEList(counts=Sorted, norm.factors=NormFacts)
yProgPerm <- estimateGLMCommonDisp(yProgPerm, designProgPerm)
yProgPerm <- estimateGLMTrendedDisp(yProgPerm, designProgPerm)
yProgPerm <- estimateGLMTagwiseDisp(yProgPerm, designProgPerm)
fitProgPerm <- glmFit(yProgPerm, designProgPerm)

lrtListProg=apply(contrasts[,1:3],2,function(x)lmLRT(fitProgPerm, contrast=x))
names(lrtListProg)=colnames(contrasts)[1:3]
lrtListProg$"progSam"=glmLRT(fitProgPerm, contrast=contrasts[,1:2])
D<-Sys.time()

save.image("MCapAthero_time_and_health_MHM_"
LibsizeNorm_GroupPermuteHalf.RData")
message(paste("Completed group ( half) permutation in ",D-C," hours",sep=""))

sessionInfo()
q(save="no")
A.2.3 Plot Density Permutation Progressors

```r
library(ggplot2)
library(edgeR)

F <- Sys.time()
load("MCapAthero_time_and_health_MHM_LibsizeNorm_{RealLabels}.RData")
Real_npv.spaavg.TEMP <- topTags(lrtListProg[[3]], n = dim(lrtListProg[[3]])[1])$table
Real_npv.t1.TEMP <- topTags(lrtListProg[[1]], n = dim(lrtListProg[[1]])[1])$table
Real_npv.t2.TEMP <- topTags(lrtListProg[[2]], n = dim(lrtListProg[[2]])[1])$table
load("MCapAthero_time_and_health_MHM_LibsizeNorm_{GroupPermuteHalf}.RData")
G <- Sys.time()
G-F

### Both Rounds ###
TEMP <- topTags(lrtListProg[[3]], n = dim(lrtListProg[[3]])[1])$table
dG <- ggplot(TEMP, aes(x=PValue)) + geom_density() + ylim(0, 3)# + scale_x_log10()
ggsave("Density_GroupDiffs_halfPermutation.pdf", width=7, height=7)
dG <- ggplot(Real npv.s.p.avg.TEMP, aes(x=PValue)) +
  geom_density() + ylim(0, 3)# + scale_x_log10()
ggsave("Density_GroupDiffs_RealLabels.pdf", width=7, height=7)
rm(dG)

### Round 1 ###
TEMP <- topTags(lrtListProg[[1]], n = dim(lrtListProg[[1]])[1])$table
dG <- ggplot(TEMP, aes(x=PValue)) + geom_density() + ylim(0, 3)# + scale_x_log10()
ggsave("Density_Round1Diffs_HalfPermutation.pdf", width=7, height=7)
```

---

A-22 APPENDIX A

---
dG <- ggplot(Real_np_vs_p.t1.TEMP, (aes(x=PValue))) + geom_density() + ylim(0, 3) #+ scale_x_log10() ggsave("Density_Round1Diffs_RealLabels.pdf", width =7, height=7) rm(dG)

### Round 2 ###
TEMP <- topTags(lrtListProg [[2]], n=dim(lrtListProg [[2]])[1])$table
dG <- ggplot(TEMP, (aes(x=PValue))) + geom_density() + ylim(0, 3) #+ scale_x_log10() ggsave("Density_Round2Diffs_HalfPermutation.pdf", width=7,height=7) dG <- ggplot(Real_np_vs_p.t1.TEMP, (aes(x=PValue))) + geom_density() + ylim(0, 3) #+ scale_x_log10() ggsave("Density_Round2Diffs_RealLabels.pdf", width =7,height=7) rm(dG)

q(save="no")
A.2.4 Paired Time Analysis

##### Differential Methylation Analysis Asklepios

→ MCap study part 3.a: time paired (basic model → )

### Read input data ###

#ZillerCounts <- read.table("./Ziller_DMR/MCapAthero_summarizedReads_DMRs_noDups_PEcorrection.txt",
→ comment.char="", header=TRUE, sep="\t",
→ stringsAsFactors=FALSE)

MHMcounts <- read.table("./MCap_Athero_MHM.txt",
→ quote="", comment.char="", skip=2,
→ header=TRUE, sep="\t", stringsAsFactors=FALSE)

totcounts <- read.table("./MCap_Athero_MHM_MappingProps.txt",
→ comment.char="", header=TRUE, row.names=1, sep="\t", stringsAsFactors=FALSE)

nonDupCounts <- read.table("./MCap_Asklepios_nonDup_MappingProps.txt",
→ comment.char="", header=TRUE, sep="\t", stringsAsFactors=FALSE,
→ colClasses=c("character","integer"))

totcounts[1,] <- as.numeric(as.character(totcounts[1,]))

### Load analysis tool ###

library(edgeR)

### Retain technical replicates with highest coverage ###

totcounts[c("AS1.20","AS1.21","AS2.29","AS2.47")]

Sorted <- MHMcounts[duplicated(MHMcounts$region_ID) == FALSE, c(13:32,34:40,54,42:53)]

rownames(Sorted) <- MHMcounts$region_ID[duplicated(MHMcounts$region_ID) == FALSE]

Libsizes <- as.data.frame(totcounts[1, colnames(Sorted)])

dim(MHMcounts)
dim(Sorted)
```r
### Normalize counts ###
NormFacts <- calcNormFactors(Sorted, method="TMM",
   refColumn = NULL, logRatioTrim = 0.3, sumTrim = 0.05, doWeighting = TRUE, Acutoff = -1e+10)
NormFacts <- c(rep(1,40))
names(NormFacts) <- colnames(Sorted)

### edgeR GLM-like approach ###
Sorted <- Sorted[rowSums(Sorted) >= 10,]
dim(MHMcounts) = dim(Sorted)

FactIndividual <- factor(c(1:20,1:20))
FactRound <- factor(c(rep("First",20),rep("Second",20)))
FactRound <- relevel(FactRound, ref="First")
design <- model.matrix(~ -1 + FactIndividual + FactRound)

A <- Sys.time()
y <- DGEList(counts=Sorted, lib.size = unlist(Libsizes), norm.factors = NormFacts)
y <- estimateGLMCommonDisp(y, design)
y <- estimateGLMTrendedDisp(y, design)
y <- estimateGLMTagwiseDisp(y, design)
fit <- glmFit(y, design)
lrt <- glmLRT(fit)
B <- Sys.time()
B - A
save.image("MCapAthero_TimePaired_MHM_LibsizeNorm.RData")
```
### Read input data ###

\[
\text{MHMcounts}\leftarrow \text{read.table}("..//MHM/MCap_Athero_MHM\_counts.txt", quote="", comment.char="", skip = 2, header=TRUE, sep="\t", stringsAsFactors=FALSE)
\]

\[
\text{totcounts}\leftarrow \text{read.table}("..//MHM/MCap_Athero_MHM\_MappingProps.txt", comment.char="", header=TRUE, row.names=1, sep="\t", stringsAsFactors=FALSE)
\]

\[
\text{totcounts}[1,]\leftarrow \text{as.numeric(\text{as.character(\text{totcounts}[1,])})}
\]

### Load analysis tool ###

\[
\text{library(edgeR)}
\]

### Retain technical replicates with highest coverage ###

\[
\text{totcounts}[\text{c("AS1.20","AS1.21","AS2.29","AS2.47")}]
\]

\[
\text{Sorted}\leftarrow \text{MHMcounts[\text{duplicated(MHMcounts$region\_ID)}==FALSE, c(13:32,34:40,54,42:53)]}
\]

\[
\text{rownames(\text{Sorted})}\leftarrow \text{MHMcounts$region\_ID[\text{duplicated}()}==FALSE]
\]

\[
\text{Libsizes}\leftarrow \text{as.data.frame(\text{totcounts}[1,\text{colnames(\text{Sorted})}]})
\]

\[
\text{dim(MHMcounts)}\quad \text{dim(\text{Sorted})}
\]

