

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/32015> holds various files of this Leiden University dissertation.

**Author:** Akker, Erik Ben van den

**Title:** Computational biology in human aging : an omics data integration approach

**Issue Date:** 2015-02-18

*Chapter 8:*

*Nederlandse samenvatting*

## 1. Introductie

### 1.1 Waarom onderzoek naar veroudering?

Volgens de voorspellingen van het Centraal Bureau van de Statistiek zal ook in de aankomende decennia de levensverwachting van de algemene Nederlandse bevolking onveranderd blijven toenemen. Helaas zullen deze gewonnen jaren doorgaans niet allen in goede gezondheid worden doorgebracht. Immers, naarmate we ouder worden neemt het risico op het krijgen van velerlei zeer algemeen voorkomende ziekten, zoals verscheidene vormen van kanker, diabetes mellitus type II en hart- en vaatziekten toe. Naar verwachting zal de aankomende vergrijzing van onze samenleving gepaard gaan met een toename van dergelijke leeftijdgerelateerde ziekten. Fundamenteel onderzoek naar de oorzakelijke samenhang tussen de verschillende aspecten van gezondheid en het verouderingsproces is dan ook noodzakelijk.

### 1.2 Wat is veroudering?

Veroudering wordt gekenmerkt door een geleidelijke maar onherroepelijke afname van controle en dus functionaliteit over alle organisatorische lagen van het menselijk lichaam. In tegenstelling tot hoe de meeste andere karakteristieken van ons lichaam tot stand komen, zoals bijvoorbeeld oogkleur of lichaamslengte, is veroudering niet vastgelegd in ons DNA, maar is het een resultaat van een passief proces. De jarenlange blootstelling aan stochastische schade bronnen, zoals ultra violet licht of vrije radicalen, zorgt voor een

opstapeling van onvolkomenheden in de cel, de kleinste levende bouwsteen van ons lichaam, en ondermijnt daarmee geleidelijk zijn efficiëntie en incasseringsvermogen. Deze slijtage slag vindt plaats in al onze cellen verspreid over al onze weefsels en vergroot daarmee op den duur de vatbaarheid van ons lichaam voor allerlei kwalen en ziekten. Dientengevolge stelt de opstapeling van onvolkomenheden in onze cellen daarmee een limiet aan de duur van ons leven.

### 1.3 Langlevendheid: de sleutel tot het verouderingsonderzoek

Alhoewel het verouderingsproces zelf onafwendbaar lijkt, is de snelheid waarmee dit optreedt niet hetzelfde voor elk individu. Zo hebben directe familie leden van negentig jarigen, vergeleken bij de algemene bevolking, een grotere kans om zelf ook deze respectabele leeftijd te bereiken. Opvallend hierbij is dat de directe nazaten van deze langlevenden al op middelbare leeftijd minder kans hebben op het krijgen van leeftijd gerelateerde kwalen en ziekten, zoals diabetes mellitus type II of een te hoge bloeddruk. Enerzijds suggereert dit dat zowel de snelheid van veroudering als de uiteindelijke levensverwachting ten minste gedeeltelijk wordt bepaald door erfelijke componenten in ons DNA. Anderzijds suggereert dit dat de erfelijke componenten die deze levensverwachting beïnvloeden ook het risico op het krijgen van leeftijd gerelateerde ziekten verkleinen. Hieruit volgt dat we door een nauwgezette bestudering van de factoren die de snelheid van het verouderingsproces bepalen, we

ook de factoren kunnen identificeren die algemeen ten grondslag liggen aan diverse leeftijd gerelateerde ziekten.

#### 1.4 Onderzoek naar de factoren van veroudering

In het humane verouderingsveld volgt men veelal twee onderzoeksstrategieën. Enerzijds wordt er gezocht naar factoren die vroegtijdig inzicht moeten geven in de mate en aard van veroudering. Het verouderingsproces verloopt bij iedereen anders en deze zogenaamde “biomarkers van biologische veroudering” meten ieder een ander aspect van dit proces. Anderzijds wordt er met verschillende technieken gekeken naar de genetische basis van factoren die de snelheid van veroudering lijken te beïnvloeden. We kijken hiervoor in het DNA van langlevenden en onderzoeken of bepaalde veranderingen in het DNA (variaties) vaker of minder vaak voorkomen in de algemene bevolking. Voor beide strategieën wordt veelal gebruik gemaakt van zogenaamde “omics” meetmethoden, die er op gericht zijn om in één enkel experiment een zo compleet mogelijk beeld te geven van alle veranderingen die de cel op dat moment ondergaat. De cel reguleert zichzelf op vele organisatorische niveaus en voor ieder niveau bestaat er een “omics” methode die met behulp van duizenden tot enkele miljoenen metingen de staat van de cel vastlegt. Een grote uitdaging ligt momenteel in het onderling relateren en interpreteren van de data verkregen met verschillende van deze omics platforms. Het werk in dit proefschrift legt zich toe op het ontwikkelen van methoden voor de geïntegreerde analyse van “omics” data