### Normalize counts: TMM not performed based on performance in technical replicates ###

\[
\text{#NormFacts}\leftarrow \text{calcNormFactors(\text{Sorted}, method="TMM", refColumn = NULL, logratioTrim = 0.3, sumTrim = 0.05, doWeighting = TRUE, A cutoff = \text{-1e+10})}
\]

\[
\text{NormFacts}\leftarrow \text{c(\text{rep}(1,40))}
\]

\[
\text{names(\text{NormFacts})}\leftarrow \text{colnames(\text{Sorted})}
\]

### edgeR GLM–like approach ###

\[
\text{Sorted}\leftarrow \text{Sorted[\text{rowSums(\text{Sorted})}>10,]}
\]
dim(MHMcounts)
dim(Sorted)

FactIndividual <- factor(c(1:20,1:20))
ProgStatus <- factor(c(rep("np",10),rep("p",10),rep("np",10),rep("p",10))
ProgStatus<- relevel(ProgStatus, ref="np")
Round <- factor(c(rep("t1",10),rep("t1",10),rep("t2" ,10),rep("t2",10))
Round<-relevel(Round, ref="t1")

set.seed(1364)
RoundPerm<-Round
ShuffleNonProg<-sample(1:10,5)
ShuffleNonProg
ShuffleProg<-sample(11:20,5)
ShuffleProg
RoundPerm[c(ShuffleNonProg,ShuffleProg)]<-"t2"
RoundPerm[c(ShuffleNonProg,ShuffleProg)+20]<-"t1"
RoundPerm

FactRPShuffled<- as.character(paste(ProgStatus ,
                  RoundPerm ,sep=" "))
FactRPShuffled[FactRPShuffled=="np_t1"]<-"t1"
FactRPShuffled[FactRPShuffled=="p_t1"]<-"t1"
FactRPShuffled<-factor(FactRPShuffled , levels=c("t1"             
                  ","np_t2","p_t2"))
FactRPShuffled

design <- model.matrix(~ -1 + FactRPShuffled +
                  FactIndividual)

A<-Sys.time()
y <- DGEList(counts=Sorted , lib.size=unlist(
                  Libsizes ) , norm.factors=NormFacts)
y <- estimateGLMCommonDisp(y,design)
y <- estimateGLMTrendedDisp(y,design)
y <- estimateGLMTagwiseDisp(y,design)
fit <- glmFit(y,design)
B <- Sys.time()
B-A
save.image("MCapAthero_TimePaired_MHM_LibsizeNorm_→WithProgStatus_PermuteHalf.RData")

sessionInfo()
q(save="no")
### Analyse validation experiment ###
### Load libraries ###

```r
library(GenomicRanges)
library(methylKit)
```

### Read data ###

```r
Samples <- read.table("SampleTable.txt", sep="\t",
                      comment.char="", header=FALSE, colClasses=c(rep("character",3)))
Index <- read.table("SampleDescriptionKey.txt", sep="\t",
                   comment.char="", header=TRUE)
DATA <- read.table("BS_table_H2G2.txt", sep="\t",
                   comment.char="", header=TRUE)
```

```r
colnames(Samples) <- c("Sample","Read1","Read2")
Samples$SampleOrder <- substr(Samples$Sample,5,9)
Samples <- Samples[!dim(Samples)[1],]
Samples <- Samples[order(as.numeric(Samples$SampleOrder)),]
```

```r
colnames(DATA)[7:94] <- paste("ABS", Samples$SampleOrder, "met", sep="")
colnames(DATA)[95:182] <- paste("ABS", Samples$SampleOrder, "cov", sep="")
dim(DATA)
```

### define regions targeted by primer pairs ###

```r
regions <- GRanges(seqnames=c("3","4","6","6","14","15","22","2","2","3","11","4","4","4","4","6","6","8","12","15","7","8","8","10"),
                   IRanges(start=c(186262961,152443952,83420357,158699387,50687972,2,2,3,11,"4","4","4","4","4","6","6","8","12","15","7","8","8","10"),
                   end=c(186263187,152444259,83420808,158699681,50688193,2,2,3,11,"4","4","4","4","4","6","6","8","12","15","7","8","8","10"))
```

```r
DATAreg <- GRanges(seqnames=DATA$chromosome, ranges=
                   IRanges(start = DATA$start, end=DATA$end))
DATA <- DATA[overlapsAny(DATAreg, regions),]
```
\texttt{dim(DATA)}

### Convert to MethylKit Input ###

\begin{verbatim}
for (i in 1:length(Samples$SampleOrder)) {
  sample <- paste("ABS_", Samples$SampleOrder[i], sep=""
                    )
  tempo <- cbind(
    chrBase = paste("chr", as.character(DATA$chromosome
                        ), ".", DATA$start, sep=""),
    chr = as.character(DATA$chromosome),
    base = DATA$start,
    strand = rep("F", dim(DATA)[1]),
    coverage = DATA[, paste(sample, ",cov", sep="")],
    freqC = DATA[, paste(sample, ",met", sep="")],
    freqT = 100 - DATA[, paste(sample, ",met", sep="")]
  )
  tempo[is.na(tempo[,5]), 5] <- "0"
  tempo[as.numeric(as.character(tempo[,5])) < 100, 5:7] <- rep(NA, 3)
  tempo[!is.na(tempo[,5]), ]
  write.table(tempo, paste("./MethylKit_noLowCov/",
                         sample, ",methylKitInput.txt", sep=""), sep="\n"
             t", quote = FALSE, col.names = TRUE, row.names =
             FALSE)
}
\end{verbatim}

### Read MethylKit Input ###

\begin{verbatim}
file.list = as.list(list.files("./MethylKit_noLowCov"
                          , full.names = TRUE))

file.list <- file.list[order(as.numeric(sub("./
                         MethylKit_noLowCov/ABS_([0-9IVMDKO]+)_
                         methylKitInput.txt", "\1", file.list)))]

file.list2 = file.list[87:88]
\end{verbatim}

##### Time difference, un-paired ######

\begin{verbatim}
myobj = read(file.list[1:86], sample.id = as.list(
               paste("ABS_", Samples$SampleOrder[1:86], sep=""),
               assembly = "hg19", treatment = c(rep(0, 46), rep
               (1, 40)))
\end{verbatim}
#myobj = read(file.list2, sample.id=list("DKO","IVM"), assembly="hg19", treatment=c(0,1))

getMethylationStats(myobj[[1]], plot=FALSE, both.strands=FALSE, na.rm=TRUE)
getMethylationStats(myobj[[2]], plot=FALSE, both.strands=FALSE, na.rm=TRUE)
getMethylationStats(myobj[[1]], plot=TRUE, both.strands=FALSE)

filtered.myobj = filterByCoverage(myobj, lo.count=100, lo.perc=NULL, hi.count=NULL, hi.perc=NULL)
meth = unite(filtered.myobj, destrand=FALSE, min.per.group=10L)  # min.per.group=2L
tail(meth)

# getCorrelation(meth, plot=TRUE)
myDiff_time = calculateDiffMeth(meth, weighted.mean=FALSE, slim=FALSE)
myDiffsign_time = get.methylDiff(myDiff_time, difference=1, qvalue=0.01)

# gene.obj = read.transcript.features("refseq.hg19.bed")
# gene.obj = read.transcript.features("ensembl.hg19.bed")
# annotate.WithGenicParts(myDiffsign, gene.obj)

###### NonProg difference, un-paired #######
myobj = read(file.list[c(1:23,47:67)], sample.id=list(paste("ABS","SampleOrder[c(1:23,47:67)], sep="")),
             treatment=c(rep(0,23), rep(1,21)))

filtered.myobj = filterByCoverage(myobj, lo.count=100, lo.perc=NULL, hi.count=NULL, hi.perc=NULL)
meth = unite(filtered.myobj, destrand=FALSE, min.per.group=2L)
group=10L)  # min. per. group=2L
tail(meth)

getCorrelation(meth, plot=TRUE)
myDiff_timeNonProg=calculateDiffMeth(meth, weighted.
  -> mean=FALSE, slim=FALSE)
myDiffsign_timeNonProg=get.methylDiff(myDiff_
  -> timeNonProg, difference=1, qvalue=0.01)

##### Prog difference, un-paired ######
myobj=read( file.list [c(24:46,68:86)], sample.id=as
  -> list (paste("ABS_", Samples$SampleOrder [c
  -> (24:46,68:86)], sep="")), assembly="hg19",
  -> treatment=c(rep(0,23),rep(1,19)))
#myobj=read( file.list2, sample.id=list("DKO","IVM
  -> "), assembly="hg19", treatment=c(0,1))

filtered.myobj=filterByCoverage(myobj, lo.count=100,
  -> lo.perc=NULL, hi.count=NULL, hi.perc=NULL)

meth=unite(filtered.myobj, destrand=FALSE, min.per.
  -> group=10L)  # min. per. group=2L
tail(meth)

getCorrelation(meth, plot=TRUE)
myDiff_timeProg=calculateDiffMeth(meth, weighted.
  -> mean=FALSE, slim=FALSE)
myDiffsign_timeProg=get.methylDiff(myDiff_timeProg,
  -> difference=1, qvalue=0.01)

##### Group difference Round 1 ######
myobj=read( file.list [1:46], sample.id=as.list(
  -> paste("ABS_", Samples$SampleOrder [1:46], sep="""
  -> )), assembly="hg19", treatment=c(rep(0,23),rep
  -> (1,23)))

filtered.myobj=filterByCoverage(myobj, lo.count=100,
  -> lo.perc=NULL, hi.count=NULL, hi.perc=NULL)
meth=unite(filtered.myobj, destrand=FALSE, min.per.
  -> group=10L)  # min. per. group=2L
tail(meth)

myDiff_R1=calculateDiffMeth(meth, weighted.mean=
FALSE, slim=FALSE)
myDiffsign_R1 = get.methylDiff(myDiff_R1, difference = 1, qvalue = 0.01)

##### Group difference Round 2 #####
myobj = read(file.list[47:86], sample.id = as.list(
  paste("ABS_", Samples$SampleOrder[47:86], sep=""))
  , assembly="hg19", treatment = c(rep(0, 21), rep(1, 19)))
filtered.myobj = filterByCoverage(myobj, lo.count = 100,
  lo.perc=NULL, hi.count=NULL, hi.perc=NULL)
meth = unite(filtered.myobj, destrand = FALSE, min.per. group = 10L)  # min.per.group = 2L
tail(meth)

myDiff_R2 = calculateDiffMeth(meth, weighted.mean = FALSE, slim=FALSE)
myDiffsign_R2 = get.methylDiff(myDiff_R2, difference = 1, qvalue = 0.01)

##### Group difference #####
myobj = read(file.list[1:86], sample.id = as.list(
  paste("ABS_", Samples$SampleOrder[1:86], sep=""))
  , assembly="hg19", treatment = c(rep(0, 23), rep(1, 19)))
filtered.myobj = filterByCoverage(myobj, lo.count = 100,
  lo.perc=NULL, hi.count=NULL, hi.perc=NULL)
meth = unite(filtered.myobj, destrand = FALSE, min.per. group = 10L)  # min.per.group = 2L
tail(meth)

myDiff_Groups = calculateDiffMeth(meth, weighted.mean = FALSE, slim=FALSE)
myDiffsign_Groups = get.methylDiff(myDiff_Groups, difference = 1, qvalue = 0.01)

##### Subset targetted regions (optional) ######
TimeTargets <- as(myDiffsign_time, "GRanges")
R1Targets <- as(myDiffsign_R1, "GRanges")
R2Targets <- as(myDiffsign_R2, "GRanges")
GroupTargets <- as(myDiffsign_Groups, "GRanges")

TimeTargets <- subsetByOverlaps(TimeTargets, regions <- [17:23])  # regions [17:23]
R1Targets <- subsetByOverlaps(R1Targets, regions <- [1:4])  # regions [1:4]
R2Targets <- subsetByOverlaps(R2Targets, regions <- [12:16])  # regions [12:16]
GroupTargets <- subsetByOverlaps(GroupTargets, regions <- [5:11])  # regions [5:11]

### Regional Level, repeat analyses but for global methylation of each amplicon #######
### Read MethylKit Input ###
file.list = as.list(list.files("./MethylKit_noLowCov", full.names=TRUE))
file.list <- file.list[order(as.numeric(sub("./MethylKit_noLowCov/ABS(\[0–9IVMDKO\]+)/methylKitInput.txt", "\\1", file.list)))]
file.list2 = file.list[87:88]

#### Time difference, unpaired ######
myobj = read(file.list[1:86], sample.id = as.list(paste("ABS_", Samples$SampleOrder[1:86], sep="")), assembly="hg19", treatment = c(rep(0, 46), rep(1, 40)))
targetRegions = regionCounts(myobj, regions)

#### Plot figures with ggplot #######
### group differences ###
library(ggplot2)
\[\text{CpG} = \text{rep}("15:22347472", \text{length}(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "15:22347472" , c(8:30,54:74)])) + \text{length}(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "11:43095973" , c(31:53,75:93)])), \text{rep}("11:43095973", \text{length}(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "11:43095973" , c(8:30,54:74)])) + \text{length}(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "3:162767494" , c(8:30,54:74)])) + \text{length}(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "3:162767494" , c(31:53,75:93)]))
\]

\[\text{Group} = \text{rep}(\text{c(rep("Non-Progressor",23+21),rep("Progressor",23+19))),3)
\]

data1$CpG <- \text{factor}(\text{data1}\$CpG, \text{levels} = \text{c}("3:162767494","11:43095973","15:22347472",3))
\]

\[\text{P1} \leftarrow \text{ggplot} (\text{data1}, \text{aes(x=factor(CpG), y=Meth, fill = Group)) + geom\_boxplot(outlier\_shape=NA) + geom\_point(alpha=0.5, position=position_jitterdodge(jitter.width =0.2)) + labs(x = "CpG\_location", y = "Methylation\_Degree\%(\%)", }\]

\[\text{ggsave("Figure\_1\_MethylationDifferences\_AllSamples.pdf",P1, width=7,height=5)}
\]