bronnen, teneinde een beter en completer beeld te krijgen van de moleculaire biologie van veroudering. Twee strategieën voor data integratie zijn toegepast, namelijk: 1) de gecombineerde analyse van meerdere data bronnen, 2) de incorporatie van voorkennis uit externe informatie bronnen.

## 2. Integratieve Analyse van Genexpressie Data Bronnen

De instructies voor het maken van alle bouwstenen van de cel, de eiwitten, liggen besloten in ons DNA, dat veilig afgeschermd ligt in de celkern. Als nieuwe eiwitten worden aangemaakt, wordt eerst het daarbij behorende stukje DNA, het gen, gekopieerd. Door deze kopietjes voor ieder van naar schatting ongeveer 18.000 unieke genen te kwantificeren kunnen we een inzicht krijgen welke genen op dat moment door de cel gebruikt worden. Deze metingen van de zogenaamde expressie van genen wordt onder andere gebruikt voor het classificeren van tumoren van borstkanker patiënten (**hoofdstuk 2** in dit proefschrift) of om veranderingen in het gebruik van genen met toenemende leeftijd te karakteriseren (**hoofdstuk 3** in dit proefschrift). In het eerste gedeelte van dit proefschrift ontwikkelen we een methode voor een integratieve analyse van meerdere genexpressie data bronnen om zo de zekerheid en interpretatie mogelijkheden van onze bevindingen te vergroten.

Uit eerdere studies is gebleken dat in het gebruik van genexpressie metingen voor onderzoek naar borstkanker tumoren of veranderingen met leeftijd er

problemen optreden met betrekking tot de interpretatie en reproduceerbaarheid van de bevindingen. Dit is enerzijds toe te schrijven aan technische aspecten van de metingen, bijvoorbeeld de lage signaal-ruis verhouding, maar anderzijds wordt dit ook veroorzaakt door de heterogeniteit tussen de onderzochte individuen en de complexiteit van de onderzochte lichaamskarakteristieken. Om zowel de reproduceerbaarheid als de interpretatie van deze genexpressie analyses te vergroten hebben we een nieuwe integratieve methode ontwikkeld. In deze methode gebruiken we meerdere studies in een gecombineerde analyse om de kracht en consistentie van onze bevindingen te bevorderen. Daarnaast incorporeren we ook informatie over welke genen hun taken gezamenlijk uitvoeren in de cel, afkomstig uit speciaal voor dit doeleinde opgerichte databases en gebruiken ook dit om de consistentie van onze observaties te verifiëren. De incorporatie van dergelijke gen-gen interacties geeft bovendien voordelen bij de interpretatie van de resultaten. De ontwikkelde methode vereenvoudigt de analyse en interpretatie door systematisch te zoeken naar groepen van genen, met een nauw onderling verbonden biologische functionaliteit, gen modules genaamd, waarvan alle genen een onderlinge consistente relatie met betrekking tot de onderzochte lichaamskarakteristieken vertonen.

In **hoofdstuk 2** passen we de ontwikkelde methodologie toe op data van borstkanker tumoren en demonstreren we de hoge mate van interpreteerbaarheid en reproduceerbaarheid van de verkregen

gen modules. Alhoewel het gebruik van gen modules niet leidt tot een verbetering ten opzichte van de gangbare classificatie van de borstkanker tumoren, komt de samenstelling van de gevonden gen modules wel sterk overeen met genen die al eerder gevonden zijn in borstkanker onderzoek. We concluderen dat toepassing van de voorgestelde methode de data van naar schatting ongeveer 18.000 onafhankelijke metingen naar enkele reproduceerbare en goed interpreteerbare moleculaire biomarkers kan reduceren.

In **hoofdstuk 3** breiden we de voorgestelde methode verder uit door zowel het samenstellen van de modules, als de daaropvolgende stap waarin hun relatie tot karakteristieken van het lichaam wordt onderzocht, wordt uitgevoerd met meerdere gen expressie data bronnen tegelijkertijd. We passen de vernieuwde methode toe op vier genexpressie data bronnen in het bloed van bijna 2.500 individuen en analyseren de gemiddelde expressie van de genen in een module voor consistente veranderingen met leeftijd. We identificeren vijf modules waarvan de expressie gezamenlijk verandert met chronologische leeftijd. In een onafhankelijke studie waarin gen expressies zijn gemeten in het bloed van ongeveer 3.500 individuen bevestigen we de door ons gevonden relaties met leeftijd voor vier van de vijf modules. Een van de gerepliceerde modules bevat het gen *ASF1A*, waarvan we eerder al hebben aangetoond dat dit in andere hoeveelheden tot expressie komt in het bloed van directe nazaten van langlevenden ten opzichte van de algehele bevolking. Bovendien laten we zien dat de genexpressie van het

*ASF1A* gen alsmede het gemiddelde van de gehele gen module indicatief is voor de levensverwachting van ouderen boven de negentig jaar, die we gedurende 10 jaar gevolgd hebben. De *ASF1A* module lijkt dus een interessante kandidaat als nieuwe moleculaire biomarker voor biologische veroudering.

### 3. Integratieve Analyse van DNA Sequentie Data Bronnen

Metingen van variaties in de volgorde en samenstelling van ons erfelijk materiaal noemen we genetische databronnen. Vaak gaat het hierbij over zeer kleine variaties waarin slechts een enkele elementaire bouwsteen van het DNA, een nucleotide, vervangen is door een andere. Genetische data bronnen kunnen echter onderling sterk verschillen in de resolutie en schaal waarop de metingen aan het DNA zijn verricht. Genetische data bronnen zijn er in vele verschillende soorten en maten en verschillen onderling sterk in de resolutie en schaal waarop de metingen aan het DNA zijn verricht. De genetische data bron met de hoogst mogelijke resolutie wordt sequentie data genoemd, omdat deze letterlijk de volgorde van alle basenparen waaruit ons DNA is opgebouwd opsomt. Deze methode rapporteert per persoon gemiddeld ongeveer 3,2 miljoen veranderingen verspreid over het gehele genoom ten opzichte van een "gemiddeld" genoom, dat gebruikt wordt als referentie. De kracht van deze methode is echter ook meteen zijn grootste zwakte. Hoe onderscheiden we de veranderingen die van belang zijn, van alle miljoenen overige veranderingen

die waarschijnlijk geen enkele of slechts zeer beperkte consequenties hebben? In het tweede gedeelte van dit proefschrift richten we ons op het ontwikkelen van innovatieve methoden voor de analyse van sequentie data bronnen.

De consequentie van een variant in het DNA voor het functioneren van een cel heeft voornamelijk te maken met de positionering ten opzichte van stukken erfelijk materiaal die coderen voor de eiwitten in de cel. Genetische varianten die de eiwitcode veranderen, of zelfs geheel verstoren hebben over het algemeen een grotere kans om het functioneren van de cel te beïnvloeden. In **hoofdstuk 4** van dit proefschrift incorporeren we informatie uit externe bronnen, die voorspelt in welke mate een genetische variant een gen verstoort. We passen dit analyse kader toe op een sequentie databron bestaande uit variaties gemeten in 218 langlevenden uit de Leiden Langleven Studie (LLS) en 98 controles uit de algemene Nederlandse bevolking. Gebruikmakende van de voorspellingen of een variatie al dan niet een gen verstoort, hebben we genetische mechanismen onderzocht die gezonde veroudering en langlevendheid zouden kunnen bevorderen.

Zo hebben we onderzocht of bepaalde genen in langlevenden vaker dan verwacht geraakt zijn door zeer ernstig versturende variaties. Uit onderzoek in dier modellen is namelijk gebleken dat als specifieke genen worden verstoord, dit kan leiden tot een significante verlenging van hun levensverwachting. We hebben echter geen bewijs gevonden dat dergelijke genen de lange levensduur in de onderzochte

families verklaarden. Wel vinden we in de hoogbejaarden veel zeer verstorende varianten in de genen *TET2* en *DNMT3A*. Beide genen vervullen een belangrijke rol bij de differentiatie van bloed stamcellen in de verschillende soorten bloed cellen en zijn verstorende varianten in deze genen geassocieerd met een afwijkend bloedbeeld.