### time differences ###
\[\text{data2} \leftarrow \text{data.frame}(\text{Meth}=\text{c(\text{unlist} (\text{DATA}[\text{DATA}\$CpG == "7:33828043", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "7:33828043", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:2511694", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:2511694", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:2511817", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:2511817", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:26669622", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:26669622", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:26669687", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "8:26669687", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "10:98145961", c(8:53)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "10:98145961", c(54:93)]), \text{unlist} (\text{DATA}[\text{DATA}\$CpG == "10:98145961", c(54:93)]))
\]
DATA$CpG=="12:44148898",c(8:53)],unlist(DATA
[DATA$CpG=="12:44148898",c(54:93)],)
CpG=c(rep("7:33828043",46+40),rep("8:2511694"
   ,46+40),rep("8:2511817",46+40),rep("8:26669622",46+40),
   rep("8:26669687",46+40), rep("10:98145961",46+40),
   rep(12:44148898"
   ,46+40),
Round=rep(c(rep("First",46),rep("Second",40)),7)
)
data2$CpG<factor(data2$CpG, levels=c("
   7:33828043","8:2511694","8:2511817","8:26669622",
   8:26669687","10:98145961","12:44148898")))
P1<ggplot(data2, aes(x=factor(CpG), y=Meth, fill
   = Round)) + geom_boxplot(outlier.shape=NA) +
   ylim(35,100) + theme(axis.text.x = element_
   text(angle = 90, hjust = 1)) +
   geom_point(alpha =0.5, position =position_
   jitterdodge(jitter.width =0.2)) + labs(x="CpGolocation",y="MethylationDegree(%)"
   )

P1
ggsave("Figure XXX MethylationDifferences Time.pdf"
   ,P1, width=8,height=8)
Supplementary Figures
Figure B.1: A CpG fragment plot showing that mapped fragments in each sample are enriched in CpG content. The average enrichment curve (blue) shows a peak around 4 CpGs/fragment.
Figure B.2: Scatterplot of peak coverage in all methylation cores of the technical replicates. The top row shows the raw count, the bottom row represents normalised counts. The left panels depict the first round replicates and the right panels contain the second round replicates.
Figure B.3: Unsupervised clustering based on the peak coverage of the 2000 most variable methylation cores shows correct pairing of first and second round samples with the exception of AS1.07-AS2.28.
Figure B.4: Density plots of the P-values for progressor versus non-progressor differential methylation obtained from the real data (first column) and the permuted data (second column) without a subject specific term in the model. The first row shows round 1 and round 2 samples combined, the second row only round 1 samples and the last row only round 2 samples.
Figure B.5: Density plots of the P-values for inter-round differential methylation obtained from the real data (first column) and the permuted data (second column) with the inclusion of a subject specific term progressor status in the model. The first row shows round 1 vs. round 2 combined, the second row only the non-progressors and the last row only the progressors.
Supplementary Tables
### Table C.1: Cardiovascular characteristics of the progressors and non-progressors used for MethylCap-seq.

| Measure         | R1 NP (Mean ± SD) | R1 P (Mean ± SD) | P-value | R2 NP (Mean ± SD) | R2 P (Mean ± SD) | P-value | R1 NP (Mean ± SD) | R1 P (Mean ± SD) | P-value | R2 NP (Mean ± SD) | R2 P (Mean ± SD) | P-value | R1 NP (Mean ± SD) | R1 P (Mean ± SD) | P-value | R2 NP (Mean ± SD) | R2 P (Mean ± SD) | P-value | R1 NP (Mean ± SD) | R1 P (Mean ± SD) | P-value |
|-----------------|-------------------|------------------|---------|-------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|
| Age (years)     | 44.4 ± 6.58       | 46.8 ± 5.03      | 0.367   | 53.1 ± 6.58       | 55.5 ± 4.99      | 0.366   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| Weight (kg)     | 60.4 ± 8.85       | 62.0 ± 8.90      | 0.688   | 64.3 ± 9.49       | 65.7 ± 9.74      | 0.749   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| Height (cm)     | 164 ± 5.24        | 164 ± 5.62       | 0.935   | 164 ± 5.09        | 163 ± 5.81       | 0.658   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| BMI (kg/m)      | 22.3 ± 2.75       | 22.9 ± 2.52      | 0.594   | 23.6 ± 2.98       | 24.4 ± 2.63      | 0.594   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| SBP (mmHg)      | 115 ± 8.51        | 120 ± 6.96       | 0.168   | 118 ± 9.99        | 120 ± 9.54       | 0.524   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| DBP (mmHg)      | 73.3 ± 7.60       | 73.3 ± 4.55      | 1.00    | 75.2 ± 5.38       | 75.6 ± 4.00      | 0.647   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| PP (mmHg)       | 42.1 ± 4.18       | 47.1 ± 5.38      | 0.044   | 43.1 ± 6.21       | 44.7 ± 7.71      | 0.608   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| Heart Rate (bpm)| 64.8 ± 6.18       | 66.3 ± 7.13      | 0.621   | 66.1 ± 7.00       | 69.2 ± 6.27      | 0.311   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| Total Cholesterol (mg/dl) | 212 ± 30.3  | 231 ± 28.5      | 0.179   | 212 ± 32.0        | 214 ± 26.7       | 0.179   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| HDL Cholesterol (mg/dl) | 81.5 ± 21.0 | 79.7 ± 13.1    | 0.821   | 81.6 ± 13.1       | 81.6 ± 13.1      | 0.821   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| LDL Cholesterol (mg/dl) | 116 ± 19.9 | 136 ± 22.6     | 0.052   | 116 ± 19.9        | 116 ± 19.9       | 0.052   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| OxLDL (mol/l)   | 71.9 ± 15.0       | 85.5 ± 46.5      | 0.579   | 96.3 ± 46.1       | 122 ± 40.3       | 0.143   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| LVM (g)         | 113 ± 18          | 136 ± 22.6       | 1.00    | 136 ± 22.6        | 136 ± 22.6       | 1.00    |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| PWV (m/s)       | 6.43 ± 1.21       | 6.3 ± 1.94       | 0.637   | 6.43 ± 1.21       | 6.3 ± 1.94       | 0.637   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| IL-6 (pg/ml)    | 0.54 ± 1.08       | 0.72 ± 1.52      | 0.704   | 0.64 ± 0.37       | 0.64 ± 0.37      | 0.704   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |
| hCRP (mg/l)     | 1.68 ± 2.56       | 1.34 ± 1.02      | 0.801   | 1.68 ± 2.56       | 1.34 ± 1.02      | 0.801   |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |                  |                  |         |

*Welch t-test*  
§ *Wilcoxon Rank Sum test*
**Table C.2**: Cardiovascular characteristics of the progressors and non-progressors used for targeted bisulphite deep sequencing.  
*Mean ± standard deviation, R1: Round 1, R2: Round 2, NP: Non-progressor, P: Progressor*