Diepere inspectie van de metingen van de meest verstorende varianten in *TET2* en *DNMT3A* liet zien dat waarschijnlijk slechts een klein gedeelte van alle cellen in het bloed van de onderzochte hoogbejaarden deze varianten draagt. Dit is een indicatie dat deze varianten vermoedelijk tijdens het leven zijn ontstaan en dit soort varianten worden somatische mutaties genoemd. In eerdere studies is van dergelijke somatische mutaties in *TET2* en *DNMT3A* al aangetoond dat deze samengaan met het onevenredig uitgroeien van het myeloïde bloedcompartiment en zelfs het risico op het ontwikkelen van Acute Myeloïde Leukemie verhogen (AML). Door de hoogbejaarde dragers van deze genafwijkingen 10 jaar lang te volgen konden we constateren dat de levensverwachting van dragers in het geheel niet wordt aangetast. Deze bevinding impliceert dat bejaarde dragers van dergelijke somatische mutaties een nog hogere leeftijd kunnen bereiken zonder een vorm van bloedkanker te ontwikkelen.

Onze zoektocht naar genetische factoren voor langlevendheid zetten we voort in **hoofdstuk 5** van dit proefschrift door een alternatieve strategie voor het prioriteren van variaties uit sequentie data toe te passen. We doen dit ten eerste door

informatie uit een additionele genetische data bron in onze analyses te incorporeren. Voor dit doeleinde gebruiken we de resultaten van een genomwijde koppeling analyse (linkage scan) naar familiare langlevendheid, waarin wordt gekeken naar de overervingspatronen binnen langlevende families. We passen deze techniek toe op de langlevende families uit de LLS, daarbij ook rekening houdende met de gemiddelde levensduur van de ouders van deze families. Gebruikmakende van dit statistisch kader voor linkage analyse, tonen we aan dat de genomische regio chr13q34 significant vaker dan verwacht in een identieke samenstelling wordt overgeërfd binnen de langlevende families en dat dit effect het sterkst is in de families waarvan de ouders ook zelf langer dan gemiddeld hebben geleefd. De resultaten van deze analyse doet vermoeden dat chr13q34 erfelijke factoren herbergt die bijdragen aan familiare langlevendheid.

Tentweede incorporeren we additionele informatie over bloedspiegels die karakteristiek zijn voor langlevenden, maar ook binnen deze groep een onderscheid maakt tussen langlevenden met langlevende ouders, FH(+), en langlevende families met ouders met een gemiddelde levensverwachting FH(-). Eerder onderzoek heeft namelijk aangetoond dat langlevende families waarvan de ouders ook langer dan gemiddeld hebben geleefd (FH(+)) een tragere schildklier functie hebben dan langlevende families met ouders met een gemiddelde levensduur (FH(-)). Dit suggereert dat de erfelijke componenten voor een tragere schildklier

functie onafhankelijk bijdragen aan langlevendheid.

Vervolgens gebruiken we de reeds in **hoofdstuk 4** van dit proefstuk beschreven sequentie data van de participanten uit de LLS voor het nader bestuderen van genomische regio chr13q34, waarvan we nu vermoeden dat het genen herbergt die betrokken zijn bij zowel een trage schildklierfunctie als langlevendheid. Aangezien deze regio nog steeds meerdere duizenden variaties telt in onze sequentie data passen we nogmaals een strategie toe ter prioritering. We doen dit door op systematische schaal voor iedere variatie te bepalen of dragerschap van deze variatie op chr13q34 in de sequentiedata samen gaat met laag fT3 een marker van de trage schildklierfunctie. Deze analyse laat zien dat dragers van het zeldzame C allel van de variant rs9515460 een veel lagere bloedspiegel van het ongebonden schildklierhormoon hebben. Mogelijkerwijs verklaart deze variant dus zowel de trage schildklier functie als de familiare langlevendheid.

In **hoofdstuk 6** van dit proefschrift presenteren we een nieuw R pakket, genaamd SATORi, waarmee vele van de integratieve berekeningen aan genetische data in dit proefschrift gedaan zijn. In dit pakket zijn enkele routines geïmplementeerd, die enkele van de meest uitgevoerde stappen in een integratieve analyse zouden moeten vergemakkelijken. Het nut en gebruik van dit pakket wordt geïllustreerd aan de hand van enkele publiekelijk beschikbare omics data sets.

## 4. Conclusie en Toekomst

Het onderzoek in dit proefschrift is er op gericht om nieuwe integratieve analyse methoden te ontwikkelen en toe te passen, teneinde een beter en completer beeld te krijgen van de moleculaire biologie van veroudering. Het proefschrift bevat dus resultaten van twee soorten. Ten eerste heeft dit proefschrift integratieve methodologie opgeleverd voor een gecombineerde analyse van meerdere genexpressie datasets. Toepassing daarvan heeft enkele robuuste en interpreteerbare moleculaire profielen opgeleverd, die verder onderzocht kunnen worden voor hun potentie als biomarker voor biologische veroudering. Ten tweede hebben we gewerkt aan methodologie voor een integratieve analyse van genetische componenten die predisponeren voor een vertraagd verouderingsproces. Dit laatste bleek moeilijker dan verwacht, niet alleen door de omvang en complexiteit van de geanalyseerde sequentie data, maar ook doordat er van het verouderingsproces op hoge leeftijd zelf weinig relatief weinig bekend is. Desalniettemin hebben we toch enkele zeer interessante observaties gedaan met betrekking tot het verouderingsproces of de mogelijke vertraging daarvan. Enerzijds vinden we in de genomen van gezonde ouderen een hoge frequentie van varianten die tijdens het leven zijn ontstaan, die vermoedelijk de samenstelling van het bloedbeeld beïnvloeden. Anderzijds vinden we dat de genetische locatie chr13q34 waarschijnlijk bijdraagt aan langlevendheid door de schildklier functie iets te matigen.



Om verder progressie in het verouderingsveld te boeken is niet alleen meer onderzoek naar integratieve analyses nodig, maar kan er ook veel winst geboekt worden met de manier waarop data gemeten wordt. Zo wordt verwacht dat vernieuwingen in de experimentele methodologie van sequentie metingen de kwaliteit van zowel de genetische als de genexpressie datasets sterk zal verbeteren. Tevens kan de waarde van de gemeten data nog sterk vergroot worden door een nog slimmere experimentele opzet te kiezen. Niet alleen kan bijvoorbeeld de invloed van de interindividuele variatie in de data sterk gereduceerd worden door dezelfde mensen herhaaldelijk te meten. Ook kan door middel van een uitgebreide fenotypering of inspectie van de sterfte en geboortecijfers van de familie over meerdere generaties uitgezocht worden in welke familie de genetische component voor langlevendheid het sterkst aanwezig is. Beide methoden worden momenteel in onze groep toegepast in de acquisitie van nieuwe data over veroudering. Een andere trend waarvan het verouderingsveld kan profiteren is dat er steeds meer grote omics datasets publiekelijk beschikbaar worden gemaakt, zoals bijvoorbeeld door het biobank initiatief BBMRI. Met behulp van deze data bronnen kan steeds beter onderzocht worden wat de samenhang is tussen de verschillende soorten metingen zonder dat daar eigen data gecreëerd voor hoeft te worden. Een uitdaging voor de nabije toekomst ligt in het identificeren van gecombineerde biomarker profielen, die rekening houden met meerdere aspecten van veroudering, bijvoorbeeld

cognitieve of cardiovasculaire, en zo dus een getrouwer beeld geeft van de algemene biologische leeftijd. Ook ligt er een uitdaging in het efficiënt incorporeren van informatie over veroudering uit dier modellen teneinde onze zoektocht naar de moleculaire biologie van menselijke veroudering te versnellen.

In dit proefschrift hebben we het belang van data integratie technieken gedemonstreerd voor het doen van onderzoek naar de moleculaire biologie van veroudering. Alhoewel we met behulp van integratieve data analyses enkele interessante aspecten van veroudering hebben ontdekt blijven er echter nog vele hiaten bestaan in ons begrip over veroudering in relatie tot gezondheid. Idealiter zouden we onze kennis over veroudering het beste kunnen verdiepen door meer omics data met meerdere soorten meetmethoden in meerdere weefsels op meerdere tijdstippen in dezelfde personen te meten. De ontwikkeling van integratieve analyse technieken zal onmisbaar zijn voor de analyse van dergelijke rijke en complexe data bronnen en zal in de toekomst dan ook een doorslaggevende rol spelen in het onderzoek naar de moleculaire biologie van veroudering.