<table>
<thead>
<tr>
<th>Measure</th>
<th>R1NP</th>
<th>R1P</th>
<th>P-value Δ R1</th>
<th>R2NP</th>
<th>R2P</th>
<th>P-value Δ R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>44.5</td>
<td>45.2</td>
<td>0.247 *</td>
<td>53.9</td>
<td>54.3</td>
<td>0.510 *</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>62.1</td>
<td>62.9</td>
<td>0.685</td>
<td>65.4</td>
<td>64.5</td>
<td>0.742</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>163</td>
<td>164</td>
<td>0.663</td>
<td>162</td>
<td>163</td>
<td>0.888</td>
</tr>
<tr>
<td>BMI (kg/m)</td>
<td>23.4</td>
<td>23.4</td>
<td>0.964</td>
<td>24.6</td>
<td>24.1</td>
<td>0.586</td>
</tr>
<tr>
<td>SBP (mmHg)</td>
<td>118</td>
<td>122</td>
<td>0.159</td>
<td>122</td>
<td>128</td>
<td>0.102</td>
</tr>
<tr>
<td>DBP (mmHg)</td>
<td>74.7</td>
<td>75.6</td>
<td>0.668</td>
<td>77.8</td>
<td>79.0</td>
<td>0.659</td>
</tr>
<tr>
<td>PP (mmHg)</td>
<td>43.5</td>
<td>46.7</td>
<td>0.101</td>
<td>44.1</td>
<td>49.1</td>
<td>0.031</td>
</tr>
<tr>
<td>Heart Rate (bpm)</td>
<td>63.8</td>
<td>64.4</td>
<td>0.545</td>
<td>64.4</td>
<td>63.3</td>
<td>0.674</td>
</tr>
<tr>
<td>Total Cholesterol (mg/dl)</td>
<td>201</td>
<td>225</td>
<td>0.020</td>
<td>212</td>
<td>237</td>
<td>0.017 *</td>
</tr>
<tr>
<td>HDL Cholesterol (mg/dl)</td>
<td>72.4</td>
<td>76.3</td>
<td>0.415</td>
<td>75.3</td>
<td>79.4</td>
<td>0.523</td>
</tr>
<tr>
<td>LDL Cholesterol (mg/dl)</td>
<td>112</td>
<td>133</td>
<td>0.035</td>
<td>119</td>
<td>142</td>
<td>0.016</td>
</tr>
<tr>
<td>OxLDL (mol/l)</td>
<td>77.5</td>
<td>85.9</td>
<td>0.648</td>
<td>80.5</td>
<td>122</td>
<td>0.038 *</td>
</tr>
<tr>
<td>LVM (g)</td>
<td>119</td>
<td>113</td>
<td>0.326 §</td>
<td>122</td>
<td>127</td>
<td>0.543</td>
</tr>
<tr>
<td>PWV (m/s)</td>
<td>5.42</td>
<td>6.01</td>
<td>0.904 *</td>
<td>7.13</td>
<td>7.11</td>
<td>0.706 *</td>
</tr>
<tr>
<td>IL6 (pg/ml)</td>
<td>1.58</td>
<td>1.41</td>
<td>0.474 *</td>
<td>2.05</td>
<td>1.50</td>
<td>0.410 *</td>
</tr>
<tr>
<td>hCRP (mg/l)</td>
<td>2.68</td>
<td>4.62</td>
<td>0.801 *</td>
<td>1.88</td>
<td>1.74</td>
<td>0.652 *</td>
</tr>
</tbody>
</table>

* Welch t-test  
* Wilcoxon Rank Sum test
Table C.3: Summary statistics for MethylCap-seq results.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Number of sequenced reads</th>
<th>Mapped Reads</th>
<th>Mapping Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1.01</td>
<td>26420862</td>
<td>16765738</td>
<td>63.5%</td>
</tr>
<tr>
<td>AS1.02</td>
<td>29335634</td>
<td>19443989</td>
<td>66.3%</td>
</tr>
<tr>
<td>AS1.03</td>
<td>56821828</td>
<td>28689322</td>
<td>50.5%</td>
</tr>
<tr>
<td>AS1.04</td>
<td>48131200</td>
<td>33229553</td>
<td>69.0%</td>
</tr>
<tr>
<td>AS1.05</td>
<td>38748299</td>
<td>24700461</td>
<td>63.8%</td>
</tr>
<tr>
<td>AS1.06</td>
<td>53831659</td>
<td>34332784</td>
<td>63.8%</td>
</tr>
<tr>
<td>AS1.07</td>
<td>49035192</td>
<td>33065763</td>
<td>67.4%</td>
</tr>
<tr>
<td>AS1.08</td>
<td>34578036</td>
<td>22878488</td>
<td>66.2%</td>
</tr>
<tr>
<td>AS1.09</td>
<td>27525918</td>
<td>18083808</td>
<td>65.7%</td>
</tr>
<tr>
<td>AS1.10</td>
<td>75169812</td>
<td>47901720</td>
<td>63.7%</td>
</tr>
<tr>
<td>AS1.11</td>
<td>30306313</td>
<td>18757175</td>
<td>61.9%</td>
</tr>
<tr>
<td>AS1.12</td>
<td>26632727</td>
<td>18970945</td>
<td>71.2%</td>
</tr>
<tr>
<td>AS1.13</td>
<td>31633415</td>
<td>21128998</td>
<td>66.8%</td>
</tr>
<tr>
<td>AS1.14</td>
<td>35419833</td>
<td>23821420</td>
<td>67.3%</td>
</tr>
<tr>
<td>AS1.15</td>
<td>24229197</td>
<td>15237524</td>
<td>62.9%</td>
</tr>
<tr>
<td>AS1.16</td>
<td>13428110</td>
<td>9210488</td>
<td>68.6%</td>
</tr>
<tr>
<td>AS1.17</td>
<td>33960434</td>
<td>21851992</td>
<td>64.4%</td>
</tr>
<tr>
<td>AS1.18</td>
<td>56375966</td>
<td>35777729</td>
<td>63.5%</td>
</tr>
<tr>
<td>AS1.19</td>
<td>53920080</td>
<td>31588567</td>
<td>58.6%</td>
</tr>
<tr>
<td>AS1.20</td>
<td>76920869</td>
<td>49577548</td>
<td>64.5%</td>
</tr>
<tr>
<td>AS1.21</td>
<td>44337119</td>
<td>29403469</td>
<td>66.3%</td>
</tr>
<tr>
<td>AS2.22</td>
<td>18994347</td>
<td>11585635</td>
<td>61.0%</td>
</tr>
<tr>
<td>AS2.23</td>
<td>45335229</td>
<td>28945794</td>
<td>63.9%</td>
</tr>
<tr>
<td>AS2.24</td>
<td>55119830</td>
<td>33519219</td>
<td>60.8%</td>
</tr>
<tr>
<td>AS2.25</td>
<td>57331447</td>
<td>35603739</td>
<td>62.1%</td>
</tr>
<tr>
<td>AS2.26</td>
<td>40902371</td>
<td>24895402</td>
<td>60.9%</td>
</tr>
<tr>
<td>AS2.27</td>
<td>41066697</td>
<td>25790668</td>
<td>62.8%</td>
</tr>
<tr>
<td>AS2.28</td>
<td>37931760</td>
<td>23209210</td>
<td>61.2%</td>
</tr>
<tr>
<td>AS2.29</td>
<td>39640041</td>
<td>25935663</td>
<td>65.4%</td>
</tr>
<tr>
<td>AS2.30</td>
<td>28162071</td>
<td>17394222</td>
<td>61.8%</td>
</tr>
<tr>
<td>AS2.31</td>
<td>26417516</td>
<td>16791956</td>
<td>63.6%</td>
</tr>
<tr>
<td>AS2.32</td>
<td>35662502</td>
<td>21197878</td>
<td>59.4%</td>
</tr>
<tr>
<td>AS2.33</td>
<td>34413172</td>
<td>22328669</td>
<td>64.9%</td>
</tr>
<tr>
<td>AS2.34</td>
<td>60115268</td>
<td>36346797</td>
<td>60.5%</td>
</tr>
<tr>
<td>AS2.35</td>
<td>32660158</td>
<td>20878919</td>
<td>63.9%</td>
</tr>
<tr>
<td>AS2.36</td>
<td>35140777</td>
<td>20893016</td>
<td>59.5%</td>
</tr>
<tr>
<td>AS2.37</td>
<td>40282893</td>
<td>22101179</td>
<td>54.9%</td>
</tr>
<tr>
<td>AS2.38</td>
<td>36065404</td>
<td>22166726</td>
<td>61.5%</td>
</tr>
<tr>
<td>AS2.39</td>
<td>24692348</td>
<td>15524679</td>
<td>62.9%</td>
</tr>
<tr>
<td>AS2.40</td>
<td>59533619</td>
<td>38031074</td>
<td>63.9%</td>
</tr>
<tr>
<td>AS2.41</td>
<td>65110198</td>
<td>43317960</td>
<td>66.5%</td>
</tr>
<tr>
<td>AS2.42</td>
<td>75340968</td>
<td>48231507</td>
<td>64.0%</td>
</tr>
</tbody>
</table>
Table C.4: Primers used for amplification in targeted bisulphite sequencing validation of atherosclerosis progression related loci.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Hypothesis</th>
<th>Forward Sequence</th>
<th>Reverse Sequence</th>
<th>chr</th>
<th>start</th>
<th>stop</th>
<th>successful amplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOS1</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>2</td>
<td>39285112</td>
<td>39285399</td>
<td>yes</td>
</tr>
<tr>
<td>YRNA</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>2</td>
<td>2326967</td>
<td>23270358</td>
<td>yes</td>
</tr>
<tr>
<td>RP1-10022.1</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>3</td>
<td>162767304</td>
<td>162767652</td>
<td>yes</td>
</tr>
<tr>
<td>ADRA1A</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>8</td>
<td>26669480</td>
<td>26669791</td>
<td>yes</td>
</tr>
<tr>
<td>intergenic_4</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>11</td>
<td>43090905</td>
<td>43091615</td>
<td>yes</td>
</tr>
<tr>
<td>SOS2</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>14</td>
<td>50687972</td>
<td>50688193</td>
<td>yes</td>
</tr>
<tr>
<td>LNC1</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>15</td>
<td>22347371</td>
<td>22347635</td>
<td>yes</td>
</tr>
<tr>
<td>ABCD1P4</td>
<td>prg vs npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>22</td>
<td>16866101</td>
<td>16866378</td>
<td>no</td>
</tr>
<tr>
<td>CRYGS</td>
<td>R1prg vs R1npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>3</td>
<td>18626296</td>
<td>18626318</td>
<td>yes</td>
</tr>
<tr>
<td>FAM160A1</td>
<td>R1prg vs R1npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>4</td>
<td>15244395</td>
<td>15244429</td>
<td>yes</td>
</tr>
<tr>
<td>intergenic_1</td>
<td>R1prg vs R1npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>6</td>
<td>83420357</td>
<td>83420808</td>
<td>yes</td>
</tr>
<tr>
<td>intergenic_2</td>
<td>R1prg vs R1npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>6</td>
<td>15869387</td>
<td>15869981</td>
<td>yes</td>
</tr>
<tr>
<td>ANTRX2</td>
<td>R2prg vs R2npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>4</td>
<td>80888013</td>
<td>80888531</td>
<td>no</td>
</tr>
<tr>
<td>MAPK10</td>
<td>R2prg vs R2npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>4</td>
<td>87251778</td>
<td>87251991</td>
<td>no</td>
</tr>
<tr>
<td>MAPK10</td>
<td>R2prg vs R2npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>4</td>
<td>87251890</td>
<td>87252209</td>
<td>yes</td>
</tr>
<tr>
<td>LNC4</td>
<td>R2prg vs R2npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>6</td>
<td>14614617</td>
<td>14614640</td>
<td>yes</td>
</tr>
<tr>
<td>LNC4</td>
<td>R2prg vs R2npr</td>
<td>TTTGTTGATAATAATAATTTTGGTGAATAA</td>
<td>AAAAAATAAACCTAAACCTCTCCAATCC</td>
<td>6</td>
<td>146146470</td>
<td>146146705</td>
<td>yes</td>
</tr>
<tr>
<td>Methylation Core PrimerPair</td>
<td>ID</td>
<td>Forward Sequence</td>
<td>Reverse Sequence</td>
<td>chr start</td>
<td>stop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>----</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERGEN5</td>
<td>42379236</td>
<td>TGGATTGGTTTGAAATAAAATTATTT</td>
<td>ATTCTTTACACCCACAAAACCTACT</td>
<td>7</td>
<td>33827898</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERGEN6</td>
<td>42536705</td>
<td>TGGGGGGTTTTTTAGATTAATTTAT</td>
<td>CTCCAACAATAAAAACAACCTTCTC</td>
<td>8</td>
<td>2511585</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADRA1A</td>
<td>42569384</td>
<td>AGGTGATGGGTATTGGAATAG</td>
<td>AAATCACCAAAAATTACCACTACTAA</td>
<td>8</td>
<td>26669480</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLL2</td>
<td>39561247</td>
<td>TGGTTTTTAGTTTTGAGTTATTTTTTAATA</td>
<td>AATAAATCCTATTTCTATTTCCCCC</td>
<td>10</td>
<td>98145748</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUS7L</td>
<td>39830328</td>
<td>GGAGTTTAAAATATTAGTTTTTTTT</td>
<td>AATTCTTTATATTTCTTTAATAATCAC</td>
<td>12</td>
<td>44148667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPR176</td>
<td>40193998</td>
<td>ATATGTTTTTTATTAGTGGATGAGAGG</td>
<td>ACTTCTATTCTCCTTTACAAAAAACATA</td>
<td>15</td>
<td>40152369</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description:Primer used for amplification in targeted bisulphite sequencing validation of ageing related loci.
Table C.6: Validation results for differential methylation between the second and first study rounds at the individual CpG level.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERGEN5_1</td>
<td>7:33827925</td>
<td>0.31</td>
<td>-0.76%</td>
<td>0.00</td>
<td>1.19%</td>
<td>0.00</td>
<td>0.18%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33827953</td>
<td>0.48</td>
<td>-0.57%</td>
<td>0.69</td>
<td>-0.27%</td>
<td>0.91</td>
<td>-0.41%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33827968</td>
<td>0.04</td>
<td>-0.07%</td>
<td>0.00</td>
<td>0.13%</td>
<td>0.00</td>
<td>0.03%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33827974</td>
<td>0.00</td>
<td>-0.50%</td>
<td>0.34</td>
<td>-0.07%</td>
<td>0.20</td>
<td>-0.29%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33828043</td>
<td>0.16</td>
<td>-0.22%</td>
<td>0.07</td>
<td>-0.16%</td>
<td>0.01</td>
<td>-0.20%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33828111</td>
<td>0.21</td>
<td>-0.48%</td>
<td>0.00</td>
<td>0.27%</td>
<td>0.00</td>
<td>-0.11%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33828145</td>
<td>0.81</td>
<td>-0.24%</td>
<td>0.62</td>
<td>0.02%</td>
<td>0.52</td>
<td>-0.13%</td>
</tr>
<tr>
<td>INTERGEN5_1</td>
<td>7:33828169</td>
<td>0.83</td>
<td>-0.11%</td>
<td>0.02</td>
<td>-0.13%</td>
<td>0.08</td>
<td>-0.12%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511598</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511626</td>
<td>0.40</td>
<td>0.64%</td>
<td>0.00</td>
<td>0.63%</td>
<td>0.06</td>
<td>0.66%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511640</td>
<td>0.00</td>
<td>0.96%</td>
<td>0.20</td>
<td>0.20%</td>
<td>0.00</td>
<td>0.61%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511674</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511665</td>
<td>0.12</td>
<td>-0.29%</td>
<td>0.12</td>
<td>-0.70%</td>
<td>0.01</td>
<td>-0.50%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511694</td>
<td>0.00</td>
<td>1.54%</td>
<td>0.97</td>
<td>0.03%</td>
<td>0.00</td>
<td>0.79%</td>
</tr>
<tr>
<td>INTERGEN6_2</td>
<td>8:2511697</td>
<td>0.12</td>
<td>0.56%</td>
<td>0.38</td>
<td>-0.73%</td>
<td>0.03</td>
<td>0.01%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511777</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511817</td>
<td>0.02</td>
<td>-0.51%</td>
<td>0.00</td>
<td>-0.72%</td>
<td>0.00</td>
<td>-0.66%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511824</td>
<td>0.00</td>
<td>-1.11%</td>
<td>0.14</td>
<td>-0.63%</td>
<td>0.00</td>
<td>-0.86%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511894</td>
<td>0.79</td>
<td>-0.04%</td>
<td>0.84</td>
<td>-0.26%</td>
<td>0.63</td>
<td>-0.16%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511933</td>
<td>0.00</td>
<td>-0.84%</td>
<td>0.98</td>
<td>0.18%</td>
<td>0.00</td>
<td>-0.34%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511938</td>
<td>0.00</td>
<td>-0.80%</td>
<td>0.51</td>
<td>-0.21%</td>
<td>0.00</td>
<td>-0.54%</td>
</tr>
<tr>
<td>INTERGEN6_1</td>
<td>8:2511948</td>
<td>0.00</td>
<td>-0.62%</td>
<td>0.00</td>
<td>-0.42%</td>
<td>0.00</td>
<td>-0.45%</td>
</tr>
<tr>
<td>ADRA1A_1</td>
<td>8:26669622</td>
<td>0.00</td>
<td>-4.42%</td>
<td>0.45</td>
<td>1.97%</td>
<td>0.01</td>
<td>-1.23%</td>
</tr>
<tr>
<td>ADRA1A_1</td>
<td>8:26669687</td>
<td>0.24</td>
<td>0.25%</td>
<td>0.00</td>
<td>1.94%</td>
<td>0.00</td>
<td>1.05%</td>
</tr>
<tr>
<td>ADRA1A_1</td>
<td>8:26669722</td>
<td>0.31</td>
<td>-0.16%</td>
<td>0.69</td>
<td>-0.39%</td>
<td>0.27</td>
<td>-0.34%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145811</td>
<td>0.35</td>
<td>-0.06%</td>
<td>0.20</td>
<td>0.61%</td>
<td>0.91</td>
<td>0.26%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145855</td>
<td>0.00</td>
<td>-0.63%</td>
<td>0.45</td>
<td>0.07%</td>
<td>0.00</td>
<td>-0.31%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145865</td>
<td>0.00</td>
<td>-0.64%</td>
<td>0.13</td>
<td>-0.08%</td>
<td>0.00</td>
<td>-0.35%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145914</td>
<td>0.78</td>
<td>0.54%</td>
<td>0.00</td>
<td>0.72%</td>
<td>0.01</td>
<td>0.65%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145961</td>
<td>0.00</td>
<td>0.98%</td>
<td>0.00</td>
<td>0.33%</td>
<td>0.00</td>
<td>0.64%</td>
</tr>
<tr>
<td>TLL2_1</td>
<td>10:98145978</td>
<td>0.77</td>
<td>0.95%</td>
<td>0.01</td>
<td>0.14%</td>
<td>0.10</td>
<td>0.51%</td>
</tr>
<tr>
<td>PUS7L_1</td>
<td>12:44148715</td>
<td>0.84</td>
<td>-0.12%</td>
<td>0.69</td>
<td>-0.10%</td>
<td>0.91</td>
<td>-0.10%</td>
</tr>
<tr>
<td>PUS7L_1</td>
<td>12:44148898</td>
<td>0.00</td>
<td>-0.44%</td>
<td>0.00</td>
<td>-0.72%</td>
<td>0.00</td>
<td>-0.56%</td>
</tr>
<tr>
<td>PUS7L_1</td>
<td>12:44148988</td>
<td>0.61</td>
<td>-0.36%</td>
<td>0.04</td>
<td>-0.05%</td>
<td>0.31</td>
<td>-0.19%</td>
</tr>
<tr>
<td>GPR176_1</td>
<td>15:40152461</td>
<td>0.00</td>
<td>-0.57%</td>
<td>0.00</td>
<td>-0.35%</td>
<td>0.00</td>
<td>-0.45%</td>
</tr>
<tr>
<td>GPR176_1</td>
<td>15:40152523</td>
<td>0.00</td>
<td>0.12%</td>
<td>0.07</td>
<td>0.18%</td>
<td>0.00</td>
<td>0.14%</td>
</tr>
<tr>
<td>GPR176_1</td>
<td>15:40152613</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table C.7: Overlap of KEGG pathways identified based on differentially methylated genes in atherosclerosis progression and inter-round ageing. The “Fraction of Affected Genes” is calculated as "Overlap/Pathway Size" divided by the size of the union of "Athero" and "Time".

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Time (n)</th>
<th>Athero (n)</th>
<th>Overlap</th>
<th>Pathway Size</th>
<th>Fraction of Affected Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreatic secretion</td>
<td>25</td>
<td>25</td>
<td>16</td>
<td>101</td>
<td>0.4706</td>
</tr>
<tr>
<td>Long-term depression</td>
<td>25</td>
<td>14</td>
<td>12</td>
<td>70</td>
<td>0.4444</td>
</tr>
<tr>
<td>Long-term potentiation</td>
<td>25</td>
<td>15</td>
<td>12</td>
<td>70</td>
<td>0.4286</td>
</tr>
<tr>
<td>Salivary secretion</td>
<td>30</td>
<td>21</td>
<td>15</td>
<td>89</td>
<td>0.4167</td>
</tr>
<tr>
<td>ErbB signaling pathway</td>
<td>29</td>
<td>22</td>
<td>15</td>
<td>87</td>
<td>0.4167</td>
</tr>
<tr>
<td>Gap junction</td>
<td>27</td>
<td>21</td>
<td>14</td>
<td>90</td>
<td>0.4118</td>
</tr>
<tr>
<td>Axon guidance</td>
<td>44</td>
<td>26</td>
<td>20</td>
<td>129</td>
<td>0.4000</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>24</td>
<td>16</td>
<td>11</td>
<td>89</td>
<td>0.3793</td>
</tr>
<tr>
<td>GnRH signaling pathway</td>
<td>31</td>
<td>24</td>
<td>15</td>
<td>101</td>
<td>0.3750</td>
</tr>
<tr>
<td>Type II diabetes mellitus</td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>48</td>
<td>0.3684</td>
</tr>
<tr>
<td>Glioma</td>
<td>18</td>
<td>12</td>
<td>8</td>
<td>65</td>
<td>0.3636</td>
</tr>
<tr>
<td>Arrhythmogenic right ventricular cardiomyopathy (ARVC)</td>
<td>29</td>
<td>20</td>
<td>13</td>
<td>74</td>
<td>0.3611</td>
</tr>
<tr>
<td>Hypertrophic cardiomyopathy (HCM)</td>
<td>25</td>
<td>17</td>
<td>11</td>
<td>83</td>
<td>0.3548</td>
</tr>
<tr>
<td>Phosphatidylinositol signaling system</td>
<td>30</td>
<td>16</td>
<td>12</td>
<td>78</td>
<td>0.3529</td>
</tr>
<tr>
<td>Regulation of actin cytoskeleton</td>
<td>51</td>
<td>31</td>
<td>21</td>
<td>213</td>
<td>0.3443</td>
</tr>
<tr>
<td>Amoebiasis</td>
<td>34</td>
<td>17</td>
<td>13</td>
<td>106</td>
<td>0.3421</td>
</tr>
<tr>
<td>Melanogenesis</td>
<td>27</td>
<td>17</td>
<td>11</td>
<td>101</td>
<td>0.3333</td>
</tr>
<tr>
<td>Neurotrophin signaling pathway</td>
<td>26</td>
<td>23</td>
<td>12</td>
<td>127</td>
<td>0.3243</td>
</tr>
<tr>
<td>Pathways in cancer</td>
<td>85</td>
<td>54</td>
<td>34</td>
<td>326</td>
<td>0.3238</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>66</td>
<td>38</td>
<td>25</td>
<td>200</td>
<td>0.3165</td>
</tr>
<tr>
<td>ECM-receptor interaction</td>
<td>30</td>
<td>16</td>
<td>11</td>
<td>85</td>
<td>0.3143</td>
</tr>
<tr>
<td>Calcium signaling pathway</td>
<td>58</td>
<td>31</td>
<td>21</td>
<td>177</td>
<td>0.3088</td>
</tr>
<tr>
<td>Wnt signaling pathway</td>
<td>35</td>
<td>25</td>
<td>14</td>
<td>150</td>
<td>0.3043</td>
</tr>
<tr>
<td>Dilated cardiomyopathy</td>
<td>30</td>
<td>22</td>
<td>12</td>
<td>90</td>
<td>0.3000</td>
</tr>
<tr>
<td>Chemokine signaling pathway</td>
<td>38</td>
<td>27</td>
<td>15</td>
<td>189</td>
<td>0.3000</td>
</tr>
<tr>
<td>Gastric acid secretion</td>
<td>29</td>
<td>15</td>
<td>10</td>
<td>74</td>
<td>0.2941</td>
</tr>
<tr>
<td>Leukocyte transendothelial migration</td>
<td>27</td>
<td>17</td>
<td>10</td>
<td>116</td>
<td>0.2941</td>
</tr>
<tr>
<td>MAPK signaling pathway</td>
<td>69</td>
<td>37</td>
<td>24</td>
<td>268</td>
<td>0.2927</td>
</tr>
<tr>
<td>Vascular smooth muscle contraction</td>
<td>34</td>
<td>21</td>
<td>12</td>
<td>116</td>
<td>0.2791</td>
</tr>
<tr>
<td>Adherens junction</td>
<td>29</td>
<td>13</td>
<td>9</td>
<td>73</td>
<td>0.2727</td>
</tr>
<tr>
<td>Tight junction</td>
<td>40</td>
<td>25</td>
<td>13</td>
<td>132</td>
<td>0.2500</td>
</tr>
<tr>
<td>Neuroactive ligand-receptor interaction</td>
<td>62</td>
<td>32</td>
<td>18</td>
<td>272</td>
<td>0.2368</td>
</tr>
<tr>
<td>Progesterone-mediated oocyte maturation</td>
<td>18</td>
<td>15</td>
<td>5</td>
<td>86</td>
<td>0.1786</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>224</td>
<td>116</td>
<td>49</td>
<td>1130</td>
<td>0.1684</td>
</tr>
</tbody>
</table>
Curriculum Vitae

Personalia

Name, surname: Simon L.I.J. Denil
Born: 06/03/1987, Leuven, Belgium
Sex: Male
Address: Oostervennestraat 85, 2400 Mol, Belgium
Phone: +32 (0) 9 264 99 22
E-mail: Simon.Denil@UGent.be

Publications

On cross-sectional associations of leukocyte telomere length with cardiac systolic, diastolic and vascular function. The Asklepios Study.
S Denil, E Rietzschel, M De Buyzere, C Van daele, ... Asklepios Investigators
PLoS ONE 9 (12), e115071

SNP-guided identification of monoallelic DNA-methylation events from enrichment based sequencing data
S Steyaert, W Van Criekinge, A De Paepe, S Denil, K Mensaert, ...
Nucleic Acids Research 42 (20), e157, 2014
Systemic Suppression of the Shoot Metabolism upon Rice Root Nematode Infection
T Kyndt, S Denil, L Bauters, W Van Criekinge, T De Meyer
PLoS ONE 9 (9), e106858, 2014

Staphylococcal enterotoxin B influences the DNA methylation pattern in nasal polyp tissue: a preliminary study
CA Prez-Novo, Y Zhang, S Denil, G Trooskens, T De Meyer, ...
Allergy, Asthma & Clinical Immunology 9 (1), 48, 2013

Nextgeneration technologies and data analytical approaches for epigenomics
K Mensaert, S Denil, G Trooskens, W Criekinge, O Thas, T De Meyer
Environmental and molecular mutagenesis 55 (3), 155-170, 2013

Family History of Cardiovascular Disease and Offspring Echocardiographic Left Ventricular Structure and Function: The Asklepios Study
T De Meyer, ML De Buyzere, TC Gillebert, S Denil, S Bekaert, ...
Journal of the American Society of Echocardiography 26 (11), 1290-1297. e2, 2013

A non-genetic, epigenetic-like mechanism of telomere length inheritance?
T De Meyer, K Vandepitte, S Denil, ML De Buyzere, ER Rietzschel, ...
European Journal of Human Genetics 22 (1), 1011, 2013

Transcriptional analysis through RNA sequencing of giant cells induced by Meloidogyne graminicola in rice roots
H Ji, G Gheysen, S Denil, K Lindsey, JF Topping, K Nahar, A Haegeman, ...
Journal of experimental botany 64 (12), 3885-3898, 2013

Addition of a Novel, Protective Family History Category Allows Better Profiling of Cardiovascular Risk and Atherosclerotic Burden in the General Population. The Asklepios Study
CM Vandeple, T De Meyer, ML De Buyzere, TC Gillebert, S Denil, ...
PLoS ONE 8 (5), e63185, 2013

Quality Evaluation of Methyl Binding Domain Based Kits for Enrichment DNA-Methylation Sequencing
T De Meyer, E Mampaey, M Vlemmix, S Denil, G Trooskens, JP Renard, ...
PLoS ONE 8 (3), e59068, 2013
No Shorter Telomeres in Subjects With a Family History of Cardiovascular Disease in the Asklepios Study
T De Meyer, ML De Buyzere, S Denil, D De Bacquer, P Segers, ...
Arteriosclerosis, Thrombosis, and Vascular Biology 32 (12), 3076-3081, 2012

Transcriptional reprogramming by root knot and migratory nematode infection in rice
T Kyndt, S Denil, A Haegeman, G Trooskens, L Bauters, W Criekinge, ...

Transcriptome analysis of rice mature root tissue and root tips in early development by massive parallel sequencing
T Kyndt, S Denil, A Haegeman, G Trooskens, T De Meyer, ...
Journal of experimental botany 63 (5), 2141-2157; 2012

Conferences

Leveraging Modest Means to Great Effect - A Case Study in Basal Cell Carcinoma
Next Generation Sequencing Data Congress, 2013, London, poster

Mining the Rest-Fraction in RNA-seq Experiments
6th BeNeLux Bioinformatics Conference, 2011, Luxemburg, poster

Telomere Length is Associated with Physical Indicators of Heart Failure
Berzelius Symposium 85, 2011, Stockholm, poster with oral presentation
Best poster award

Education

Master of applied biological sciences: Cell and gene biotechnology
Ghent University, Belgium
2008-2010

Erasmus program exchange semester
Universität für Bodenkultur Wien, Austria
2009
**Bachelor of applied biological sciences: Cell and gene biotechnology**
Ghent University, Belgium
2005-2008

**Design and Analysis of Clinical Trials**
Institute for Continuous Education in Science
2011

**General secondary education: Latin – Sciences**
Sint-Jan-Berchmans College Mol, Belgium
1999-2005

**Work experience**

**Doctor of applied biological sciences: Cell and gene biotechnology**
Ghent University, Belgium
2010-2014

**Volunteer work**

From 2008-2010 Simon Denil was an elected students’ representative on the board of the faculty of Bioscience Engineering and participated in various committees.